MAPSI

Modèles et **A**lgorithmes **P**robabilistes et **S**tatistiques pour l'**I**nformatique

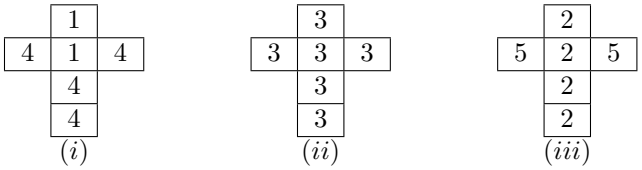
Fascicule de TD

Année 2017-2018

Semaine 1 - Notions élémentaires de probabilités

Exercice 1 – Dés de Gardner

Dans un numéro de la revue *Scientific American* de 1974, M. Gardner proposait un jeu consistant à choisir un dé parmi les trois dés à 6 faces non pipés ci-dessous, de manière à essayer d'obtenir le nombre le plus élevé en lançant le dé une seule fois.



(i) (ii) (iii)

Q 1.1 On vous propose de jouer au jeu à 2 joueurs suivant : chaque joueur mise M euros. Puis on vous demande de choisir un des dés ci-dessus, votre adversaire en choisit ensuite un autre et enfin chacun lance son dé. Celui qui obtient le nombre le plus élevé remporte la mise.

Q 1.1.1 Calculez, pour chaque couple (x, y) de dés la probabilité qu'en jouant avec le dé x on obtienne un résultat plus élevé qu'avec y.

Q 1.1.2 Sachant que la mise est de 30 euros, devez-vous accepter de jouer et, le cas échéant, quel dé devez-vous choisir ? Formellement, quel critère vous permet de statuer ?

Exercice 2 – Indépendence

Soit deux dés à six faces non pipés, un de couleur blanc et un de couleur noir. Les deux sont jetés une fois. On définit les événement suivants :

- le dé blanc donne 1, 2 ou 3.
- le dé blanc donne 2, 3 ou 6.
- la somme des deux dés est égal à 9.
- les deux dés donnent deux nombres égaux, dont la somme est inférieure à 9.

Q 2.1 Quel est la probabilité des ces événements ?

Q 2.2 Quels événements sont deux-à-deux indépendants ?

Q 2.3 Sont-ils mutuellement indépendants ? Si non, trouvez les groupes (à trois ou quatre variables) qui sont mutuellement indépendants.

Exercice 3 – La roulette

Dans les casinos, la roulette contient 37 numéros : 18 rouges, 18 noirs et un vert. Quand la roulette tourne, la bille a autant de chances de tomber sur chacun des 37 numéros. Si l'on mise €1 sur le rouge et que ce dernier sort, on gagne €1, sinon on perd la mise de €1.

Q 3.1 La roulette vous sera t-elle profitable ?

Q 3.1.1 Soit X la variable aléatoire représentant le résultat d'une mise de €1. Quelle est la distribution de probabilité de X ? Quelle est l'espérance de X ?

Q 3.1.2 En moyenne combien gagnerez-vous ou perdrez-vous par mise ?

Q 3.1.3 Combien gagnerez-vous ou perdrez-vous si vous jouez 100 fois en misant €1 à chaque fois ? 1000 fois ? Peut-on en déduire que la roulette n'est pas un jeu profitable ? Justifiez votre réponse.

Exercice 4 – Paradoxe de Simpson

Le recensement des jugements prononcés dans l'état de Floride entre 1973 et 1978 a permis d'établir le tableau suivant, qui présente les sentences en fonction de la couleur de peau de l'accusé :

meurtrier	peine de mort	autre sentence
noir	59	2547
blanc	72	2185

Q 4.1 Calculez la probabilité d'obtenir la peine de mort sachant que l'on est noir, puis sachant que l'on est blanc. Qu'en concluez-vous ?

Q 4.2 En fait le tableau ci-dessus est une synthèse du tableau ci-dessous :

victime	meurtrier	peine de mort	autre sentence
blanche	noir	48	238
	blanc	72	2074
noire	noir	11	2309
	blanc	0	111

Calculez la probabilité d'obtenir la peine de mort conditionnellement à la couleur de peau de l'accusé et de la victime. La justice est-elle clémente envers les noirs dans l'état de Floride ? Justifiez votre réponse.

Exercice 5 – Sport et age

Dans un échantillon aléatoire de 240 personnes, on a recueilli l'information suivante sur l'âge et sur le type de sport le plus fréquemment pratiqué à Jussieu :

âge / activité sportive	moins de 20 ans	[20; 25[ans	[25; 30[ans	plus de 30 ans
jogging	15	20	15	30
natation	15	10	20	25
ping pong	20	10	30	30

Q 5.1 Quelles sont les deux variables aléatoires étudiables ?

Q 5.2 Estimer la loi jointe de ces deux variables.

Q 5.3 Calculer la probabilité qu'a un individu de faire de la natation (dans cet échantillon). Quelle est la probabilité qu'un individu de cet échantillon tiré au hasard ait entre 20 et 25 ans ?

Q 5.4 Calculer la probabilité qu'a un individu qui fait du jogging d'avoir plus de 30 ans (dans cet échantillon).

Q 5.5 Ces deux variables aléatoires semblent-elles indépendantes ?

Exercice 6 – Notes d'examen

Les étudiants d'un même groupe de TD ont obtenu chacun à une UE une note de partiel (sur 50) et une note globale (sur 100, qui intègre note de CC, de partiel et d'examen) ; les données sont les suivantes :

N°Etudiant	*partiel*	*global*
1	45	92
2	23	86
3	50	97
4	46	95
5	33	87
6	21	76
7	13	72
8	30	84
9	34	85
10	50	98

Q 6.1 Calculer le coefficient de corrélation linéaire r entre X (note de partiel) et Y (note globale).

Q 6.2 Représenter graphiquement le nuage de points correspondant aux données. Est-ce que la relation entre les variables semble être linéaire ? Y a-t-il des points aberrants ?

Q 6.3 On a oublié les notes d'un $11^{\text{ème}}$ étudiant qui sont : 40 au partiel et 70 de note globale. Si on les incorpore, quel sera l'effet sur r ?

Exercice 7 – Estimation de la variance d'une loi de probabilité

Soit $X = (X_1, X_2, \ldots, X_k, \ldots, X_n)$ l'échantillon i.i.d. empirique tiré de X_0, variable dont l'espérance $E(X_0) = m$ et la variance $V(X_0) = \sigma^2$ sont deux paramètres inconnus ; on note $\theta = (m, \sigma^2)$ le paramètre bi-dimensionnel.

Q 7.1 Quelle est l'espérance $E_\theta(\bar{X})$ de la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$?

Q 7.2 Quelle est sa variance $V_\theta(\bar{X})$?

Q 7.3 En déduire que $\bar{X}$ est un estimateur sans biais et convergent de m, c'est-à-dire que :

$$E_\theta(\bar{X}) = m \text{ et } \lim_{n \to \infty} E_\theta[(\bar{X}_n - m)^2] = 0.$$

[*rappel* : lorsque n varie, on écrit $\bar{X}_n$ au lieu de $\bar{X}$].

Q 7.4 Montrer que $\frac{1}{n} E_\theta[\sum_{k=1}^{n}(X_k - m)^2] = \sigma^2$.

Q 7.5 Pourquoi ne peut-on pas prendre $\frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2$ comme estimateur de σ^2 ?

Q 7.6 On considère la statistique $Y = \sum_{k=1}^{n}(X_k - \bar{X})^2$. En utilisant la décomposition $X_k - \bar{X} = (X_k - m) - (\bar{X} - m)$, montrer que $Y = \sum_{k=1}^{n}(X_k - m)^2 - n(\bar{X} - m)^2$ puis que $E_\theta(Y) = (n-1)\sigma^2$.

Q 7.7 En déduire que la *variance empirique corrigée* $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$ est un estimateur sans biais de σ^2.

Exercice 8 – Inégalités aux États-unis

Le Monde, 5/09/2014 - ***Les inégalités continuent de se creuser aux États-Unis***

Les inégalités se sont encore accrues aux États-Unis, selon une étude publiée jeudi 4 septembre par la Réserve Fédérale (Fed). Les revenus des 10% les plus riches ont augmenté de 10% entre 2010 et 2013 pour s'inscrire à 397 500 dollars par an (307 000 euros). Dans le même temps, ceux des 40% les moins aisés, ajustés de l'inflation, ont décliné, indique le rapport publié tous les trois ans. Pour les vingt premiers centiles situés au bas de l'échelle, la chute atteint 8% à 15 200 dollars annuels. Si le revenu moyen global a augmenté de 4% au cours des trois dernières années, le revenu médian [...], lui a chuté de 5%. Une tendance qui « correspond à un accroissement de la concentration des revenus durant cette période », indique la Fed.
Ainsi, les 3% les plus riches américains concentrent 30,5% du revenu total en 2013 contre 27,7% en 2010, tandis que la part des 90% les moins riches, elle, a reculé. Par ailleurs, cette catégorie des 3% les plus riches détient 54,4% de la richesse globale (revenu plus patrimoine) contre 44,8% en 1989. A l'autre bout de l'échelle, les 90% les moins riches ont vu leur part tomber à 24,7% contre 33,2% en 1989.
[...]
ORIGINES DES MÉNAGES
Lorsqu'on regarde l'origine des ménages, les inégalités sont encore plus criantes. Le revenu moyen de la population blanche, propriétaire et diplômée a augmenté entre 2010 et 2013, tandis que celui des noirs, des hispaniques, des locataires et des sans diplôme a baissé dans le même temps. De la même façon, le revenu médian des noirs et des hispaniques a chuté de 9% sur la période, quand il ne baissait que de 1% pour les blancs.
Par ailleurs, le rapport indique que le taux de propriétaires de leur logement parmi les ménages américains est tombé à 65,2%. Il s'agit du plus bas niveau constaté depuis 1995. Quand aux familles propriétaires de leurs entreprises, le pourcentage est tombé à 11,7%. Du jamais vu depuis 25 ans.
La thèse de l'économiste français Thomas Piketty développée dans son livre Le capital au XXIe siècle sur l'accroissement des inégalités, a beau avoir été contestée par une partie de la doxa libérale, les chiffres semblent têtus.

Q 8.1 Notons, R_a la variable aléatoire du revenu des salariés américains, indexée par l'année concernée. Nous avons un tirage aléatoire uniforme sur les individus de la population américaine et que nous nous intéressons au revenu de la personne tirée. Nous avons alors :

$$P(R_{2010} > \alpha_{10}^{2010}) = 0.1, \qquad P(R_{2013} > \alpha_{10}^{2013}) = 0.1$$

Que valent α_{10}^{2010} et α_{10}^{2013} ?

Q 8.2 Sans tenir compte de l'inflation, donner une traduction probabiliste de la phrase concernant les 4 premiers déciles des distributions de R_{2010} et R_{2013}.

Q 8.3 Même question sur les 2 premiers déciles.

Q 8.4 En imaginant que nous disposons d'une formule analytique pour $P(R_a), a \in \{2010, 2013\}$ exprimer l'espérance de R_a.

Q 8.5 Donner une modélisation de la phrase suivante : *les 3% les plus riches américains concentrent 30,5% du revenu total en 2013*

Q 8.6 Introduire de nouvelles variables aléatoires (*Origine*, 3 modalités et *Diplome*, 2 modalités) et utiliser les probabilités conditionnelles pour modéliser le premier paragraphe de la seconde partie du texte.

Semaine 2 - Rappels de probabilités

Exercice 9 – Indépendence et conjonction

Soit trois variables aléatoires X, Y, Z. Montrer que si X est indépendante du couple (Y, Z), et Y est indépendante de Z, alors Z est indépendante du couple (X, Y).

Exercice 10 – Indépendances conditionnelles

La loi de probabilité jointe de 3 variables aléatoires X, Y et Z, est donnée par le tableau suivant dans lequel, par exemple, la case 1/12 représente la probabilité $P(X = x_2, Y = y_1, Z = z_1)$:

		$Y = y_1$	$Y = y_2$	$Y = y_3$
$Z = z_1$	$X = x_1$	1/24	1/15	1/8
	$X = x_2$	1/12	7/120	1/8
		$Y = y_1$	$Y = y_2$	$Y = y_3$
$Z = z_2$	$X = x_1$	3/40	1/20	13/120
	$X = x_2$	1/20	3/40	17/120

On note respectivement $X \perp\!\!\!\perp Y$ et $X \perp\!\!\!\perp Y|Z$ l'indépendance probabiliste entre X et Y, et l'indépendance probabiliste entre X et Y conditionnellement à Z.

Q 10.1 D'un point de vue probabiliste, a-t-on $X \perp\!\!\!\perp Y$, $X \perp\!\!\!\perp Z$, $Z \perp\!\!\!\perp Y$? Rappel : si A et B sont indépendants, $P(A, B) = P(A) \times P(B)$.

Q 10.2 A-t-on $X \perp\!\!\!\perp Y|Z$, $X \perp\!\!\!\perp Z|Y$, $Z \perp\!\!\!\perp Y|X$?

Exercice 11 – Indépendances conditionnelles (2)

Soit trois variables aléatoires X, Y, Z, ayant respectivement pour domaines $\{x_1, x_2\}$, $\{y_1, y_2\}$ et $\{z_1, z_2\}$. On a pu déterminer la probabilité jointe $P(X, Y, Z)$ de ces 3 variables :

	y_1		y_2	
	x_1	x_2	x_1	x_2
z_1	0,060	0,140	0,060	0,140
z_2	0,192	0,048	0,288	0,072

Montrez que X est indépendante de Y conditionnellement à Z.

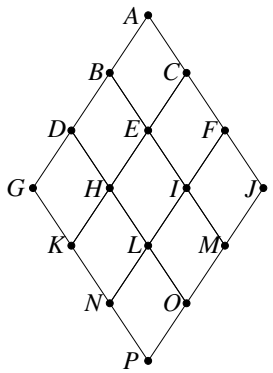
Exercice 12 – Aviation et loi normale

Un pilote de ligne assure régulièrement le trajet Paris-Montpellier. Il s'est amusé à calculer le temps qu'il passe entre le moment où il part de chez lui (à Paris, huit heures du matin) et le moment où il arrive à l'aéroport de Montpellier. Voici le résultat de ses observations : le temps passé dans le RER pour aller jusqu'à Orly suit une loi normale $\mathcal{N}(35 \text{ min}, 8 \text{ min}^2)$; le temps pour préparer le vol/inspecter l'appareil suit une loi $\mathcal{N}(1 \text{ heure}, 16 \text{ min}^2)$; enfin, le temps de vol suit une loi $\mathcal{N}(1 \text{ heure } 10 \text{ min}, 25 \text{ min}^2)$.
Le pilote a décidé de donner rendez-vous à l'aéroport de Montpellier à un de ses collègues. Il ne voudrait pas le fixer trop tôt pour ne pas être en retard, ni le fixer trop tard car cela l'obligerait à attendre. Pour l'aider à choisir l'heure du rendez-vous, calculez la probabilité que le pilote arrive :
1/ entre 10h31 et 10h52.
2/ après 11h.
3/ avant 10h40.
Vous détaillerez les calculs avant de les instancier numériquement.

Indication : lorsque des variables aléatoires $X_1, \ldots, X_n$ suivant des lois normales sont indépendantes, leur somme est une variable aléatoire d'espérance la somme des espérances des X_i et de variance la somme des variances des X_i.

Exercice 13 – Robotique

Un robot doit se rendre du point A au point P en passant par les arêtes du graphe ci-dessous. Le robot est limité dans ses mouvements, aussi ne peut-il que descendre (par exemple, lorsqu'il est en E, il ne peut aller qu'en H ou en I, mais pas en B). Lorsqu'il est sur un nœud du graphe, il peut descendre soit sur l'arête de gauche, soit sur celle de droite. Son programme lui fait choisir 7 fois sur 10 l'arête de gauche et 3 fois sur 10 celle de droite.



Q 13.1 Calculez la probabilité que le robot passe en B pour aller vers P. Faites de même avec C. Soit X_1 la variable aléatoire de modalités $\{B, C\}$ représentant le *point de passage du robot sur le niveau en dessous de A*. Quel est le type de distribution de probabilité suivie par X_1 (binomiale, Poisson, normale, etc) ? Quels sont les paramètres de cette loi ?

Q 13.2 Notez, pour chaque chemin menant à D, le nombre de fois où le robot a été à gauche ou à droite. Faites de même avec E et F. Déduisez-en la probabilité que le robot passe en D, E et F pour aller vers P.

Q 13.3 Calculez la probabilité que le robot passe en G pour aller vers P. Faites de même avec H, I, et enfin J.

Q 13.4 Soit X une variable aléatoire valant 0 si le robot est passé en G, 1 s'il est passé en H, 2 en I et 3 en J. Quelle est la loi de probabilité suivie par X ? Justifiez votre réponse.

Exercice 14 – Modélisation

Nous nous intéressons à la modélisation de phénomène réels par des lois de probabilités standard.

Q 14.1 Dans une image $\mathbf{x}$, nous avons 256 pixels x_j noirs ou blancs. Quelle loi utiliser pour modéliser un pixel ? Que signifient le ou les paramètres de cette loi ?

Q 14.2 Imaginons que nous sommes dans un problème bi-classe, impliquant des chiens et des chats et que nous disposons de 2 modèles optimisés pour chaque classe. Nous faisons aussi l'hypothèse que tous les pixels sont indépendants. Une nouvelle image arrive dont le pixel visible, x_{18}, est allumé : comment déterminer s'il s'agit d'un chien ou d'un chat ?

Q 14.3 Un expert nous indique l'importance des profils dans les images en noir et blanc : c'est à dire l'indice sur chaque ligne où se trouve le premier pixel allumé. Comment modéliser une ligne de l'image ? En imaginant que chaque ligne de l'image est indépendante, exprimer la probabilité d'observation de l'image $\mathbf{x}$ contenant 16

lignes $\{x_1, \ldots, x_{16}\}$.

Q 14.4 Nous disposons de 10 images de chats. Comment vérifier l'hypothèse d'indépendance précédente pour les deux premières lignes ? Indiquer les dimensions des tableaux à introduire.

Q 14.5 Nous modélisation maintenant une image $\mathbf{x}$ dont les pixels x_j peuvent prendre 16 niveaux de gris différents. Quelle loi utiliser pour modéliser un pixel ? Que signifient le ou les paramètres de cette loi ?

Q 14.6 Nous cherchons à modéliser une station Vélib en fonction du nombre de vélos décrochés toutes les 15 minutes. Quelle loi choisir ? Nous voulons pouvoir caractériser chaque station en distinguant 4 périodes dans la journée et en séparant les jours de semaine et les week-end. Formaliser les probabilités que nous cherchons à calculer. Combien faut-il de paramètres ?

Q 14.7 Nous nous intéressons à la durée de vie d'ampoules basse consommation. Nous avons obtenu d'un fabricant le tableau suivant :

durée (dizaines d'années)	0.05	0.1	0.2	0.25	0.3	0.35
nb ampoules	200	100	300	100	200	100

En considérant les lampes modernes comme des objets sans mémoire (pouvant claquer n'importe quand...), quelle loi utiliser pour modéliser cette durée de vie ? Comment estimer le ou les paramètres de cette loi en utilisant ses propriétés ?

Semaine 3 - Max de vraisemblance et max *a posteriori*

Exercice 15 – MAP

Soit X une variable aléatoire définie sur l'ensemble des nombres entiers positifs. X suit la loi géométrique de paramètre $p \in [0, 1]$ si $P(X = n) = (1 - p)^{n-1}p$. On a observé 5 réalisations (obtenues indépendamment les unes des autres) d'une variable X suivant la loi géométrique :

4	2	6	5	8

.

Q 15.1 Estimez par maximum de vraisemblance la valeur du paramètre $\boldsymbol{\theta} = p$ de la loi.

Q 15.2 Avant le tirage de l'échantillon, nous avions une connaissance a priori sur le paramètre $\boldsymbol{\theta}$: ce dernier suivait a priori une loi Beta de paramètres 4 et 5, autrement dit $\pi(\theta) \propto \theta^3(1 - \theta)^4$. Estimez la valeur du paramètre $\boldsymbol{\theta} = p$ par maximum a posteriori.

Exercice 16 – MAP 2

Soit X une variable aléatoire suivant la loi binomiale $\mathcal{B}(K, \theta)$, où K est une constante supposée connue. On observe un échantillon $\{x_1, \ldots, x_n\}$ de taille n d'instanciations de cette variable aléatoire.

Q 16.1 Calculez la valeur de θ par maximum de vraisemblance. Bien entendu, vous démontrerez mathématiquement votre résultat.

Q 16.2 Des études statistiques nous indiquent que θ suit une loi Beta *a priori* $\pi(\theta) = \text{Beta}(\theta, a, b)$. Quelle est la valeur a posteriori de θ ? Justifiez mathématiquement votre réponse.

Exercice 17 – Max de vraisemblance

Dans une urne se trouvent des boules de 4 couleurs différentes : rouge (R), bleues (B), vert (V) et jaune (J). On ne connaît pas la quantité de boules dans l'urne ni la proportion des différentes couleurs. Soit la variable aléatoire $X =$ Â« couleur d'une boule tirée au hasard dans l'urne Â». On se propose de représenter la distribution de probabilité de X par une distribution catégorielle de paramètres $\theta = \{p_R, p_B, p_V, p_J\}$, c'est-à-dire :

$$P(X = R) = p_R \quad P(X = B) = p_B \quad P(X = V) = p_V \quad P(X = J) = p_J$$

avec, bien entendu, $p_R, p_B, p_V, p_J \geq 0$ et $p_R + p_B + p_V + p_J = 1$.
Afin d'estimer les paramètres de la distribution, on a tiré avec remise un échantillon des boules de l'urne et on a observé leurs couleurs, que l'on a retranscrites dans le tableau suivant :

R	R	R	R	B	B	V	V	V	J

Q 17.1 Estimez par maximum de vraisemblance les paramètres de la distribution $P(X)$. Vous justifierez votre réponse.

Q 17.2 Un expert, qui a pu observer brièvement l'urne, propose une information *a priori* sur la distribution des couleurs sous la forme d'un *a priori* de Dirichlet d'hyperparamètres $\alpha = \{\alpha_R = 3, \alpha_B = 2, \alpha_V = 4, \alpha_J = 3\}$. Une distribution de Dirichlet d'hyperparamètres $\alpha_1, \ldots, \alpha_K$ est définie de la manière suivante : pour tout K-uplet $(x_1, \ldots, x_K)$ tel que $x_i \in]0, 1[$ pour tout $i \in \{1, \ldots, K\}$ et tel que $\sum_{i=1}^{K} x_i = 1$, on a :

$$Dir(x_1, \ldots, x_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

où $\Gamma(\cdot)$ est la fonction Gamma usuelle.
Estimez par maximum a posteriori les paramètres de la distribution $P(X)$ sur les couleurs des boules de l'urne. Vous justifierez votre réponse.

Exercice 18 – Maximum a posteriori, maximum de vraisemblance

Une pièce de monnaie peut être plus ou moins biaisée en faveur de *Pile* ou de *Face*.
On prend pour paramètre θ la probabilité de *Pile* :

$$P_\theta(Pile) = \theta.$$

L'ensemble des valeurs possibles pour θ est $\Theta = \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$; les probabilités a priori $\pi(\theta)$ de la v.a. $\tilde{\theta}$ sont :

θ	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$
$\pi(\theta)$	0.1	0.2	0.4	0.2	0.1

On effectue 5 lancers indépendants de la pièce et on observe le nombre x de résultats *Pile* obtenus ; la v.a. X a donc pour valeurs possibles $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$.

Q 18.1 Quelle est la loi suivie par X conditionnellement à l'hypothèse $\tilde{\theta} = \theta$? Calculer tous les éléments du tableau des probabilités conditionnelles $P(x|\theta),\ (x, \theta) \in \mathcal{X} \times \Theta$. (on pourra se servir d'une table et mettre à profit les symétries des données)

Q 18.2 Déduire de la question précédente les valeurs des éléments du tableau des probabilités jointes $\pi(x, \theta),\ (x, \theta) \in \mathcal{X} \times \Theta$. À partir de ce tableau, comment peut-on retrouver la loi a priori $\{\pi(\theta)\}$ de la v.a. $\tilde{\theta}$? comment trouve-t-on la loi a priori de X ? Calculez-la.

Q 18.3 Déduire de ce qui précède les valeurs des éléments du tableau des probabilités a posteriori $\pi(\theta|x),\ (x, \theta) \in \mathcal{X} \times \Theta$.

Q 18.4 Donner les valeurs d'acceptation des diverses hypothèses sur la valeur du paramètre :

Q 18.4.1 quand la règle de décision est celle de la *probabilité d'erreur minimum* ; Cette règle équivaut à une règle de la probabilité maximum d'une décision juste : dans chaque ligne du tableau des $\pi(\theta|x)$ il faut choisir

$$d(x) = \operatorname*{Argmax}_\theta \pi(\theta|x).$$

Q 18.4.2 quand la règle de décision est celle du *maximum de vraisemblance*.

Q 18.4.3 Quand ces deux règles donnent-elles le même résultat ?

Exercice 19 – Loi exponentielle et MAP

La loi exponentielle est une loi continue dont la fonction de densité est : $f(x) = \lambda e^{-\lambda x}$ pour tout $x > 0$. Elle sert, entre autres, pour caractériser la durée de vie des composants électroniques. Le tableau suivant recense les durées de vie (en années) observées pour un échantillon de 10 composants électroniques :

2	7	3	4	1	2	6	5	1	9

Q 19.1 On suppose que la distribution des durées de vie est effectivement une loi exponentielle. Estimez par maximum de vraisemblance la valeur de λ.

Q 19.2 Après discussion avec un expert en électronique, on a un *a priori* sous la forme d'une loi Gamma de densité $g(x) = \frac{1}{\Gamma(5)} x^4 e^{-x}$. Estimez par maximum a posteriori la valeur de λ.

Exercice 20 – Loi géométrique et maximum de vraisemblance

Un robot effectue des actions et, afin de déterminer son efficacité, un observateur a noté les temps d'exécution (en secondes) de 100 tâches qu'il a effectuées. Ces temps sont indiqués dans le tableau ci-dessous :

temps (en secondes)	1	2	3	4	5	6	7	8	9
nb observations	31	22	15	11	7	5	4	2	3

Q 20.1 L'observateur pense que la variable aléatoire X=Â« temps d'exécution Â» suit une loi géométrique. On rappelle que la loi géométrique de paramètre p est telle que $P(X = k) = p(1-p)^{k-1}$, pour tout entier $k \geq 0$. Déterminez la valeur du paramètre p par maximum de vraisemblance.

Exercice 21 – Codage de textes et approche *Naive Bayes*

Soit un ensemble des documents $\{d_i\}_{i=1,\ldots,N}$, chacun de ces documents étant composé d'une suite $|d_i|$ mots $w_j : d_i = (w_1, \ldots, w_{|d_i|})$. Nous souhaitons classer ces documents dans des classes bien identifiées (par exemple, la classe des documents relatifs aux automobiles ou bien celle relative à la biologie). Pour cela, nous allons construire un modèle Θ_m pour chaque classe de documents. Nous nous appuierons ensuite sur ces modèles pour construire un classifieur de documents.

Q 21.1 Donnez un cas d'usage classique pour ce type de classifieur de textes. Imaginez un modèle simple pour répondre à ce problème.

Q 21.2 Nous allons construire un modèle Θ_m basé sur la probabilité d'apparition des mots : une classe de document sera donc caractérisée par des mots ayant une forte probabilité d'apparition et des mots ayant une faible probabilité d'apparition.
Calculez la probabilité $P(d_i|\Theta_m)$ d'observer un document d_i en fonction des $P(w_j|\Theta_m)$ (probabilité d'observation d'un mot w_j) en faisant l'hypothèse que les tirages des w_j sont indépendants. Expliquez pourquoi cette hypothèse est (très) forte en vous basant sur un exemple.

Q 21.3 Introduisons la variable x_i^j qui décrit le nombre d'apparitions du mot j dans le document i. Introduisons également l'ensemble $D = \{w_1, \ldots, w_{|D|}\}$ contenant tout le vocabulaire utilisé dans un corpus.
Écrivez $P(d_i|\Theta_m)$ comme une fonction de x_i^j.

Q 21.4 Pour trouver les paramètres Θ_m, nous allons maximiser la log-vraisemblance de ces paramètres sur l'ensemble de la classe m du corpus.
Montrez que le problème d'optimisation permettant de trouver Θ_m en fonction des $P(w_j|\Theta_m)$ s'écrit :

$$\Theta_m = \arg\max_{\Theta} \sum_{i=1}^{|C_m|} \sum_{j=1}^{|D|} x_i^j \log P(w_j|\Theta),$$

La voiture est au garage. La voiture est sur la route. La jeep roule sur la route.	**Classe 1** Chaque ligne est un document
Le lion est dans la savane. Le lion guette sa proie. L'antilope est une proie pour le lion.	**Classe 2** Chaque ligne est un document

TABLE 1 – Base de donnée d'apprentissage de phrases relatives aux voitures et aux lions.

où, par abus de notation, l'ensemble des documents de la classe m dans le corpus est noté $C_m = \{d_1, \ldots, d_{|C_m|}\}$.

Q 21.5 Simplifions les notations $P(w_j|\Theta_m) \to \theta_j$. Le modèle devient : $\Theta_m = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_j \\ \vdots \\ \theta_{|D|} \end{bmatrix} = \begin{bmatrix} \vdots \\ P(w_j|\Theta_m) \\ \vdots \end{bmatrix}$. En introduisant la contrainte $\sum_j \theta_j = 1$, les paramètres optimaux (maximisant la vraisemblance) sont les suivants : $\theta_j = \frac{\sum_{d_i \in C_m} x_i^j}{\sum_{d_i \in C_m} \sum_{j \in D} x_i^j}$ Pourquoi avoir introduit la contrainte $\sum_j \theta_j = 1$? A quoi correspondent chacun des paramètres θ_j ?
Soit la base de donnée de la table 1, établissez le dictionnaire D puis calculez Θ_1 et Θ_2.

Q 21.6 A votre avis, pourquoi cet algorithme s'appelle-t-il *Naive Bayes* ?
Intuitivement, pensez-vous qu'il donne de bons résultats ?

Q 21.7 Calculez et comparez les probabilités $P(\Theta_1|d_i)$ et $P(\Theta_2|d_i)$ pour les documents de la base d'apprentissage.

Q 21.8 Analyse des résultats :
Êtes-vous satisfaits des résultats obtenus ?
Dans la pratique, on redéfinit les paramètres comme :

$$\theta_j = \frac{\sum_{d_i \in C_m} x_i^j + 1}{\sum_{d_i \in C_m} \sum_{j \in D} x_i^j + |D|}$$

Expliquez les raisons de ce choix en analysant les résultats précédents et les résultats sur les phrases suivantes :
- Le monospace roule sur la route
- Le lion apprécie également les gazelles

Q 21.9 Quels sont les mots qui *participent* le plus à la classification des phrases ? Quels traitements effectuer pour avoir des *mots-clés* plus pertinents ?

Semaine 4 - l'algorithme EM

Exercice 22 – Algorithme des K-moyennes

L'algorithme des K-moyennes procède selon les étapes suivantes pour partitionner un ensemble de données en K clusters :

- Initialiser aléatoiremet les K prototypes.
- Répéter jusqu'à stabilité des clusters :
  - Partitionner les exemples en les affectant aux prototypes dont ils sont le plus proche en terme de distance euclidienne
  - Redéfinir les prototypes de manière à ce qu'ils correspondent aux centres de gravité des partitions

Q 22.1 Soit l'ensemble d'exemples en dimension 2 :
$D = \{(0,-4),(0,-3),(1,-3),(1,-2),(0,4),(-1,1),(-1,2),(0,3)\}$
Faire tourner l'algorithme des K-moyennes en prenant comme point de départ les prototypes $(0,-6)$ et $(-1,1)$.
On sait maintenant que les données sont issues d'une mixture de 2 Gaussiennes multivariées de paramètres respectifs (μ_1,Σ_1) et (μ_2,Σ_2).

Q 22.2 Par quoi pourrait-on remplacer la distance euclidienne de l'algorithme donné ci-dessus pour que les données soient toutes affectées à la gaussienne qui les a le plus probablement générées ?

Q 22.3 Comment doit-on modifier l'étape de mise à jour des prototypes dans ce cas ?

Q 22.4 Quelles différences alors avec l'algorithme EM vu en cours ? En quoi un algorithme EM pourrait-il être supérieur à celui-ci ?

Exercice 23 – EM et mixture de gaussiennes

Les prix fonciers d'un quartier suivent une mixture de 2 gaussiennes, de paramètres respectifs (μ_1,σ_1^2) et (μ_2,σ_2^2). Le tableau ci-dessous recense les prix en 100K€ de quelques transactions immobilières :

8	1	4	3	3	5	7	5	4	5

On appellera π_1 et π_2 les coefficients des 2 gaussiennes dans la mixture.

Q 23.1 Triez les éléments de l'échantillon par ordre croissant et servez-vous des 5 plus petites valeurs pour estimer par maximum de vraisemblance (μ_1,σ_1^2) et des 5 plus grandes pour estimer, toujours par maximum de vraisemblance (μ_2,σ_2^2). Dans ces conditions, quelles valeurs faudrait-il logiquement affecter aux poids π_1 et π_2 ?
On rappelle que la fonction de densité de la loi normale est :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

Q 23.2 En partant du $\Theta^0 = \{\mu_1,\sigma_1^2,\pi_1,\mu_2,\sigma_2^2,\pi_2\}$ obtenu à la question précédente, estimez la valeur de Q_i^1 selon l'algorithme EM.

Q 23.3 Estimez la valeur du paramètre Θ^1.

Exercice 24 – EM et loi jointe de deux variables

Soit deux variables aléatoires discrètes A et B dont les domaines respectifs sont $\{a_1,a_2\}$ et $\{b_1,b_2\}$. On cherche à estimer la distribution jointe de A et B en utilisant l'algorithme EM sur l'échantillon suivant dont certaines valeurs sont manquantes (Â≪ ? Â≫) :

(a_1,b_1)	(a_2,b_1)	$(?,b_2)$	$(a_2,?)$	(a_2,b_2)	(a_1,b_2)	$(?,b_1)$	$(a_1,?)$	(a_1,b_2)	(a_2,b_1)

Q 24.1 Que représente, dans ce cas, le paramètre Θ^t de l'algorithme EM ?

Q 24.2 Supposons que l'on démarre l'algorithme EM avec une loi jointe $P(A,B)$ estimée uniforme. Quelles sont les valeurs des $Q_i^{t+1}(x_i^h)$, $i = 1,\ldots,10$, si l'on applique une étape E de l'algorithme EM ?

Q 24.3 En utilisant les valeurs des $Q_i^{t+1}(x_i^h)$ de la question précédente, donnez l'expression $\log L^{t+1}(\mathbf{x}^o,\Theta)$ en fonction des paramètres de Θ.

Q 24.4 Calculez Θ^1, c'est-à-dire appliquez l'étape M de EM.

Q 24.5 Quelles sont les nouvelles valeurs des $Q_i^{t+1}(x_i^h)$, $i = 1,\ldots,10$, si l'on applique à nouveau une étape E de l'algorithme EM ?

Q 24.6 En utilisant les valeurs des $Q_i^{t+1}(x_i^h)$ de la question précédente, donnez l'expression $\log L^{t+1}(\mathbf{x}^o,\Theta)$ en fonction des paramètres de Θ.

Exercice 25 – EM et loi exponentielle

Une entreprise s'intéresse à la durée de vie d'un composant informatique. Pour cela, elle a fait fonctionné 10 composants et a noté les temps (en mois) au bout desquels lesdits composants ont cessé de fonctionner. Les résultats sont répertoriés dans le tableau ci-dessous. Pour ces tests, l'entreprise a imposé un timeout de 20 mois et, lorsque les composants continuaient à fonctionner après ce délai, elle a noté dans le tableau un Â« ? Â».

1	2	2	3	3	7	10	?	?	?

La loi classiquement utilisée en statistiques pour modéliser la durée de vie de composants est la loi exponentielle (de paramètre λ) dont la fonction de densité est $f(x|\lambda) = \lambda e^{-\lambda x}$, pour tout $x \geq 0$.

Q 25.1 En ne prenant en compte que les données numériques du tableau (c'est-à-dire sans tenir compte des Â« ? Â»), estimez la valeur du paramètre λ par maximum de vraisemblance.

Q 25.2 La troncature de la loi exponentielle sur l'intervalle $[20, +\infty[$ est la loi $g(x|\lambda) = \begin{cases} 0 & \text{si } x < 20 \\ \mu e^{-\lambda x} & \text{si } x \geq 20 \end{cases}$
Donnez une expression du paramètre μ en fonction de λ (suggestion : l'intégrale d'une fonction de densité sur tout son domaine de définition est égale à 1). On rappelle que la dérivée de $e^{\alpha x}$ par rapport à x est égale à $\alpha e^{\alpha x}$.

Q 25.3 On va maintenant exécuter l'algorithme EM afin de déterminer le paramètre λ de la loi exponentielle $f(x|\lambda)$ en tenant compte des Â« ? Â». Donnez une expression de $Q_i^1(x_i^h) = p(x_i^h|x_i^o, \lambda)$, $i = 8, 9, 10$, en fonction d'une valeur λ_0 (qui sera, par la suite égale au λ estimé à la question 25.1). On comprendra le conditionnement par x_i^o comme Â« étant donné que l'on n'a pas observé l'arrêt du composant Â».

Q 25.4 Que vaut $p(x_i^o|\lambda)$ pour $i = 8, 9, 10$, c'est-à-dire la probabilité de ne pas observer l'arrêt du composant. Sachant que $p(x_i^o, x_i^h|\lambda) = p(x_i^o|\lambda) \times p(x_i^h|x_i^o, \lambda)$, donnez l'expression de $p(x_i^o, x_i^h|\lambda)$.

Q 25.5 Donnez une expression de $Q_i^1(x_i^h) \log\left(\frac{p(x_i^o, x_i^h|\lambda)}{Q_i^1(x_i^h)}\right)$, $i = 8, 9, 10$, en fonction de λ et de l'expression obtenue dans la question précédente.

Q 25.6 Donnez une expression de $\int Q_i^1(x_i^h) \log\left(\frac{p(x_i^o, x_i^h|\lambda)}{Q_i^1(x_i^h)}\right) dx_i^h$, pour $i = 8, 9, 10$.

On rappelle que, pour $\alpha > 0$, $\int_{20}^{+\infty} x e^{-\alpha x} dx = \frac{1}{\alpha}\left(20 + \frac{1}{\alpha}\right) e^{-20\alpha}$ (on le démontre aisément par intégration par parties).

Q 25.7 En déduire une expression pour $\log L^{t+1}(\mathbf{x}^o, \lambda)$

Q 25.8 On peut maintenant appliquer EM. En supposant que $\Theta^0 = \lambda_{ML}$, le lambda estimé par maximum de vraisemblance à la question 25.1, estimez Θ^1.

Q 25.9 En utilisant l'expression obtenue à la question précédente, pour quelle valeur de λ aura-t-on convergence ?

Semaine 5 - Tests d'hypothèses, d'ajustement et d'indépendance

Exercice 26 – Test entre hypothèses simples

Parmi les personnes atteintes d'une certaine maladie, que l'on ne sait pas traiter, 36% guérissent spontanément, les 64% restant devenant des malades chroniques.
Un laboratoire pharmaceutique propose un remède très coûteux avec lequel, affirme-t-il, le pourcentage de guérison passe à 50%.
Un service hospitalier doute de l'efficacité de ce remède ; pour le tester, il l'administre à un échantillon de 100 patients atteints de la maladie ; les patients sont numérotés de $k = 1$ à $k = 100$.

Q 26.1 – Mise en place du test d'hypothèse :

Q 26.1.1 Quelles sont les hypothèses simples en présence ? (on appellera θ le paramètre). Laquelle doit-on prendre comme hypothèse H_0 ?

Q 26.1.2 Au patient k est associée la variable X_k qui prend la valeur 1 si ce patient guérit et la valeur 0 sinon. Quel est le type de loi suivie par X_k dans les deux hypothèses H_0 et H_1 ?
Vérifier que les probabilités élémentaires des deux lois peuvent se mettre sous la forme :

$$P_\theta(X_k = x_k) = \theta^{x_k}(1-\theta)^{1-x_k},\ x_k \in \{0,1\}$$

avec $\theta = \theta_0$ pour l'une, $\theta = \theta_1$ pour l'autre.

Q 26.1.3 En déduire l'expression de la vraisemblance [= la probabilité d'obtenir l'échantillon $(x_1, \ldots, x_n)$ conditionnellement à θ],

$$L(\mathbf{x},\, \theta) = L(x_1, \ldots,\, x_k,\, \ldots,\, x_n,\, \theta) = \prod_{k=1}^{n} P_\theta(X_k = x_k).$$

Montrer qu'elle s'exprime comme une fonction de la moyenne empirique $\bar{x}$ et de θ.

Q 26.1.4 En déduire que, pour tout test du rapport de vraisemblance $L(\mathbf{x},\, \theta_0)/L(\mathbf{x},\, \theta_1)$, il existera un nombre positif λ tel que :

$$\bar{x} < \lambda \ \Rightarrow \text{accepter } H_0$$
$$\bar{x} > \lambda \ \Rightarrow \text{rejeter } H_0.$$

Q 26.2 – Réalisation effective du test :

Q 26.2.1 Que représente la variable $Y = n\bar{X}$? Quelle est la loi de Y dans l'hypothèse H_0 ?

Q 26.2.2 La table ci-dessous donne les probabilités exactes d'observer k guérisons au moins parmi les 100 malades sous l'hypothèse H_0 (de $k = 42$ à $k = 50$) ; les valeurs obtenues par l'approximation normale sont données au-dessous.

k	42	43	44	45	46	47	48	49	50
$P_{\theta_0}(Y \geq k)$	0.126	0.089	0.060	0.040	0.025	0.015	0.009	0.0052	0.0029
val. appr.	0.125	0.089	0.059	0.038	0.023	0.014	0.008	0.0047	0.0025

Au niveau de signification $\alpha = 0.05$, quelles sont les valeurs de k pour lesquelles on doit : accepter l'hypothèse H_0 ? rejeter H_0 ? rejeter H_0 avec une certaine probabilité ? (que l'on précisera).

Q 26.2.3 Sous l'hypothèse H_1,

$$P_{\theta_1}(Y \leq 43) = 0.0967 \text{ et } P_{\theta_1}(Y = 44) = 0.039$$
$$P_{\theta_1}(Y \leq 46) = 0.242 \text{ et } P_{\theta_1}(Y = 47) = 0.066$$

En déduire la puissance du test de niveau de signification $\alpha = 0.05$.

Q 26.2.4 Mêmes questions que précédemment mais cette fois au niveau de signification $\alpha = 0.01$.

Q 26.2.5 L'approximation normale est-elle bonne ici ? Quel est le théorème de convergence qui laisse prévoir ce fait ?

Q 26.3 On suppose que le chef du service hospitalier est capable :
— d'attribuer des probabilités *a priori* aux deux hypothèses

$$\pi_0 = P(H_0)\,;\, \pi_1 = P(H_1) = 1 - \pi_0\,;$$

— et d'estimer le coût d'une erreur de $1^{\text{ère}}$ espèce, C_0 et de 2^{eme} espèce C_1, ce qui permet de se placer dans le cadre de la statistique bayésienne.

Soit un test T_W, caractérisable par sa région critique W, c'est-à-dire tel que :

$$H_0 \text{ rejeté} \Leftrightarrow x \in W$$
$$H_0 \text{ accepté} \Leftrightarrow x \notin W$$

Q 26.3.1 Montrer que l'espérance mathématique du coût de ce test est :

$$\pi_0 C_0 \sum_{x \in W} L(\mathbf{x},\, \theta_0) + \pi_1 C_1 \sum_{x \notin W} L(\mathbf{x},\, \theta_1).$$

Q 26.3.2 En déduire que, s'il existe $x \in W$ tel que :

$$L(\mathbf{x},\, \theta_0) > \frac{\pi_1 C_1}{\pi_0 C_0} L(\mathbf{x},\, \theta_1),$$

alors il existe un autre test d'espérance de coût strictement inférieure à celle du test T_W.

Q 26.3.3 Montrer de même que s'il existe $x \notin W$ tel que

$$L(\mathbf{x},\, \theta_0) < \frac{\pi_1 C_1}{\pi_0 C_0} L(\mathbf{x},\, \theta_1),$$

alors il existe un autre test d'espérance de coût strictement inférieure à celle du test T_W.

Q 26.3.4 En déduire qu'un test optimal bayésien est nécessairement un test du rapport de vraisemblance. Comment λ varie-t-il avec π_0 et π_1 d'une part et C_0 et C_1 d'autre part ?
Donner un test optimal bayésien lorsque $\pi_0 = 0.95$, $C_0 = 1$ et $C_1 = 10$ (unités monétaires).

Exercice 27 – Investissement à la bourse

Vous voulez investir à la bourse. Afin d'optimiser vos profits, vous relevez pendant 16 jours le cours du CAC40. Au début de ces deux semaines, celui-ci vaut 5715 points. Dans l'échantillon de 16 jours, le CAC40 vaut en moyenne 5726,025 points, avec un écart-type de 6 points. Vous ne voulez investir que si le CAC40 est à la hausse.

Q 27.1 Sachant que la variable $X =$ *valeur du CAC40* suit une loi normale de variance 36, effectuez un test d'hypothèse de niveau de confiance 99% pour savoir si le CAC40 a augmenté. Vous préciserez bien les hypothèses H_0 et H_1.

Q 27.2 D'après le test précédent, peut-on conclure que le CAC40 a augmenté ?

Q 27.3 Calculez la puissance du test pour $\mu = 5726,025$. Pour vous aider, vous pourrez supposer que si une variable $Y \sim \mathcal{N}(0,1)$, alors :

$$P(Y > -1) = 0,8413 \qquad P(Y > -2) = 0,9772 \qquad P(Y > -3) = 0,9986 \qquad P(Y > -4) \approx 1.$$

Exercice 28 – Test d'ajustement du χ^2

Dans un supermarché, on maintient 8 caisses de plus de 10 articles en opération durant les nocturnes du jeudi. Normalement, la clientèle devrait se répartir uniformément entre les caisses. Afin de vérifier cela, on a recensé le nombre de clients passés à chacune des caisses un jeudi soir. Les résultats observés ont été les suivants :

Numéro de la caisse	Nombre de clients
1	72
2	70
3	71
4	52
5	45
6	59
7	67
8	48
Total	484

Hypothèse H_0 à tester : *la clientèle se répartit uniformément entre les 8 caisses.*

Q 28.1 Sous l'hypothèse H_0, quels sont les effectifs théoriques ν_i dans chaque classe ?

Q 28.2 Calculer la statistique d'ajustement :

$$A = \sum_{i=1}^{I} \frac{(n_i - \nu_i)^2}{\nu_i}.$$

Q 28.3 Quel est le nombre de degrés de liberté ? Au niveau de signification $\alpha = 0.05$, doit-on accepter H_0 ?

Exercice 29 – Boules de couleur

Soit une urne contenant des boules de 5 couleurs différentes : (R)ouges, (B)leues, (V)ertes, (J)aunes, (N)oires. On suspecte que la distribution de probabilité sur les couleurs des boules de l'urne est la suivante :

$$P(R) = 0,2 \quad P(B) = 0,4 \quad P(V) = 0,1 \quad P(J) = 0,2 \quad P(N) = 0,1.$$

Par ailleurs, on a tiré un échantillon i.i.d. de 20 boules et on a noté le nombre de boules de chaque couleur :

Couleur	R	B	V	J	N
Nb boules	2	9	4	5	0

Faites un test d'ajustement avec un niveau de confiance $1 - \alpha = 90\%$ pour déterminer si, oui ou non, la distribution de probabilité sur les couleurs des boules est celle indiquée ci-dessus.

Exercice 30 – Test d'indépendance du χ^2

Un échantillon de 200 contribuables est prélevé afin de vérifier si le revenu brut annuel d'un individu est un caractère dépendant du niveau de scolarité de l'individu. Les observations recueuillies sont données dans le tableau suivant :

scolarité (années) $\rightarrow$ revenu (kF) $\downarrow$	$[0;7[$	$[7;12[$	$[12;14[$	$[14;\rightarrow[$	*total*
$[0;75[$	17	14	9	5	45
$[75;120[$	12	37	11	5	65
$[120;200[$	7	20	20	8	55
$[200;\rightarrow[$	4	9	10	12	35
total	40	80	50	30	200

Q 30.1 On admet que les fréquences relatives déduites des marges du tableau donnent les vraies lois de probabilité, $p_{r.}$ et $p_{.s}$ des variables $R(evenu)$ et $S(colarité)$. Donner le tableau des fréquences théoriques, $200 \times p_{rs}$, correspondantes en cas d'indépendance des deux variables.

Q 30.2 Calculer le χ^2. Expliquez pourquoi il y a 9 degrés de liberté. Doit-on rejeter l'hypothèse d'indépendance au risque $\alpha = 5\%$?

Exercice 31 – Notation en MAPSI

On sait, par expérience, que les notes de partiel de MAPSI suivent une loi normale $\mathcal{N}(\mu; 6^2)$. On considère l'échantillon de notes i.i.d. suivant :

10	8	13	20	12	14	9	7	15

.

Q 31.1 Par expérience, les années précédentes, la moyenne au partiel de MAPSI était égale à 14. Dressez un test d'hypothèse de niveau de confiance $1 - \alpha = 95\%$ pour confronter les hypothèses $H_0 =$ *la moyenne est égale à 14* et $H_1 =$ *la moyenne a baissé, i.e., elle est inférieure à 14.*

Q 31.2 Calculez la puissance du test pour une moyenne de 12 (H_1 : la moyenne est égale à 12).

Exercice 32 – Il faut assurer

La loi oblige tout automobiliste à contracter une assurance. La prime exigée annuellement d'un assuré dépend de plusieurs facteurs : la zone habitée, le type de véhicule, l'utilisation à des fins commerciales ou non, la distance estimée que parcourra l'assuré... Il est presque impossible d'estimer la distance parcourue par un automobiliste pour une année donnée. Voilà pourquoi tous les assurés d'un véhicule non utilisé à des fins commerciales se voient imposer le même montant sur ce point. Celui-ci est fonction de la distance moyenne parcourue annuellement par les automobilistes de cette catégorie. Des études ont montré par le passé que celle-ci était de 18000 km avec un écart-type de 5000 km. Le montant que l'on prévoit d'exiger est de 2 centimes du km, autrement dit $18000 \times 0,02 = 360$€. Le montant de cette prime a continuellement augmenté ces dernières années, de telle sorte que l'opinion publique commence à être très mécontente et à exercer de fortes pressions sur les compagnies d'assurances pour qu'elles baissent leurs tarifs.
C'est ainsi que la MAIF est priée de réévaluer tous les facteurs considérés dans le calcul de la prime. Le plus vulnérable de ces facteurs est précisément la distance parcourue annuellement. Un statisticien est donc chargé de réexaminer le bien-fondé de l'estimation à 18000 km de la moyenne contestée. La démarche qu'il compte suivre est de prélever rapidement un échantillon de 400 individus afin de tester si la moyenne a effectivement diminué. Si tel est le cas, une étude plus exhaustive, menée sur un grand nombre d'assurés, sera entreprise afin d'estimer très précisément la valeur de la moyenne. Sinon, ce facteur ne sera pas révisé.
La variable qu'étudie le statisticien est X : la distance parcourue en 2016 (dernière année complète sur laquelle on peut fonder l'étude), par un véhicule utilisé à des fins non commerciales. Il décide pour l'instant de ne pas remettre en cause l'estimation de l'écart-type σ de X (5000 km). En revanche, il veut réestimer la moyenne μ et vous demande donc de l'aider :

Q 32.1 Dans un test d'hypothèses, quelles hypothèses H_0, H_1 formuleriez-vous pour tester s'il faut revoir les tarifs de l'assurance à la baisse ? Quelle serait la forme de la région critique ?

Q 32.2 Le statisticien s'interroge sur les conséquences qu'auraient le fait de rejeter H_0 alors que celle-ci est vraie. Cela entrainerait la réalisation de l'étude exhaustive pour rien, donc une dépense inutile, et porterait atteinte à la réputation du statisticien. Il décide donc de ne pas prendre de risque et de fixer la probabilité de commettre une telle erreur à $\alpha = 0,01$. Exprimez α en fonction de la région critique.

Q 32.3 Quelle est la loi suivie par $\overline{X}$ sous H_0 ?

Q 32.4 À partir de quelle valeur le test nous indique-t-il de rejeter H_0 ?

Q 32.5 Notre statisticien veut maintenant examiner la puissance de son test afin de voir si sa règle de décision est *solide*. Il réfléchit alors sur les conséquences de rejeter H_1 alors que H_1 est vraie. Si une telle erreur se produisait, les automobilistes n'obtiendraient pas une réduction du prix de la prime alors qu'il y auraient droit.

Si la diminution à laquelle ils avaient droit se chiffrait à 20 euros ou moins, on ne pourrait pas parler de conséquences sérieuses. à quel nombre moyen k de kilomètres parcourus correspond une baisse de 20 euros de l'assurance ?

Q 32.6 Calculez la puissance du test pour k. Cette puissance nous indique la probabilité que la règle de décision soit *fiable* pour une moyenne de k kilomètres, c'est-à-dire lorsque les assurés devraient commencer à percevoir une différence au niveau du prix de leur assurance. Est-ce que le statisticien peut procéder au recueil des données auprès des 400 personnes ou bien son test d'hypothèses n'est-il pas sûr ?

Exercice 33 – Sécurité sociale

On souhaite comparer l'efficacité de deux médicaments censés combattre la même maladie. Le premier médicament est générique et son prix est réduit, le deuxième est un médicament de marque de prix beaucoup plus élevé. La Sécurité Sociale a effectué une enquête sur les guérisons obtenues grâce à chacun de ces médicaments. Le nombre de guérisons et de non guérisons (sur les 250 personnes testées) sont consignés dans le tableau ci-dessous :

	générique	marque
guérisons	44	156
non guérisons	6	44

À un niveau de risque de 5%, peut-on estimer que le taux de guérison dépend du médicament (générique ou marque) ? Justifiez votre réponse mathématiquement.

Exercice 34 – À l'attaque !

L'autorité maritime d'un certain pays souhaite évaluer un logiciel qui analyse les données de navigation de vaisseaux (satellitaires, enregistrées dans les ports, etc.) pour détecter des attaques de piraterie. D'après son constructeur, une attaque est détectée dans 90% des cas. Malheureusement, il y a aussi 20% de chances que le logiciel identifie une attaque lorsqu'il n'y en a pas. Pour l'évaluation, l'autorité se concentre sur un trajet en particulier qui, dans la dernière année, a enregistré 200 attaques des pirates sur 4000 passages de vaisseaux. On note L la variable aléatoire pour la prédiction du logiciel et A pour l'attaque.

Q 34.1 Probabilité à posteriori Le logiciel signale une attaque en ce moment. En utilisant le nombre d'attaques observé dans la dernière année pour estimer la probabilité a priori, calculer la probabilité qu'il y ait effectivement une attaque. Écrire la formule correspondant à cette probabilité, puis calculer sa valeur.

Q 34.2 Information complémentaire L'autorité a à disposition des outils supplémentaires : depuis quelques années elle a mis en place un réseau d'observateurs permanents (choisi parmi des pécheurs et d'autres navires civils), dotés d'un appareil spécial pour notifier en temps réels l'occurrence de mouvements suspects. Selon une statistique interne, ce réseau a permis de reconnaitre 40% des attaques en avance. Malheureusement, cette méthode donne aussi 30% de faux positifs (mouvements suspects sans attaque). En supposant les deux notifications (logiciel et réseaux d'observateurs) indépendantes, calculer la probabilité qu'il y ait une attaque lorsque les deux notifications sont actives (utiliser R pour la variable aléatoire de prédiction venant du réseau).

Q 34.3 Recalibration ? Après une recherche qualitative, nous nous sommes aperçus que la statistique des attaques utilisée pour estimer la valeur des probabilités a priori était non optimale. En effet, les pirates n'attaquent pas tous les navires, mais principalement ceux qui ont un certain tonnage.
Supposons que les navires se distribuent en 3 classes (I, II, III). D'après les statistiques historiques, la probabilité d'attaque en fonction de la classe est la suivante : $P(\text{I}) = 0.1$, $P(\text{II}) = 0.5$ $P(\text{III}) = 0.4$. D'un autre coté, les données détaillées de l'année passée donne :

Classe de navire	I	II	III
Nombre d'attaques	40	120	40

Faire un test d'ajustement avec un niveau de confiance de 90% pour déterminer si les observations correspondent toujours à la distribution historique sur les classes de navires.

Semaine 6 - Chaînes de Markov

On considère des chaînes de Markov permettant de modéliser la météo. Une chaîne permet de modéliser la météo dans une ville. Un état d'une chaîne correspond au climat observé pour un jour donné (Soleil, Nuage ou Pluie) dans la ville. Chaque jour, on change d'état suivant la loi de probabilités de transitions associée à l'état courant. On prendra comme convention que l'état 1 correspond à Soleil, l'état 2 à Nuage, l'état 3 à Pluie.

Exercice 35 – Probabilité d'une séquence, génération aléatoire d'une séquence

On suppose que les paramètres de la chaîne de Markov pour Paris sont les suivants (dans l'ordre les observations sont S N P) :

- probabilités initiales : $\Pi = [0.2, 0.3, 0.5]$
- probabilités de transitions : $A = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$

Q 35.1 Calculez la probabilité de la séquence d'états suivante : N, N, S, N, N, P, P, N, P, S, S, P. Généralisez au cas quelconque d'une séquence.

Q 35.2 On souhaite utiliser la chaîne de Markov précédente pour générer aléatoirement une séquence de climats journaliers.
Pour cela, on utilise la procédure suivante : on considère une distribution de probabilités sur un ensemble fini d'événements $E = \{e_1, \ldots, e_N\}$ possibles. Cette distribution est donc définie par des probabilités associées aux événements $p(e_1), \ldots p(e_N)$, avec $\sum p(e_i) = 1$.
Pour tirer un événement au hasard « informatiquement » avec une distribution de ce type (tirage type *roulette*), on découpe le segment [0,1] en autant de tranches qu'il y a d'événements, la tranche correspondant à e_i ayant une largeur égale à $p(e_i)$. Ensuite, on utilise un générateur aléatoire uniforme entre 0 et 1, et on regarde dans quelle tranche on tombe. L'événement tiré aléatoirement est celui correspondant à la tranche dans laquelle on « tombe ». On utilise cette procédure pour tirer au hasard le premier état, puis la transition à partir de cet état, etc ... Les nombres donnés par le générateur aléatoire (entre 0 et 1) sont : $0.21, 0.63, 0.92, 0.87, 0.01, 0.35, 0.01, 0.43, 0.55$. Quelle est la séquence de climats journaliers générée avec ces tirages ?

Q 35.3 Quelle est la longueur moyenne d'une séquence consécutive de nuage avec ce modèle ?
Indice : l'espérance d'une loi géométrique de paramètre p est $1/p$.

Exercice 36 – Exemple de classification avec des CMs

On vous donne la séquence de climats journaliers suivante (S, S, P, P, N, S). On dispose également de deux chaînes de Markov, l'une correspondant au climat de Paris, l'autre au climat de Marseille. Les paramètres de ces deux chaînes sont les suivants :

$$\text{Paris} : \quad \Pi = [0.2, 0.3, 0.5] \quad A = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

$$\text{Marseille} : \quad \Pi = [0.5, 0.3, 0.2] \quad A = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

Q 36.1 En quelle ville (Paris ou Marseille) cette séquence a été observée ?

Q 36.2 On se place dans l'état initial *Soleil* : quelle est la distribution de probabilité des états au bout de 2 jours dans les deux villes ?

Exercice 37 – Apprentissage des paramètres d'une chaîne de Markov

On observe une séquence d'observations et on souhaite apprendre les paramètres de la chaîne de Markov qui a généré cette séquence d'observations. Soit la séquence de symboles suivante : P N S P N S P.

Q 37.1 Déterminez les fréquences d'apparition des symboles et celle des bigrammes (suite de deux symboles). En déduire les paramètres de la chaîne de Markov permettant de modéliser le processus sous jacent qui a généré la séquence précédente. Dressez la matrice de transition d'ordre 1.

Exercice 38 – Apprentissage des paramètres d'un mélange de chaînes de Markov

L'algorithme des K-Moyennes est une version simplifiée de l'algorithme EM pour la classification de données qui fonctionne comme suit pour des données vectorielles :

1. Initialiser k prototypes (par exemple, des individus tirés dans la base ou des tirages aléatoires)
2. Tant que *critère de convergence* non atteint
   — Affecter à chaque point de la base la classe du prototype le plus proche ($\sim$E)
   — Re-estimer les prototypes comme la moyenne des points composants chaque classe ($\sim$M)

On travaille maintenant des séquences d'observations (eg 1 séquence = 1 mouvement ou 1 séquence = 1 série de lancers de dé pipé) et on souhaite apprendre les paramètres d'un mélange de chaîne de Markov (k mouvements distincts ou k dés différents) qui a généré cet ensemble de séquences d'observations.

Q 38.1 En vous inspirant de l'algorithme des K-Moyennes imaginer une stratégie pour réaliser un tel apprentissage.

Q 38.2 Ecrire le code python correspondant.

Exercice 39 – Classes sociales (exercice de LI323)

D'éminents sociologues rangent les individus de notre société dans trois classes sociales : (B)ourgeoisie, (C)lasse moyenne et (P)rolétariat. On s'intéressera, dans ce modèle simpliste, à la classe sociale qu'atteint un individu à la fin de sa vie. On supposera que celle-ci dépend uniquement de la classe sociale de son père (et pas de celles de ses ancêtres).

— Si le père appartient à la classe sociale (B) son fils appartiendra aux classes (B), (C) avec des probabilités respectives 0.5 et 0.5.
— Si le père appartient à la classe sociale (C) son fils appartiendra aux classes (B), (C) et (P) avec des probabilités respectives 0.2, 0.7 et 0.1.
— Si le père appartient à la classe sociale (P) son fils appartiendra aux classes (B), (C) et (P) avec des probabilités respectives 0.1, 0.3 et 0.6.

Q 39.1 Montrer que ce comportement sociologique peut être modélisé par une chaîne de Markov (CM). Donner le graphe associé, ainsi que la matrice des transitions.

Q 39.2 Cette chaîne est-elle irréductible ? Apériodique ? Quelle est la nature des différents états de la CM ?

Q 39.3 Déterminer les probabilités stationnaires des trois classes sociales. Quelles sont, à (très) long terme, les proportions de gens de chacune des trois classes ?

Q 39.4 Quelle est la probabilité d'être d'une classe sociale différente de celle de son grand-père (paternel) ?

Exercice 40 – Météo (bis) (*FM Diener*)

Le magicien d'Oz a comblé tous les désirs des habitants du pays d'Oz, sauf peut-être en ce qui concerne le climat : au pays d'Oz en effet, s'il fait beau un jour, il est certain qu'il pleuvra ou neigera le lendemain, avec une probabilité égale qu'il pleuve ou qu'il neige. Et si le temps d'un jour est pluvieux ou neigeux, alors il reste

inchangé dans 50% des cas le lendemain et ne devient beau que dans 25% des cas. Les habitants se sont plaint auprès du magicien, affirmant que, ce faisant, ils n'ont qu'un beau jour sur cinq, ce à quoi il a répondu qu'il s'agit d'une impression mais qu'en réalité il y a bien plus d'un beau jour sur cinq. Qu'en est-il ? Pour le savoir, on se propose de modéliser l'évolution du climat au pays d'Oz par une chaîne de Markov à 3 états, $\{P, B, N\}$ (pour Pluvieux, Beau, et Neigeux).

Q 40.1 Donner la matrice de transition A de ce modèle.

Q 40.2 Donner un exemple de trajectoire de probabilité nulle.

Q 40.3 Donner la probabilité que le surlendemain d'un jour neigeux soit neigeux.

Q 40.4 Le carré de A vaut : $\begin{bmatrix} 0,438 & 0,188 & 0,375 \\ 0,375 & x & 0,375 \\ 0,375 & 0,188 & 0,438 \end{bmatrix}$

Q 40.4.1 Que représente A^2 concrètement ?

Q 40.4.2 Donner le coefficient x manquant.

Q 40.5 En calculant les puissances successive de A^k, on trouve une valeur stable à partir de $k = 6$:

$$A^{k>5} = \begin{bmatrix} 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

Q 40.5.1 En déduire la valeur de la distribution stationnaire μ

Q 40.5.2 En déduire la réponse à la question initiale sur le nombre de jours de beau temps.

Exercice 41 – Analyse du mental des joueurs de tennis

Nous proposons de modéliser un jeu de tennis comme un enchainement de points de l'un ou l'autre des joueurs. Nous voulons voir comment évolue la probabilité de remporter un point au cours d'un jeu. Proposer un modèle de Markov permettant d'analyser l'évolution des points dans un jeu.
Détailler les mécanismes d'apprentissage du modèle et expliquer la ou les approches qui vous semblent intéressantes : un modèle global, un modèle par joueur, plusieurs modèles par joueur...
Si le mental ne rentrait pas en ligne de compte, que devrait-on observer ?

Semaine 7 - Modèles de Markov Cachés

Rappel des notations utilisées pour les TD sur les Modèles de Markov Cachés

Modèle	
π_i	$p(s_1 = i\|\lambda)$
a_{ij}	$p(s_t = j\|s_{t-1} = i, \lambda)$
$b_i(x_t)$	$p(x_t\|s_t = i, \lambda)$

Forward	
$\alpha_t(i)$	$p(x_1^t, s_t = i\|\lambda)$
Backward	
$\delta_t(i)$	$\max_{s_1^{t-1}} p(s_1^{t-1}, s_t = i, x_1^t\|\lambda)$
$\Psi_t(j)$	$\underset{i\in[1,\ N]}{\operatorname{argmax}}\, \delta_{t-1}(i)a_{ij}$

Exercice 42 – Modélisation avec des HMM crédit : F. Galisson

Dans un casino aux pratiques douteuses, les croupiers utilisent le plus souvent des dés normaux mais introduisent parfois des dés pipés (dont la probabilité de faire 6 vaut 0.5 et la probabilité d'obtenir un autre score vaut 0.1).

La probabilité de passer à des dés pipés est de 0.2. Afin de ne pas trop attirer l'attention, la probabilité de revenir aux dés normaux est plus grande (une chance sur deux) et la probabilité d'utiliser le dé pipé initialement est de 0.1.

Q 42.1 Donner le HMM représentant ce processus.

Q 42.2 Est-il possible de modéliser ce système sans état caché ?

Q 42.3 Dans le cadre de la modélisation HMM, calculer les probabalités des évènements suivants :
- utiliser le dé non pipé puis deux fois le dé pipé,
- tirer la séquence 1, 2, 6 étant donnée la séquence d'états précédente,

Q 42.4 Un MMC est un modèle de génération aléatoire de séquences. On peut générer une séquence d'observations aléatoirement. On vous donne la séquence de nombres tirés aléatoirement avec un générateur aléatoire informatique (uniforme entre 0 et 1) : 0.1 0.55 0.45 0.3 0.95 0.23
Déterminez la séquence d'états et d'observations générées.
NB : En règle générale, on observe un phénomène (séquence de numéros de dés) mais on ne connaît pas la séquence d'états sous-jacente. C'est pourquoi on dit que ce sont des Modèles de Markov Cachés.

Q 42.5 Exploitation du modèle : calcul de la probabilité d'une séquence d'observations (méthode α, forward). Calculer la probabilité de la séquence 2, 6, 6, 6, 3.

Q 42.5.1 Définition : Rappeler l'interprétation des α
- *Définition* : $\alpha_t(j) = p(x_1^t, s_t = j|\lambda)$
  Rappeler l'interprétation des α

- *Initialisation* :
  $$\alpha_{t=1}(j) = p(x_1^1, s_1 = j|\lambda) = \pi_j b_j(x_1), \quad b_j \text{ correspond au modèle d'émission dans l'état } j$$
- *Recursion* – Exprimer $\alpha_t(j)$ en fonction des $\alpha_{t-1}(i)$ et des paramètres du modèle.
  - Arriver de n'importe quel état en $t-1$
  - Faire la transition vers j
  - Observer x_t
- *Terminaison* :
  Exprimer $p(x_1^T|\lambda)$ en fonction des α_T

Q 42.6 Comment évolue les α_t en fonction de t ? Etudier les variations de $\sum_i \alpha_t(i)$ entre 2 pas de temps. Qu'en déduire sur l'évaluation du tableau des α ?

Q 42.7 Pourquoi le passage au log n'est pas une solution ?

Q 42.8 Afin de résoudre le problème des approximations numériques, définissons
- $\alpha_t^\dagger(j) = p(x_t, s_t = j|x_1^{t-1}, \lambda)$
- $\Omega_t = p(x_t|x_1^{t-1}, \lambda)$

Exprimez Ω en fonction de $\alpha^\dagger$, puis la récursion en fonction des $\alpha^\dagger$ et Ω. Montrez alors comment calculer $p(x_1^T)$ en évitant les problèmes d'approximation numérique.

Q 42.9 Exploitation du modèle : décodage.
On cherche maintenant à calculer la séquence d'états la plus probable pour la séquence d'observations :

2 6 6 6 3

$$\delta_t(i) = \max_{s_1^{t-1}} p(s_1^{t-1}, s_t = i, x_1^t|\lambda)$$

1. Initialisation
$$\begin{aligned} \delta_1(i) &= \pi_i b_i(x_1) \\ \Psi_1(i) &= 0 \end{aligned}$$

2. Récursion
$$\begin{aligned} \delta_t(j) &= \left[\max_i \delta_{t-1}(i) a_{ij}\right] b_j(x_t) \\ \Psi_t(j) &= \operatorname*{Argmax}_{i \in [1,\ N]} \delta_{t-1}(i) a_{ij} \end{aligned}$$

3. Terminaison
$$S^\star = \max_i \delta_T(i)$$

4. Chemin
$$\begin{aligned} s_T^\star &= \arg\max_i \delta_T(i) \\ s_t^\star &= \Psi_{t+1}(s_{t+1}^\star) \end{aligned}$$

Q 42.9.1 On suppose que Δ est une matrice $N \times T$ où chaque ligne correspond à un état et chaque colonne à un pas de temps avec
$$\Delta_{it} = \log \delta_t(i) = \log \max_{s_1^{t-1}} p(s_1^{t-1}, s_t = i, x_1^t|\lambda)$$

On suppose que l'on a calculé les valeurs suivantes :

$$\Delta = \begin{pmatrix} -1.9 & -3.9 & -5.9 & -7.9 & -9.5 \\ -4.6 & -4.2 & -5.6 & -7 & -10 \end{pmatrix}, \begin{pmatrix} -0.2 & -1.6 \\ -0.7 & -0.7 \end{pmatrix}$$

et

$$\begin{pmatrix} 0 & 1 & 1 & ? & ? \\ 0 & 1 & 2 & ? & ? \end{pmatrix}$$

Compléter la matrice Ψ et donner la séquence d'états la plus probable

Q 42.9.2 Est-il possible de passer cet algorithme au log facilement ?

Q 42.10 Maximisation de la vraisemblance (connaissant les états). Etant donné un ensemble de séquences d'observations et l'ensemble associé des séquences d'états, proposer un algorithme de maximisation de la vraisemblance de $\lambda = \{\Pi, A, B\}$.

Q 42.11 Apprentissage (Baum-Welch simplifié). Vous disposez maintenant de deux méthodes :
— Décodage (estimation des états et de la probabilité d'une séquence d'observation)
— Maximisation de la vraisemblance connaissant les états
Proposer un algorithme simple (type k-means, avec affectation dure des états) pour apprendre itérativement un modèle de Markov caché. Comment définir un critère d'arrêt des itérations ?

Exercice 43 – Modélisation de séquences d'observations

On ne considère que des observations discrètes, i.e. appartenant à un ensemble fini Σ d'observations possibles. On considère un alphabet à 3 symboles $\Sigma = \{a, b, c\}$ et une base de données d'apprentissage constituée de 4 séquences $X = \{aaba, aabc, aaca, aacb\}$.

Q 43.1 Dessinez un modèle de Markov caché (en explicitant les probabilités de transition et les lois de probabilités d'émission) qui maximise la vraisemblance de X. Ce modèle est-il unique ? Que vaut la vraisemblance de chacune des séquences calculée par votre modèle ? Que vaut la vraisemblance de X calculée par votre modèle ?

Q 43.2 De manière générale, et en anticipant le prochain cours, donner des indications succinctes sur la façon de construire un MMC maximisant la vraisemblance sur un ensemble de séquences X quelconque.

Q 43.3 Expressivités et limites des MMC
On considère maintenant les ensembles de séquences $E_1 = \{a^*b\}$, $E_2 = \{(ab)^*\}$, $E_3 = \{(ab^*)^*\}$, $E_4 = \{a^nba^n, n \in \mathbb{N}\}$, où x^* représente l'ensemble des séquences constituées d'un nombre quelconque de répétitions de x, et x^n représente la séquence constituée de n répétitions de x.
On dit qu'un MMC accepte une séquence s particulière si la probabilité de s calculée par le MMC est non nulle. Peut-on construire un modèle de Markov (chaîne de Markov ou MMC) acceptant l'ensemble de séquences E_1 ? Si la réponse est oui, explicitez le MMC, sinon expliquez succinctement pourquoi.

Q 43.4 Idem pour E_2 ? Idem pour E_3 ? Idem pour E_4 ?

Semaine 8 - Modèles de Markov Cachés avancés

Exercice 44 – Synthèse sur les MMC

Q 44.1 Formalisation de l'apprentissage d'un MMC.
Généralement on apprend un MMC à partir d'une base de données d'apprentissage non étiquetée, c'est-à-dire constituée d'un ensemble de séquences d'observations, mais sans les séquences d'états associées. On commence par se placer dans ce cadre.

Q 44.1.1 On suppose que l'on dispose d'une base d'apprentissage d'une seule séquence d'observations $X = \{\mathbf{x}^{(1)}\}$. Quelle propriété satisfait le modèle λ qui maximise la vraisemblance des données d'apprentissage ? Quel algorithme utiliser pour faire l'apprentissage ?

Q 44.1.2 On suppose que l'on dispose d'une base d'apprentissage de N séquences $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)} ..., \mathbf{x}^{(N)}\}$. Quelle propriété satisfait le modèle λ qui maximise la vraisemblance des données d'apprentissage ?

Q 44.1.3 On considère maintenant le cas d'une base de données d'apprentissage étiquetée, c'est-à-dire constituée d'un ensemble de couples (séquence d'observations, séquence d'états). On suppose que l'on dispose d'une base d'apprentissage étiquetée de N séquences $XS = \{(\mathbf{x}^{(1)}, \mathbf{s}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{s}^{(2)}), ..., (\mathbf{x}^{(N)}, \mathbf{s}^{(N)})\}$. Quelle propriété satisfait le modèle λ qui maximise la vraisemblance des données d'apprentissage ?

Q 44.2 Difficulté de l'apprentissage d'un MMC.
On considère le cas d'une base de données d'apprentissage non étiquetée. On vous fournit la séquence d'observations $\mathbf{x} = (1, 2, 1, 1, 3, 2)$ produite par un modèle Markovien, mais on ne vous dit pas par quel type de modèle (nombre d'états etc) cette séquence a été produite, ni la séquence d'états correspondante.

Q 44.2.1 Quel modèle Markovien maximise la vraisemblance de la séquence $\mathbf{x}$ (nombre d'états, lois de probabilité de transitions et d'émission) ? Quel est son pouvoir de généralisation ?

Q 44.2.2 En supposant que la séquence a été générée par un modèle MMC à 1 état, quels sont les paramètres de ce modèle ?

Q 44.2.3 On suppose que cette séquence a été générée par un MMC à deux états. Proposez des paramètres pour ce modèle. Pouvez-vous prouver que votre modèle est localement optimal ? Vous commencerez par définir ce que signifie localement optimal.

Q 44.3 Apprentissage en présence de données étiquetées.
On change de cadre maintenant et on suppose que l'on vous fournit comme corpus d'apprentissage des données étiquetées, c'est-à-dire un ensemble de couples (séquence d'observations, séquence d'états). On considère une base d'apprentissage constituée d'une séquence $XS = \{(\mathbf{x} = (1, 2, 1, 1, 3, 2), \mathbf{s} = (1, 1, 2, 2, 1, 2))\}$ et on vous demande le MMC qui maximise la vraisemblance de cette base d'apprentissage.

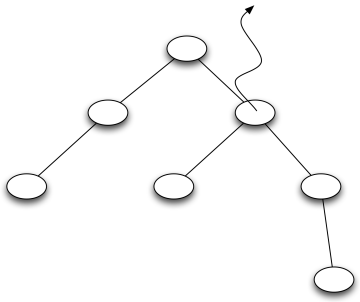
Q 44.3.1 Quel est le nombre d'états du MMC ?

Q 44.3.2 Quels sont les paramètres du modèle optimal ? Pouvez-vous démontrer son optimalité ?
On généralise maintenant en considérant une base d'apprentissage étiquetée $XS = \{(\mathbf{x}^1, \mathbf{s}^1), \ldots, (\mathbf{x}^N, \mathbf{s}^N)\}$.

Q 44.3.3 Comment trouve-t-on le nombre d'états du modèle optimal ?

Q 44.4 Dans les stratégies d'apprentissage des MMC, quelle est la différence entre l'algorithme Baum-Welch simplifié et la version complète ? Sur quel variable intermédiaire repose la version complète ?

Exercice 45 – Viterbi dans les arbres binaires



Dans cet exercice, nous allons ajouter une dimension à l'algorithme de Viterbi pour l'appliquer dans les arbres. L'idée est de caractériser une structure arborescente comme une page HTML, un document XML. En analyse linguistique, les phrases sont également transformées en arbres. L'idée est donc de concevoir un algorithme de type HMM capable de modéliser des états cachés au niveaux des noeuds de l'arbre, les observations étant des émissions depuis ces états. Les applications associées peuvent être la détection et le blocage de publicité dans les pages web, la classification des noeuds dans l'analyse d'une phrase. Ce type d'approche est également utilisé en image : en découpant un image en région puis en construisant un arbre entre ces régions. Dans tous ces exemples, le but est de caractériser les enchainements de noeuds (ie les transitions du processus markovien) pour améliorer la classification des noeuds.

Cependant l'approche *forward-backward* de Viterbi n'est pas applicable directement. En particulier, le backtracking dans le tableau ψ pose problème avec les embranchements [1].

Nous allons donc généraliser le forward et le backward (Viterbi) aux arbres en utilisant les notations suivantes :

- n est un nœud de l'arbre, son état est s_n et x_n est l'observation
- Ses enfants sont $\underline{n}$, leurs états sont $s_{\underline{n}} = \{s_c\}_{c\in\underline{n}}$ et les observations associées $x_{\underline{n}} = \{x_c\}_{c\in\underline{n}}$
- Ses descendants sont $\underline{\underline{n}}$, leurs états sont $s_{\underline{\underline{n}}}$ et les observations associées $x_{\underline{\underline{n}}}$
- r est la racine de l'arbre

Les probabilités de transition, d'état initial et d'observation sont les mêmes que pour les MMC : l'état d'un enfant ne dépend que de celui du parent et peut donc être décrit par la même matrice de transition A.

Q 45.1 Forward

Calculer la probabilité $\alpha'_n(i) = p(x_n, x_{\underline{\underline{n}}}|s_n = i)$ de l'ensemble des observations $x_{\underline{\underline{n}}}$ des descendants du nœud n ainsi que lui même, et en déduire la probabilité $p(x_{\underline{\underline{r}}})$ de l'ensemble des observations. Comparez avec les résultats obtenus pour le HMM. Dire pourquoi cela pose problème numériquement et proposer une solution.

Q 45.2 Viterbi dans les arbres binaires.

Calculer les états les plus probables $s^\star_{\underline{\underline{r}}}$ sachant les observations. Comparez à ce que vous aviez trouvé dans le cas du HMM.

Semaine 9 - Monte-Carlo par chaînes de Markov (MCMC)

Exercice 46 – Rejection sampling

Supposons qu'un phénomène réel peut être modélisé par une variable aléatoire $X \in [0, 1]$ qui suit une loi normale tronquée proportionnelle à $\mathcal{N}(\frac{3}{4}, 1)$. La fonction de densité d'une loi normale tronquée proportionnelle à $\mathcal{N}(\mu, \sigma^2)$ définie sur $[a, b]$ peut être écrite comme :

$$f(x) = \begin{cases} C \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} & \text{pour } a \le x \le b \\ 0 & \text{pour } x < a \text{ et } x > b \end{cases}$$

où C est un facteur de normalisation.

Q 46.1 Déterminer la fonction de densité $f(x)$ de X en calculant C.

Note : utiliser la table de la loi normale.

Q 46.2 Imaginons que la fonction $f(x)$ est difficile à échantillonner. Nous allons donc utiliser la méthode Monte Carlo appelée *rejection sampling*.

1. Rappel : dans l'algorithme sur les séquences, $\psi_t(i)$ contient la valeur de l'état $t-1$ si on était en i à l'instant t. Ce n'est pas applicable dans les arbres où il y a potentiellement plusieurs instants t pour un $t-1$.

Rappel de cours : Choisir une distribution $q(\cdot)$, facile à échantillonner, telle qu'il existe un facteur k satisfaisant $\forall x, k \cdot q(x) \geq f(x)$. L'algorithme d'échantillonnage est constitué de quatre étapes : (1) tirer un nombre z selon $q(\cdot)$ (*pre-échantillonnage*) ; (2) calculer $m_q = k \cdot q(z)$; (3) tirer un nombre u selon la distribution uniforme sur $[0, m_q]$; (4) accepter z comme échantillon si $u \leq f(x)$.

Calculez le *taux d'acceptation* (la proportion de pre-échantillons acceptés) lorsque $q(\cdot)$ est une loi uniforme sur $[0, 1]$, et $k \cdot q(\cdot) = \max_{x \in [0,1]} f(x)$.

Q 46.3 Supposons que le calcul de $f(x)$ est relativement coûteux. Une méthode pour augmenter l'efficacité (appelée *compression*) est de proposer une fonction $r(x)$ simple—e.g. une droite—qui est une limite inférieure de $f(x)$. L'algorithme modifié prend en compte un *pré-filtrage* des éléments qui sont certainement acceptés, quand $u \leq r(z)$. Calculez le taux des préfiltrage si on prend $r(x) = \min_{x \in [0,1]} f(x)$.

Exercice 47 – Estimation de π par une méthode de Monte Carlo

Les méthodes de Monte Carlo permettent de calculer de manière approchée des intégrales (ou d'estimer des espérances). Pour comprendre comment elles fonctionnent, nous allons les appliquer à un exemple très simple (pour lequel une méthode de Monte Carlo n'est clairement pas la méthode la plus efficace). Ces méthodes de Monte Carlo se révèlent en fait particulièrement utiles et performantes en grande dimension quand le calcul analytique ou les techniques numériques classiques ne sont plus possibles.

Dans cet exercice, on cherche à estimer π par une méthode de Monte Carlo. Dans $\mathbb{R}^2$, considérons le cercle de rayon 1 et le carré $[-1, 1]^2$.

Q 47.1 Donner les aires du cercle et du carré.

L'aire du cercle peut être calculée par l'intégrale suivante :

$$\pi = \int_{-1}^{1} \int_{-1}^{1} \mathbb{I}(\sqrt{x^2 + y^2} \leq 1) dx dy$$

où $\mathbb{I}(\sqrt{x^2 + y^2} \leq 1)$ est la fonction indicatrice qui vaut 1 si le point de coordonnées (x, y) est à une distance inférieure ou égale à 1 de l'origine.

Q 47.2 Soit la loi uniforme $\mathcal{U}([-1, 1])$ définie sur l'intervalle $[-1, 1]$. Quelle est la fonction de densité $u(\cdot) : [-1, 1] \mapsto \mathbb{R}$ de cette distribution de probabilité ? De même, quelle est la fonction de densité de la loi uniforme $\mathcal{U}([-1, 1]^2)$ définie sur le carré $[-1, 1] \times [-1, 1]$?

Q 47.3 En observant que $\mathbb{I}(\sqrt{x^2 + y^2} \leq 1) = \mathbb{I}(x^2 + y^2 \leq 1)$, montrer que cette aire peut se réécrire comme une espérance en introduisant la densité de loi uniforme $\mathcal{U}([-1, 1]^2)$:

$$\pi = 4 \times \mathbb{E}_{X,Y}\big(\mathbb{I}(X^2 + Y^2 \leq 1)\big)$$

où X, Y sont deux variables aléatoires de loi uniforme $\mathcal{U}([-1, 1])$.

Q 47.4 La valeur π peut donc être estimée par une méthode de Monte Carlo. Expliquer le principe de l'algorithme. Quelle loi justifie cette approche ?

Un des problèmes quand on applique une méthode de Monte Carlo est de savoir quelle valeur choisir pour N, le nombre de tirages aléatoires. La technique suivante, fondée sur une estimation de l'erreur, peut nous aider à choisir approximativement N.
Nous cherchons à estimer pour un certain niveau de confiance l'erreur qui peut être faite par une procédure de Monte Carlo. Notons ε la variable aléatoire définie par :

$$\varepsilon = 4 \times \mathbb{I}(X^2 + Y^2 \leq 1) - \pi$$

où $X \sim U([-1,1]), Y \sim U([-1,1])$.

Q 47.5 Quelle interprétation peut-on donner à ε ? Donner l'expression de l'erreur $\overline{\varepsilon}_N$ (en tant que variable aléatoire) faite par la méthode de Monte Carlo précédente et son espérance.

Q 47.6 Calculer la variance σ^2 de ε.

Q 47.7 En supposant connue la variance σ^2 de ε, vers quelle loi tend $\overline{\varepsilon}_N$? Que ferait-on si σ^2 n'était pas calculable ?

Q 47.8 En exploitant la loi suivie par $\overline{\varepsilon}_N$, déterminez l'intervalle $[-\varepsilon^*, \varepsilon^*]$ pour lequel la probabilité que $\overline{\varepsilon}_N$ soit inférieur à $-\varepsilon^*$ ($P(\overline{\varepsilon}_N \leq -\varepsilon^*)$) ou bien supérieur à ε^* ($P(\overline{\varepsilon}_N \geq \varepsilon^*)$) est égale à un nombre α donné.

Q 47.9 Déduisez-en la valeur de N minimale pour que ε^* soit inférieur à une valeur ε_α fixée (par exemple $\varepsilon_\alpha = 10^{-3}$). L'interprétation de tous ces calculs est la suivante :

1. dans la question précédente, on a choisi un risque d'erreur de 2α, c'est-à-dire que l'on a 2α chances de générer une chaîne de Markov pour laquelle l'erreur entre notre estimation de π et la véritable valeur de π est supérieure à ε^*. En pratique, on se fixe donc un α très petit, souvent de l'ordre de 10^{-2}, de manière à avoir presque la certitude que notre chaîne génère une erreur d'estimation inférieure à ε^*.
2. dans la question actuelle, on fixe en outre l'erreur maximale $\varepsilon^* = \varepsilon_\alpha$ que l'on est prêt à accepter et l'on calcule donc la valeur de N qui nous permet d'assurer cela. Ici, on va donc choisir une valeur de ε_α très petite (elle est souvent plus petite que la valeur de α).
3. en combinant ces deux propriétés, on en déduit que, pour la valeur de N calculée, on a une chance de 2α d'obtenir une chaîne dont l'erreur d'estimation est supérieure à ε_α.

Exercice 48 – Estimation de π par MCMC

Pour appliquer une méthode de Monte Carlo, il est nécessaire de savoir échantillonner selon une certaine loi. Dans l'exercice précédent, on échantillonnait directement selon une loi uniforme. Dans certains problèmes, l'échantillonnage direct peut ne pas être applicable, soit parce qu'on utilise des lois plus complexes pour lesquelles un échantillonnage est difficile voire impossible, soit parce qu'on est en très grandes dimensions. En cours, vous avez vu la méthode d'échantillonnage par rejet. Elle peut se révéler efficace en petites dimensions. Cependant, en très grandes dimensions, il est nécessaire d'échantillonner à l'aide d'une chaîne de Markov en utilisant l'idée suivante.
Soit $S = S_1 \times \cdots \times S_k$ l'ensemble dans lequel on souhaite échantillonner selon la distribution de probabilité $\mathcal{P}$. Supposons qu'on ait une chaîne de Markov dont les états sont S et qui admette une distribution stationnaire. Alors, en partant d'un état initial quelconque, la distribution de l'état tend vers $\mathcal{P}$ après un grand nombre de transitions. Donc, quand on décide de s'arrêter, l'état courant correspond à un tirage selon la distribution voulue $\mathcal{P}$. Une méthode de Monte Carlo où on échantillonne grâce à une chaîne de Markov est appelée méthode de Monte Carlo par chaîne de Markov (MCMC).
Nous allons illustrer le principe d'une méthode MCMC pour estimer π.

Q 48.1 Quel est l'ensemble des états de la chaîne de Markov que nous pourrions utiliser ?

Bien qu'en TME, on fera l'hypothèse que cela ne pose aucun problème technique de travailler avec des chaînes de Markov avec un espace continu d'états, en TD, pour comprendre l'algorithme MCMC avec les outils que nous connaissons, nous allons discrétiser les états pour éviter les problèmes techniques dus à l'espace continu d'états. En prenant une discrétisation suffisamment fine, on pourrait "simuler" le cas de l'espace continu.
On découpe le carré $[-1,1]^2$ en une grille composée de cases de côté ε (on suppose que $2/\varepsilon \in \mathbb{N}$). Cette grille constituera l'espace d'états de notre chaîne de Markov. Un état est caractérisé par les coordonnées du centre de la case. L'idée est d'échantillonner une case de la grille de manière uniforme et de tester si la case est à l'intérieur du cercle ou non. Ce test peut se faire en vérifiant que le centre de la case tirée est à distance inférieure de 1 de l'origine.
Depuis un état donné (case), une transition possible correspond à un déplacement aléatoire $(\delta_x, \delta_y) \in [-m, m]^2$ où $m \in [0,1]$. La transition est acceptée si $(x + \delta_x, y + \delta_y) \in [-1,1]^2$ et on détermine alors la case d'arrivée qui

constitue l'état suivant. Si la transition n'est pas acceptée, l'état ne change pas.

Q 48.2 En partant d'une case centrale (où toutes les transitions seraient acceptées), quelle condition sur m garantit une distribution uniforme sur les cases accessibles ?

Par la suite, on suppose que ce choix de m est fait.

Q 48.3 En partant d'une case centrale, quelle distribution de probabilité définit-on sur les cases accessibles ?

Q 48.4 Quelle est la distribution de probabilité de transition quand on est proche du bord du carré $[-1,1]^2$?

Q 48.5 La chaîne de Markov ainsi obtenue est-elle irréductible ? Déduisez-en que la chaîne admet une distribution de probabilité stationnaire.

Une succession de telles transitions amène à un tirage aléatoire d'une case dans le carré $[-1,1]^2$. Intuitivement, si on se déplace aléatoirement pendant suffisamment longtemps, on a une probabilité égale de se retrouver dans n'importe quelle case quelle que soit la case de départ.

Q 48.6 Montrez que la distribution uniforme sur les états est la distribution stationnaire de cette chaîne de Markov.

Replaçons-nous dans le cas général où l'ensemble d'états est l'ensemble des points du carré $[-1,1]^2$. Une des difficultés des méthodes MCMC est de savoir quand on est proche de la distribution stationnaire (c'est-à-dire combien de transitions faut-il faire ?). Dans ce problème simple, il est possible de donner une borne inférieure au nombre de transitions nécessaires avant convergence vers la distribution uniforme.

Q 48.7 Quel est la distance moyenne d'un déplacement (en supposant qu'il est accepté) ?

Q 48.8 À partir d'un point (x_0, y_0), donnez une borne inférieure en espérance sur le nombre de transitions avant convergence vers la distribution stationnaire. *Indication :* considérez la distance minimale à parcourir pour garantir que tous les points soient atteignables à partir d'un état courant (x_0, y_0).

Exercice 49 – Décodage par la méthode de Métropolis-Hastings

Dans cet exercice, nous allons appliquer une méthode de type MCMC, l'algorithme de Métropolis-Hastings, au problème de décodage d'un texte codé par substitution. Nous supposons la langue du texte connue. Nous supposons également que nous avons à notre disposition une modélisation de cette langue sous forme de bigramme : plus formellement, cette langue s'écrit avec l'alphabet fini Λ. Par exemple, en français, Λ contient les lettres minuscules et majuscules, les lettres accentuées, les signes de ponctuation, les chiffres, etc... Le modèle bigramme est donné par μ et M où μ est une distribution de probabilité sur Λ et M est une matrice stochastique qui donne pour chaque lettre de Λ la probabilité de la lettre suivante. Ce modèle peut facilement être estimé à partir d'un grand corpus de texte. Une fonction d'encodage (ou de décodage) par substitution est une fonction bijective τ de Λ dans Λ. Si T' est un texte, le texte encodée $T = \tau(T')$ est obtenu en remplaçant chaque lettre c de T' par $\tau(c)$.
Le problème que nous souhaitons résoudre ici est, étant donné un texte encodé $T = (c_1, c_2, \ldots, c_{|T|})$ (où $c_i \in \Lambda$, $\forall i$), de retrouver le texte initial décodé.

Q 49.1 Comment peut-on mesurer la vraisemblance d'une fonction d'encodage en utilisant le modèle bigramme ?

Une méthode de type Monte Carlo pour résoudre ce problème serait de tirer au hasard une fonction d'encodage avec une probabilité proportionnelle à sa vraisemblance (c'est-à-dire $\mathcal{P}(\tau) = L(\tau(T), \mu, M)/(\sum_{\tau'} L(\tau'(T), \mu, M))$) et de répéter un grand nombre de fois cette opération en gardant le texte décodé le plus vraisemblable.

Q 49.2 Combien y-a-t-il de fonctions d'encodage ? Est-ce qu'une méthode de Monte Carlo est réalisable ici ?

Comme il n'est pas possible d'échantillonner directement une fonction d'encodage selon la loi $\mathcal{P}$, on souhaite recourir à un échantillonnage par chaîne de Markov. Cette méthode ne nécessite de connaître les probabilités

de tirage qu'à un facteur de normalisation près, ce qui est le cas ici.

Q 49.3 Définir une chaîne de Markov (sans donner les distributions de probabilité) qui nous permettrait de réaliser cette échantillonnage.

Étant donné la matrice stochastique A de transition de la chaîne de Markov, la méthode MCMC de Métropolis-Hastings se définit comme suit :
Répéter N fois les étapes suivantes à partir d'un état initial τ_0 choisi de manière quelconque :

- Calculer τ à partir de τ_t en échangeant deux lettres c_1, c_2 dans l'encodage, c'est-à-dire $\tau(c) = \tau_t(c)$ pour $c \neq c_1$ et $c \neq c_2$, $\tau(c_1) = \tau_t(c_2)$ et $\tau(c_2) = \tau_t(c_1)$.
- Accepter la transition de τ_t vers τ avec la probabilité $\alpha(\tau_t, \tau) = \min(1, \frac{L(\tau(T),\mu,M)A(\tau_t,\tau)}{L(\tau_t(T),\mu,M)A(\tau,\tau_t)})$ et $\tau_{t+1} = \tau$, sinon, $\tau_{t+1} = \tau_t$.

où $A(\tau, \tau')$ est la probabilité de la transition vers τ' depuis τ.
Après avoir itéré un nombre suffisamment grand de fois, la fonction de décodage τ_N correspond à un tirage aléatoire selon $\mathcal{P}$. Pour obtenir d'autres tirages selon $\mathcal{P}$, on répète ces opérations en gardant les τ_{N+kh} pour un entier $k > 0$ fixé et $h \in \mathbb{N}^*$. Le paramètre k permet d'espacer les tirages pour éviter les auto-corrélations dans l'échantillonnage.

Q 49.4 Montrer que le log de la probabilité d'acceptation s'écrit finalement :

$$\log \alpha(\tau_t, \tau) = \min(0, \mu(\tau(c_1)) + \sum_{i=1}^{|T|} \log M(\tau(c_{i-1}), \tau(c_i)) - \mu(\tau_t(c_1)) - \sum_{i=1}^{|T|} \log M(\tau_t(c_{i-1}), \tau_t(c_i)))$$

Q 49.5 Montrer que la distribution de probabilité $\mathcal{P}$ est bien la distribution stationnaire de la chaîne de Markov.

Semaine 10 - Regression

Exercice 50 – Régression simple et indicateurs statistiques

Une entreprise veut analyser ses coûts de production de son produit principal et en particulier les décomposer en coûts fixes et coûts variables et vérifier si ceux-ci sont, ou non, proportionnels aux quantités produites.

Elle postule donc un modèle linéaire $Y = \alpha + \beta X + \varepsilon$ où : X est la quantité produite (en milliers d'unités) ; Y le coût de production total (en milliers d'euros) ; β est le coût marginal de production (= coût nécessaire pour produire une unité supplémentaire) ; α représente les coûts fixes ; et ε est le résidu aléatoire.

Il dispose de données sur les $n = 10$ derniers mois :

$mois_i$	1	2	3	4	5	6	7	8	9	10
x_i	100	125	175	200	500	300	250	400	475	425
y_i	2 000	2 500	2 500	3 000	7 500	4 500	4 000	5 000	6 500	6 000

Q 50.1 Calculer les moyennes empiriques $\bar{x}$ et $\bar{y}$, les écarts-types empiriques s_x et s_y, la covariance empirique $cov(x, y)$ et le coefficient de corrélation linéaire r.

Q 50.2 Retrouver les expressions de α et β en calculant l'espérance de Y puis la covariance de X, Y.
Estimer a et b en fonction de ces expressions. Exprimer b en fonction de r.

Exercice 51 – Régression linéaire

Q 51.1 Régression linéaire 1D
Nous disposons d'un ensemble de N données $\{(x_i, y_i)_{i=1,\dots,N}\}$ à partir duquel nous souhaitons apprendre une droite de régression de Y sur X. Notre estimateur aura donc la forme suivante : $\hat{y}_i = f(x_i) = ax_i + b$. Notre but est de trouver les meilleurs coefficients a et b.
Pour une droite donnée l'erreur de régression cumulée au sens des moindres carrés est déterminée par $\sum_i e_i^2$ où $e_i = f(x_i) - y_i$, et $f(x_i)$ est l'ordonnée du point d'abscisse x_i.
Notons X la matrice $N \times 2$ des entrées avec ajout d'une colonne de termes constants : $X = \begin{bmatrix} x_1 & 1 \\ \vdots & \\ x_N & 1 \end{bmatrix}$ et $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$. Notons $D = \begin{bmatrix} a \\ b \end{bmatrix}$ le vecteur des paramètres de la droite de régression.

Q 51.1.1 Montrer que l'ensemble des estimations pour les entrées X peut être calculé matriciellement en utilisant la formule suivante : $\hat{Y} = XD$ (vérifier les dimensions et détailler le calcul d'une ligne).

Q 51.1.2 Montrer que l'erreur cumulée est calculée matriciellement en utilisant la formule suivante : $E = (XD - Y)^t(XD - Y)$
NB : dans un premier temps, détailler les dimensions de chaque matrice et calculer sur la dimension de E. Développer ensuite la formulation A^tA à l'aide d'une somme pour revenir à la formulation classique de l'erreur.

Q 51.1.3 Une fois le critère d'erreur E établi, quel problème d'optimisation devons nous résoudre pour trouver la droite de régression optimale ?
NB : nous sommes dans un cadre convexe : la fonction E de paramètres D admet un seul optimum global qui est un minimum. Rappeler la manière de trouver un optimum.

Q 51.1.4 Montrer que la dérivée de l'erreur, par rapport à D, s'écrit sous la forme matricielle suivante : $\nabla_D E = 2X^t(XD - Y)$
NB : détailler le calcul de chaque dérivée partielle et refactoriser pour obtenir la forme matricielle.

Q 51.1.5 Calculer les paramètres optimaux en résolvant analytiquement le problème sous forme matricielle.

Q 51.1.6 Simplifions temporairement le problème en considérant un biais nul. Quelle est la forme de la fonction $E(a)$? Tracer sommairement $E(a)$. Quelles sont les propriétés de $E(a)$ (combien de minimum...) ?

Q 51.2 Algorithme itératif pour la régression linéaire
Dans le cas général, on cherche à optimiser une fonction continue dérivable $C(W)$ d'un vecteur de paramètres W. Pour résoudre un tel problème, on peut utiliser un algorithme de gradient comme celui proposé ci-dessous :

Initialisation des W;
$t = 1$;
repeat
  $W_{t+1} = W_t - \varepsilon \frac{\partial C}{\partial W}$;
  $t = t + 1$;
until *($C(W)$ n'évolue plus)*;

Algorithm 1: Descente de gradient

Q 51.2.1 Quel est l'intérêt d'utiliser un algorithme itératif (dont la solution est une approximation du point

optimal) alors que nous disposons d'une solution analytique ?

Q 51.2.2 Adapter l'algorithme de descente de gradient pour la régression linéaire.

Q 51.2.3 L'algorithme est initialisé avec les paramètres suivants : $D^0 = (b^0, a^0)$. Si a^0 est plus grand que le $a^\star$ optimal, le gradient $\frac{\partial E}{\partial a}$ est-il positif ou négatif ? Idem si a^0 est plus petit que le a optimal.
Ces résultats vous semblent-ils cohérents avec l'algorithme de descente de gradient ?

Q 51.2.4 Même question avec b^0.

Q 51.2.5 Imaginez ce qui se passe si l'on optimise uniquement par rapport à a, en supposant que la valeur optimale de b est connue, pour différentes valeurs d'ε (valeur très grande, valeur très petite) : l'algorithme précédent peut-il diverger ou converge-t-il toujours vers la bonne solution ?

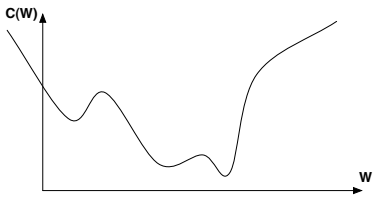


FIGURE 1 – Exemple de fonction coût $C(W)$ non-convexe

Q 51.2.6 Nous avons utilisé jusqu'ici des estimateurs linéaires. Donner un exemple d'estimateur non linéaire pour la régression 1D. Dans le cas non-linéaire, la fonction $C(W)$ est parfois non convexe (cf figure 1).

Q 51.2.7 Que pensez-vous de l'algorithme de gradient discuté précédemment dans ce cas ? L'algorithme de gradient converge-t-il toujours ? vers la solution optimale ?

Exercice supplémentaire

Exercice 52 – Approche discriminante : régression logistique

Jusqu'ici, nous avons toujours travaillé sur le critère de la vraisemblance selon le schéma :

1. Modélisation probabiliste d'une situation = 1 classe de données (chiffres manuscrits, mouvements du stylo sur des lettres...), paramètre θ
2. Optimisation des θ = trouver $\theta^\star$ maximisant la vraisemblance

Pourtant, ce type d'approche présente une faiblesse évidente dans les problèmes de classification : les classes sont apprises de manière isolées et on ne peut pas se focaliser sur l'information discriminante (ce qui distingue une classe d'une autre). Une autre classe de modèles permet de palier cette faiblesse : sur les données multi-variées, il s'agit de la régression logistique (ou classifieur de maximum d'entropie).
Nous notons les observations $\mathbf{x}_i \in \mathbb{R}^d$ et les étiquettes binaires associées $y_i \in \mathcal{Y} = \{0, 1\}$. Nous faisons l'hypothèse que les couples $(\mathbf{x}_i, y_i)$ sont tirés de manière i.i.d. et suivent une loi inconnue $P(X, Y)$.

Q 52.1 Formulation discriminante : afin de se focaliser sur ce qui distingue une classe de données d'une autre, nous modélisons directement $p(Y = 1|X = \mathbf{x})$.

Q 52.1.1 Est-il possible d'utiliser la fonction paramétrique f définie ci-dessous pour modéliser cette probabilité a posteriori ?

$$f(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{x}\mathbf{w} + b))}, \quad \mathbf{x},\ \mathbf{w} \text{ respectivement en ligne et colonne}$$

Q 52.1.2 Identifier les paramètres à apprendre et leurs dimensions respectives.

Q 52.1.3 En déduire une règle d'affectation à une classe pour un exemple $\mathbf{x}$.

Q 52.1.4 Quelle est la forme de la frontière de séparation des classes ? **Q 52.1.5** Par exemple, dans le cas où $d = 2$, $\mathbf{w} = [-2\ 1]$ et $b = 1$, représenter graphiquement la frontière de décision.

Q 52.2 Soit un ensemble d'apprentissage étiqueté $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. Après avoir identifié la loi suivie par Y, exprimer la vraisemblance jointe d'un couple $(\mathbf{x}_i, y_i)$ en fonction de $p(X = \mathbf{x}_i)$, $f(\mathbf{x}_i)$ et y_i.

Q 52.3 Rappeler l'hypothèse faite sur le tirage des couples et exprimer la vraisemblance jointe de l'ensemble de l'échantillon.
Passer au log. et simplifier la formulation du maximum de vraisemblance en expliquant comment supprimer le terme en $P(X = \mathbf{x}_i)$.

Q 52.4 Donner l'expression de $\frac{\partial L_{\log}}{\partial w_j}$, pour $j = 1, \ldots, d$ ainsi que l'expression de $\frac{\partial L}{\partial b}$.

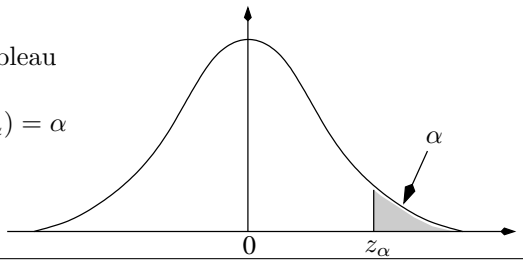
Q 52.5 Le gradient de $L_{\log}$ peut-il s'annuler directement ? Proposer une équation de mise à jour (type gradient) permettant de produire une suite de paramètres menant à un maximum de vraisemblance.

Q 52.6 En considérant l'ensemble d'apprentissage $S = \left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, 0 \right) \right\}$ les valeurs initiales $w^0 = [0\ 1]^T$ et $b^0 = -1$ et un pas d'apprentissage fixe $\varepsilon = 0.3$, faire deux itérations des algorithmes d'apprentissage proposés.

Tables des lois usuelles

Table de la loi normale centrée réduite

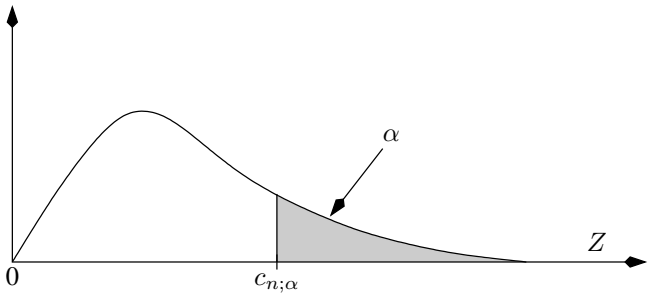
valeurs dans le tableau ci-dessous : les α tels que $P(Z > z_\alpha) = \alpha$



z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0859	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0466	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0063
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0047
2,6	0,0047	0,0045	0,0044	0,0043	0,0042	0,0041	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0014	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
3,6	0,0002	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,7	0,0001	0,0001	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Table de la loi du χ^2

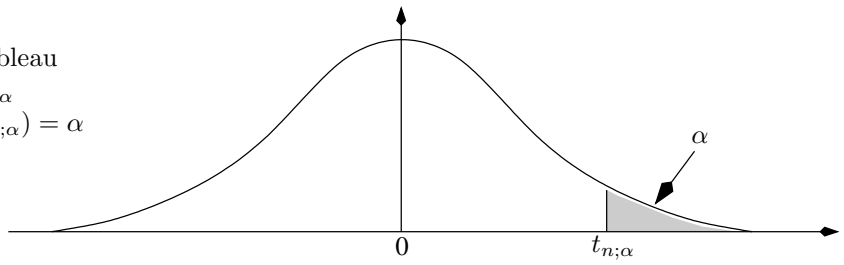
valeurs dans le tableau ci-dessous : les $c_{n;\alpha}$ tels que $P(Z > c_{n;\alpha}) = \alpha$



$n \setminus \alpha$	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,0000393	0,000157	0,000982	0,00393	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2
11	2,60	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2	28,3
13	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8
14	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8
16	5,14	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0	34,3
17	5,70	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4	35,7
18	6,26	7,01	8,23	9,39	10,9	26,0	28,9	31,5	34,8	37,2
19	6,84	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2	38,6
20	7,43	8,26	9,59	10,9	12,4	28,4	31,4	34,2	37,6	40,0
21	8,03	8,90	10,3	11,6	13,2	29,6	32,7	35,5	38,9	41,4
22	8,64	9,54	11,0	12,3	14,0	30,8	33,9	36,8	40,3	42,8
23	9,26	10,2	11,7	13,1	14,8	32,0	35,2	38,1	41,6	44,2
24	9,89	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43,0	45,6
25	10,5	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3	46,9
26	11,2	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6	48,3
27	11,8	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47,0	49,6
28	12,5	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3	51,0
29	13,1	14,3	16,0	17,7	19,8	39,1	42,6	45,7	49,6	52,3
30	13,8	15,0	16,8	18,5	20,6	40,3	43,8	47,0	50,9	53,7

Table de la loi de Student

valeurs dans le tableau ci-dessous : les $t_{n;\alpha}$ tels que $P(Z > t_{n;\alpha}) = \alpha$



$n \setminus \alpha$	0,10	0,05	0,025	0,01	0,005	0,001
1	3,078	6,314	12,706	31,821	63,657	318,309
2	1,886	2,920	4,303	6,965	9,925	22,327
3	1,638	2,353	3,182	4,541	5,841	10,215
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
31	1,309	1,696	2,040	2,453	2,744	3,375
32	1,309	1,694	2,037	2,449	2,738	3,365
33	1,308	1,692	2,035	2,445	2,733	3,356
34	1,307	1,691	2,032	2,441	2,728	3,348
35	1,306	1,690	2,030	2,438	2,724	3,340
36	1,306	1,688	2,028	2,434	2,719	3,333
37	1,305	1,687	2,026	2,431	2,715	3,326
38	1,304	1,686	2,024	2,429	2,712	3,319
39	1,304	1,685	2,023	2,426	2,708	3,313
40	1,303	1,684	2,021	2,423	2,704	3,307

$n \setminus \alpha$	0,10	0,05	0,025	0,01	0,005	0,001
41	1,303	1,683	2,020	2,421	2,701	3,301
42	1,302	1,682	2,018	2,418	2,698	3,296
43	1,302	1,681	2,017	2,416	2,695	3,291
44	1,301	1,680	2,015	2,414	2,692	3,286
45	1,301	1,679	2,014	2,412	2,690	3,281
46	1,300	1,679	2,013	2,410	2,687	3,277
47	1,300	1,678	2,012	2,408	2,685	3,273
48	1,299	1,677	2,011	2,407	2,682	3,269
49	1,299	1,677	2,010	2,405	2,680	3,265
50	1,299	1,676	2,009	2,403	2,678	3,261
51	1,298	1,675	2,008	2,402	2,676	3,258
52	1,298	1,675	2,007	2,400	2,674	3,255
53	1,298	1,674	2,006	2,399	2,672	3,251
54	1,297	1,674	2,005	2,397	2,670	3,248
55	1,297	1,673	2,004	2,396	2,668	3,245
56	1,297	1,673	2,003	2,395	2,667	3,242
57	1,297	1,672	2,002	2,394	2,665	3,239
58	1,296	1,672	2,002	2,392	2,663	3,237
59	1,296	1,671	2,001	2,391	2,662	3,234
60	1,296	1,671	2,000	2,390	2,660	3,232
61	1,296	1,670	2,000	2,389	2,659	3,229
62	1,295	1,670	1,999	2,388	2,657	3,227
63	1,295	1,669	1,998	2,387	2,656	3,225
64	1,295	1,669	1,998	2,386	2,655	3,223
65	1,295	1,669	1,997	2,385	2,654	3,220
66	1,295	1,668	1,997	2,384	2,652	3,218
67	1,294	1,668	1,996	2,383	2,651	3,216
68	1,294	1,668	1,995	2,382	2,650	3,214
69	1,294	1,667	1,995	2,382	2,649	3,213
70	1,294	1,667	1,994	2,381	2,648	3,211
71	1,294	1,667	1,994	2,380	2,647	3,209
72	1,293	1,666	1,993	2,379	2,646	3,207
73	1,293	1,666	1,993	2,379	2,645	3,206
74	1,293	1,666	1,993	2,378	2,644	3,204
75	1,293	1,665	1,992	2,377	2,643	3,202
∞	1,282	1,645	1,960	2,326	2,576	3,090