Examen de 1ère session de MAPSI

Durée : 2 heures

Seuls documents autorisés : Calculatrices et votre antisèche recto-verso
– Barême indicatif –

Exercice 1 (2.5 pts) – Modèle de langue

On souhaite réaliser un classifieur automatique de langue (anglais, français, italien par exemple) basé sur les enchainements entre 3 types de caractères : les voyelles, les consonnes et les autres caractères (espace, ponctuation...). On utilisera les notations suivantes. Corpus documentaire associé à la langue ℓ : $X_\ell = \{d_i\}_{i=1,\dots,N_\ell}$. Chaque document étant composé de caractères : $d_i = \{c_{ij}\}_{j=1,\dots,|d_i|}$

Q 1.1 (1 pt) Nous optons pour un modèle de Markov simple. Combien de paramètres sont nécessaires en tout (ensembles des paramètres de tous les modèles) ? Pour apprendre un tel modèle à partir de documents de quelques centaines de caractères chacun : diriez-vous qu'il est nécessaire d'avoir des dizaines, des centaines, des milliers ou des millions de documents disponibles dans chaque langue ?

Modèle de markov simple à 3 états : consonne, voyelle, autre : $\forall i, j\ c_{ij} \in \mathcal{C} = \{1 = C, 2 = V, 3 = A\}$
Pour chaque langue, le modèle est composé de $\Pi \in \mathbb{R}^3, A \in \mathbb{R}^{3\times 3}$
Au total, il y a donc $12 \times 3 = 36$ paramètres
A partir de quelques dizaines (ou éventuellement centaines) de documents, on est en mesure d'évaluer les paramètres du modèle convenablement (toutes les transitions auront été observées un grand nombre de fois).

Q 1.2 (1 pt) Comment apprendre un tel modèle à partir d'un corpus de documents dans les 3 langues ? Formuler le problème d'optimisation en fonction des documents d_i et rappeler la définition des paramètres optimaux (sans le calcul d'optimisation).

— Apprendre 3 modèles (1 pour chaque langue) : max de vraisemblance (ou max de log-vraisemblance)

$$\forall \ell, \arg\max_{A_\ell, \Pi_\ell} \sum_{i=1} \log p(d_i | A_\ell, \Pi_\ell)$$

Bien vérifier que les étudiants ont identifié le arg max et les paramètres à optimiser ainsi que le fait que le comptage se fait sur chacun des sous-corpus indépendamment.

— Chaque modèle est appris par comptage (comme vu en cours, TD, TME) : comptage des transitions entre caractères et comptage (A) des caractères initiaux des documents (Π)

$$A = \{a_{nm}\}, \qquad a_{nm} = \frac{|\{(c_{ij}, c_{i,j+1}) | c_{ij} = n, c_{i,j+1} = m\}|}{|\{(c_{ij}, c_{i,j+1})\}|}$$

Note : les étudiants peuvent exprimer le comptage des transitions comme ils le souhaitent (explication texte, variante de formulation...). On vérifie juste que c'est raisonnable.

Q 1.3 (0.5 pt) Soit la série de 3 caractères pris au début d'un document : `abr`
Expliquer comment utiliser le modèle précédemment appris pour classer cette chaine (formuler le problème d'inférence en fonction des paramètres des modèles).

Inférence : faire passer les nouveaux textes dans les 3 modèles pour trouver le plus vraisemblable. Avec un codage $c_{ij} \in \mathcal{C} = \{1 = C, 2 = V, 3 = A\}$

$$classe = \arg\max_{\ell} p(d = "abr"|\Pi_\ell, A_\ell) = \arg\max_{\ell} \log(\Pi_\ell(2)) + \log(A_\ell(2,1)) + \log(A_\ell(1,1))$$

Q 1.4 ((bonus) à faire à la fin) Comment procéder si la séquence de 3 caractères est tirée du centre du document ?

Il faut utiliser la distribution stationnaire μ au lieu de Π. μ donnant la distribution des caractères sur l'ensemble des documents.

Exercice 2 (5pts) – Indépendances conditionnelles

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 :

	z_1		z_2	
	y_1	y_2	y_1	y_2
x_1	$0,064$	$0,036$	$0,016$	$0,084$
x_2	$0,448$	$0,072$	$0,112$	$0,168$

Q 2.1 Est-ce que X est indépendant de Z ? Vous justifierez votre réponse.

On calcule :

$$P(X, Z) = \sum_{Y} P(X, Y, Z) =$$

	z_1	z_2
x_1	0.10	0.10
x_2	0.52	0.28

Si X et Z étaient indépendants, la deuxième ligne serait proportionnelle à la première. Donc X et Z ne sont pas indépendants.

Q 2.2 Est-ce que X est indépendant de Z conditionnellement à Y ? Vous justifierez votre réponse.

Pour qu'il y ait indépendance, il faut montrer que i) $P(X, Z|Y) = P(X|Y)P(Z|Y)$ ou bien ii) $P(Z|X, Y) = P(Z|Y)$.

Pour l'alternative i), on commence par calculer :

$$P(X,Y) = \sum_Z P(X,Y,Z) =$$

	y_1	y_2
x_1	0.08	0.12
x_2	0.56	0.24

puis $P(Y) = \sum_X P(X,Y) = [0.64,\ 0.36]$. On peut maintenant diviser $P(X,Y)$ par $P(Y)$ pour obtenir $P(X|Y)$:

$$P(X|Y) =$$

	y_1	y_2
x_1	0.125	0.34
x_2	0.875	0.66

Pour calculer $P(Z|Y)$, on commence par calculer $P(Z,Y) = \sum_X P(X,Y,Z)$:

$$P(Z,Y) =$$

	y_1	y_2
z_1	0.512	0.108
z_2	0.128	0.252

Ensuite, on divise par $P(Y)$:

$$P(Z|Y) =$$

	y_1	y_2
z_1	0.8	0.3
z_2	0.2	0.7

Si l'on multiplie $P(X|Y)$ par $P(Z|Y)$, on obtient :

$$P(X|Y) \times P(Z|Y) =$$

	z_1		z_2	
	y_1	y_2	y_1	y_2
x_1	0.1	0.1	0.025	0.233
x_2	0.7	0.2	0.175	0.466

Il faut comparer cette table à celle de $P(X,Y,Z)/P(Y)$. Si l'on calcule le résultat de cette division, on tombe sur la même table. Donc X et Z sont bien indépendants conditionnellement à Y.
L'alternative ii) est plus simple : on divise $P(X,Y,Z)$ par $P(X,Y)$ que l'on a calculé plus haut :

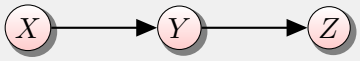
$$P(Z|X,Y) =$$

	z_1		z_2	
	y_1	y_2	y_1	y_2
x_1	0,8	0,3	0,2	0,7
x_2	0,8	0,3	0,2	0,7

On remarque que la première ligne est égale à la deuxième, donc $P(Z|X,Y)$ ne dépend pas de X. Donc X est indépendant de Z conditionnellement à Y.

Q 2.3 On sait que $P(X,Y,Z) = P(Z|X,Y)P(Y|X)P(X)$. En utilisant les questions précédentes, simplifiez cette expression. Quel réseau bayésien en déduit-on ?

Puisque X est indépendant de Z conditionnellement à Y, $P(Z|X,Y) = P(Z|Y)$. On obtient donc le réseau bayésien :



Exercice 3 (5pts) – 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 $X =$ « couleur d'une boule tirée au hasard dans l'urne ». 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 3.1 Estimez par maximum de vraisemblance les paramètres de la distribution $P(X)$. Vous justifierez votre réponse.

La vraisemblance de l'échantillon est $L(\mathbf{x}, \theta) = p_R^4 \times p_B^2 \times p_V^3 \times p_J$. Donc la log-vraisemblance est égale à : $\log L(\mathbf{x}, \theta) = 4 \log p_R + 2 \log p_B + 3 \log p_V + \log p_J$. Sachant que $p_J = 1 - p_R - p_B - p_V$, c'est équivalent à :

$$\log L(\mathbf{x}, \theta) = 4 \log p_R + 2 \log p_B + 3 \log p_V + \log(1 - p_R - p_B - p_V).$$

Pour obtenir le max de vraisemblance, on dérive $\log L(\mathbf{x}, \theta)$:

$$\frac{\partial \log L(\mathbf{x}, \theta)}{\partial p_R} = \frac{4}{p_R} - \frac{1}{1 - p_R - p_B - p_V} = 0 \Longleftrightarrow 4 - 4p_R - 4p_B - 4p_V - p_R = 0 \qquad (\alpha)$$

$$\frac{\partial \log L(\mathbf{x}, \theta)}{\partial p_B} = \frac{2}{p_B} - \frac{1}{1 - p_R - p_B - p_V} = 0 \Longleftrightarrow 2 - 2p_R - 2p_B - 2p_V - p_B = 0 \qquad (\beta)$$

$$\frac{\partial \log L(\mathbf{x}, \theta)}{\partial p_V} = \frac{3}{p_V} - \frac{1}{1 - p_R - p_B - p_V} = 0 \Longleftrightarrow 3 - 3p_R - 3p_B - 3p_V - p_V = 0 \qquad (\gamma)$$

En soustrayant 2 fois l'équation (β) à (α), on obtient $p_R = 2p_B$. En calculant $3(\beta) - 2(\alpha)$, on obtient $p_V = \frac{3}{2} p_B$. Par conséquent, en remplaçant dans (β) les valeurs de p_R et p_V par leur équivalent en p_B, on obtient $2 - 10p_B = 0$, d'où :

$$p_R = 0.4 \qquad p_B = 0.2 \qquad p_V = 0.3 \qquad p_J = 0.1$$

Q 3.2 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.

Le maximum a posteriori est défini par :

$$\begin{aligned}\theta_{MAP} &= \text{Argmax}_\theta P(\mathbf{x}, \theta) = \text{Argmax}_\theta L(\mathbf{x}, \theta)\pi(\theta) \\ &= \text{Argmax}_\theta (p_R^4 \times p_B^2 \times p_V^3 \times p_J) \times (p_R^2 \times p_B \times p_V^3 \times p_J^2) \\ &= \text{Argmax}_\theta (p_R^6 \times p_B^3 \times p_V^6 \times p_J^3).\end{aligned}$$

On procède maintenant comme dans la question précédente :

$$\frac{\partial \log P(\mathbf{x}, \theta)}{\partial p_R} = \frac{6}{p_R} - \frac{3}{1 - p_R - p_B - p_V} = 0 \iff 6 - 6p_R - 6p_B - 6p_V - 3p_R = 0 \qquad (\alpha)$$

$$\frac{\partial \log L(\mathbf{x}, \theta)}{\partial p_B} = \frac{3}{p_B} - \frac{3}{1 - p_R - p_B - p_V} = 0 \iff 3 - 3p_R - 3p_B - 3p_V - 3p_B = 0 \qquad (\beta)$$

$$\frac{\partial \log L(\mathbf{x}, \theta)}{\partial p_V} = \frac{6}{p_V} - \frac{3}{1 - p_R - p_B - p_V} = 0 \iff 6 - 6p_R - 6p_B - 6p_V - 3p_V = 0 \qquad (\gamma)$$

On obtient donc que $p_R = p_V = 2p_B$. Par conséquent :

$$p_R = \frac{1}{3} \qquad p_B = \frac{1}{6} \qquad p_V = \frac{1}{3} \qquad p_J = \frac{1}{6}$$

Exercice 4 (5 pts) – Paris sportifs

Nous souhaitons construire une machine efficace pour prendre des paris sportifs. Afin de simplifier le problème, nous avons opté pour le basket-ball où chaque match est gagné ou perdu (pas de match nul). Dans un premier temps, nous avons récupéré un historique de données sur les matchs passés et nous disposons d'une base matricielle X où chaque ligne décrit une équipe e à la veille du match cible :

- X_1 : pourcentage de victoire de e sur les 20 dernières rencontres,
- X_2 : pourcentage de réussite au tir de e sur les 20 dernières rencontres,
- X_3 : moyenne de points marqués (divisée par 100) pour l'équipe e sur les 20 dernières rencontres,
- X_4 : moyenne de points encaissés (divisée par 100) pour l'équipe e sur les 20 dernières rencontres.

Nous disposons également du résultat du match de l'équipe e : $Y \in \{0, 1\}$, le succès étant codé par 1. On notera $\mathbf{x}_i \in \mathbb{R}^4$ la description de e pour le match i et y_i le résultat. Nous choisissons de modéliser

$p(Y = 1|X)$ par une fonction logistique paramétrée par $\mathbf{w}$:

$$p(Y = 1|X = \mathbf{x}_i) = \sigma_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}$$

Q 4.1 (0.5 pt) Quelle est la dimension du vecteur de paramètre à estimer ?

Q 4.2 (0.5 pt) Que signifie une probabilité de 0.5 pour $p(Y = 1|X = \mathbf{x}_i)$? Quand cette valeur est-elle atteinte ?

$$\mathbf{w} \in \mathbb{R}^4$$

La valeur de 0.5 correspond à la frontière de décision, la zone où un match est parfaitement indécis, correspond à $p(Y = 1|X = \mathbf{x}_i) = 0.5$ càd $\mathbf{w}\mathbf{x}_i = \sum_j w_j x_{ij} = 0$ (l'équation d'un hyper-plan)

Q 4.3 (0.5 pt) Montrer que : $p(Y = y_i|X = \mathbf{x}_i) = \sigma_{\mathbf{w}}(\mathbf{x}_i)^{y_i}(1-\sigma_{\mathbf{w}}(\mathbf{x}_i))^{(1-y_i)}$. Quelle loi reconnaissez-vous alors ?

Loi de Bernoulli. La formule permet de retrouver la probabilité de victoire si $y_i = 1$ et la proba de défaite si $y_i = 0$

Q 4.4 (2 pt) Exprimer la vraisemblance puis la log-vraisemblance sur l'ensemble des données sous la forme $\sum_i (\alpha \log(\sigma_{\mathbf{w}}(\mathbf{x}_i)) + \beta \mathbf{w}\mathbf{x}_i)$.

$$\mathcal{L} = \prod_i p(Y = y_i|X = \mathbf{x}_i) = \prod_i \left(\frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}\right)^{y_i} \left(1 - \frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}\right)^{1-y_i}$$

$$\log \mathcal{L} = \sum_i \log p(Y = y_i|X = \mathbf{x}_i) = \sum_i y_i \log \left(\frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}\right) + (1 - y_i) \log \left(1 - \frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}\right)$$

$$1 - \frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)} = \frac{\exp(\mathbf{w}\mathbf{x}_i)}{1 + \exp(\mathbf{w}\mathbf{x}_i)}$$

$$\log \mathcal{L} = \sum_i \log \left(\frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}\right) (y_i + 1 - y_i) + (1 - y_i)\mathbf{w}\mathbf{x}_i$$

$$\log \mathcal{L} = \sum_i \log \left(\frac{1}{1 + \exp(\mathbf{w}\mathbf{x}_i)}\right) + (1 - y_i)\mathbf{w}\mathbf{x}_i, \quad \text{soit : } \alpha = 1,\ \beta = 1 - y_i$$

Q 4.5 (1.5 pts) Calculer le gradient et indiquer brièvement la procédure à suivre pour obtenir les paramètres optimaux $\mathbf{w}^\star$.
Note pour le calcul du gradient : $(\log(1 + \exp(u)))' = \frac{u' \exp(u)}{1+\exp(u)}$

Calcul du gradient + Annulation du gradient (qui ne sera pas possible analytiquement ⇒ descente de gradient)

$$\nabla_w \log \mathcal{L} = \begin{bmatrix} \dots \\ \frac{\partial \log \mathcal{L}}{\partial w_j} \\ \dots \end{bmatrix}, \qquad \frac{\partial \log \mathcal{L}}{\partial w_j} = \sum_i -\frac{x_{ij} \exp(\mathbf{w}\mathbf{x}_i)}{1+\exp(\mathbf{w}\mathbf{x}_i)} + (1-y_i)x_{ij}$$

$$\frac{\partial \log \mathcal{L}}{\partial w_j} = \sum_i x_{ij}\left(\frac{1}{1+\exp(\mathbf{w}\mathbf{x}_i)} - y_i\right)$$

Vue la forme du gradient, on ne peut pas l'annuler analytiquement, il faut procéder par descente de gradient pour obtenir $\mathbf{w}^\star$.

Q 4.6 (bonus, à faire à la fin) Notre modèle présente une faiblesse évidente : il ne prend pas en compte l'équipe adverse dans la prédiction de victoire. Proposer un modèle (éventuellement très proche du modèle actuel) pour prendre en compte l'opposition adverse. Proposer une technique d'apprentissage de ce modèle.

Il suffit de reprendre exactement de le même modèle, mais sur une différence entre les caractéristiques des deux équipes :

$$p(Y=1|X_1=\mathbf{x}_i, X_2=\mathbf{x}_j) = \sigma_{\mathbf{w}}(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{1+\exp(\mathbf{w}(\mathbf{x}_i-\mathbf{x}_j))}$$

Tout le processus d'optimisation est identique, seules les données à fournir en entrée du modèle sont à retravailler.

Exercice 5 (2.5 pts) – Analyse du gain

L'exercice précédent est un jalon nécessaire mais pas suffisant : les probabilités de victoire ne sont pas suffisantes, il faut connaitre les cotes et calculer un gain pour savoir s'il est intéressant de parier. Nous disposons des cotes de victoire c_i, associées aux matchs décrits précédemment par $(\mathbf{x}_i, y_i)$. On fait l'hypothèse que les paramètres du modèle sont optimisés et fixés, ainsi, la probabilité de victoire associé au match est simplement écrite p_i.
Rappel : lorsque la cote de victoire c_i associée à un match vaut 1.2, cela signifie que si l'on mise 1 sur la victoire, on reçoit 1.2 en cas de victoire effective et rien en cas de défaite (à vous d'en déduire le gain).

Q 5.1 (0.5 pt) Donner la distribution de la variable aléatoire discrète de gain G_i associé à une mise unitaire sur le match i.

G_i :	(c_i-1)	-1
	p_i	$1-p_i$

Q 5.2 (1 pt) A quelle condition est-il intéressant de parier ? Exprimer le critère en fonction de p_i et c_i. Parmi tous les matchs de la soirée, sur lequel pariez-vous de façon prioritaire ?

Si l'espérance du gain est positive :

$$E[G_i] = p_i(c_i - 1) + (-1)(1 - p_i) = p_i c_i - 1$$

Il faut parier si : $E[G_i] > 0 \iff p_i > 1/c_i$
On parie en premier sur le match associé à l'espérance de gain la plus forte.

Q 5.3 (1 pt) Deux matchs *intéressants* (et indépendants) ont des cotes et des probabilités de victoire très proches : $c_i \approx c_j, p_i \approx p_j$. On hésite entre deux stratégies de gain S_1 et S_2 : miser 1 euros sur le *meilleur* match ou miser 0.5 euros sur chacun match. Comparer les espérances et les variances de S_1 et S_2 et déterminer l'approche la plus robuste.

$$E[S_1] = E[G_i] = p_i c_i - 1$$
$$E[S_2] = E[1/2G_i + 1/2G_j] = 0.5(p_i c_i - 1) + 0.5(p_j c_j - 1) \approx p_i c_i - 1$$

$$V[S_2] = V[1/2G_i + 1/2G_j] = V[1/2G_i] + V[1/2G_j], \qquad \text{variables indep.}$$
$$V[S_2] \approx 2 * 1/4 V[G_i] = 1/2 V[G_i] = V[S_1]/2$$

Très facile à re-démontrer avec :

$$V[S_1] = V[G_i] = p_i((c_i - 1) - E[G_i])^2 + (1 - p_i)((-1) - E[G_i])^2$$

Mais pas obligatoire...
La seconde stratégie présente la même espérance de gain mais une variance deux fois moins élevée : elle est plus robuste.

Q 5.4 (bonus, à faire à la fin) Pourrait-on utiliser le théorème central limite pour calculer le gain moyen sur un ensemble de match et en déduire un intervalle de confiance sur un gain cible ?

Non, car chaque match correspond à une variable aléatoire tirée selon un loi spécifique.