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

Exercice 1 – 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 1.1 Estimez par maximum de vraisemblance la valeur du paramètre $\theta = p$ de la loi.

Q 1.2 Avant le tirage de l'échantillon, nous avions une connaissance a priori sur le paramètre θ : 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 $\theta = p$ par maximum a posteriori.

Exercice 2 – 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 2.1 Calculez la valeur de θ par maximum de vraisemblance. Bien entendu, vous démontrerez mathématiquement votre résultat.

Q 2.2 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 3 – 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 3.1 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 3.2 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 3.3 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 3.4 Donner les valeurs d'acceptation des diverses hypothèses sur la valeur du paramètre :

Q 3.4.1 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

$$d(x) = \underset{\theta}{\text{Argmax}}\, \pi(\theta|x).$$

Q 3.4.2 quand la règle de décision est celle du *maximum de vraisemblance*.

Q 3.4.3 Quand ces deux règles donnent-elles le même résultat ?

Exercice 4 – 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 4.1 On suppose que la distribution des durées de vie est effectivement une loi exponentielle. Estimez par maximum de vraisemblance la valeur de λ.

Q 4.2 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 5 – 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 5.1 L'observateur pense que la variable aléatoire X=« temps d'exécution » 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 6 – 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 6.1 Donnez un cas d'usage classique pour ce type de classifieur de textes. Imaginez un modèle simple pour répondre à ce problème.

Q 6.2 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 6.3 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 6.4 Pour trouver les paramètres Θ_m, nous allons maximiser la log-vraisemblance de ces paramètres sur l'ensemble de la classe m du corpus.

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.

Montrez que le problème d'optimisation permettant de trouver Θ_m en fonction des $P(w_j|\Theta_m)$ s'écrit :

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

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 6.5 Simplifions les notations $P(w_j|\Theta_m) \rightarrow \theta_j$. 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 6.6 A votre avis, pourquoi cet algorithme s'appelle-t-il *Naive Bayes* ?
Intuitivement, pensez-vous qu'il donne de bons résultats ?

Q 6.7 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 6.8 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 6.9 Quels sont les mots qui *participent* le plus à la classification des phrases ? Quels traitements effectuer pour avoir des *mots-clés* plus pertinents ?