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 : $X_\ell = \{d_i\}_{i=1,\ldots,N_\ell}$. Chaque document étant composé de caractères : $d_i = \{c_{ij}\}_{j=1,\ldots,|d_i|}$

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 ?

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).

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).

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

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.

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

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 ?

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.

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.

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 ?

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 ?

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

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 : $(\log(1 + \exp(u)))' = \frac{u' \exp(u)}{1+\exp(u)}$

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.

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.

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 ?

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.

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 ?