IFT 6390A Fondements de l'apprentissage machine
Prof. Guillaume Rabusseau

Devoir 1 - Partie Théorique

- Ce devoir doit être fait et envoyé sur Gradescope individuellement. Vous pouvez discuter avec d'autres étudiants mais les réponses que vous soumettez doivent être les vôtres. A noter que nous utiliserons l'outil de détection de plagiat de Gradescope. Tous les cas suspectés de plagiat seront enregistrés et transmis à l'Université pour vérification.
- Vous devez soumettre vos solutions au format pdf sur Gradescope en utilisant le devoir intitulé `Devoir 1 / Homework 1 - Pratique/Practical - 6390A/B`.

1. **Rappels de probabilités: probabilité conditionnelle et règle de Bayes** [5 points]

   (a) Donnez la définition de la probabilité conditionnelle de la variable aléatoire discrète X sachant la variable aléatoire discrète Y.

   (b) Soit une pièce déséquilibrée dont la probabilité d'obtenir face est $2/3$ et la probabilité d'obtenir pile est $1/3$. Cette pièce est lancée à trois reprises. Quelle est la probabilité d'obtenir exactement deux faces (parmis les trois lancers), sachant que le premier lancer a fait face?

   (c) Donnez deux expressions équivalentes de $P(X, Y)$:

   (i) en fonction de $\mathbb{P}(X)$ et $\mathbb{P}(Y|X)$
   (ii) en fonction de $\mathbb{P}(Y)$ et $\mathbb{P}(X|Y)$

   (d) Prouvez le théorème de Bayes:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

   (e) Un sondage des étudiants Montréalais est fait, où 60% des élèves sondés sont affiliés à l'UdeM alors que les autres sont affiliés à McGill. Un étudiant est choisi aléatoirement parmis ce groupe.

   i. Quelle est la probabilité que l'étudiant soit affilié à McGill?

ii. Considérons maintenant que l'étudiant est bilingue, et que 60% des étudiants de l'UdeM sont bilingues alors que seulement 30% des étudiants de McGill le sont. Étant donné cette information, quelle est la probabilité que cet étudiant soit affilé à McGill?

Réponses

(a) La probabilité conditionnelle de la variable aléatoire discrète X sachant la variable aléatoire discrète Y est la probabilité d'un évènement X sachant q'un autre évènement Y a eu lieu. Elle s'écrit : $\mathbb{P}(X|Y)$. Par définition : $\mathbb{P}(X|Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$

(b) On pose :
F : l'évènement obtenir face
P : l'évènement obtenir pile.
avec $\mathbb{P}(F) = \frac{2}{3}$ et $\mathbb{P}(P) = \frac{1}{3}$. On recherche la probabilité d'obtenir exactement deux faces parmi 3 lancers, sachant que le premier lancer a fait face.
On cherche donc $\mathbb{P}(FP|F) + \mathbb{P}(PF|F)$.

$$\mathbb{P}(FP|F) + \mathbb{P}(PF|F) = \frac{\mathbb{P}(FP, F)}{\mathbb{P}(F)} + \frac{\mathbb{P}(PF, F)}{\mathbb{P}(F)}$$

Or, puisque les évènements sont indépendants, on a :

$$\mathbb{P}(FP, F) = \mathbb{P}(FP) \times \mathbb{P}(F), \qquad \mathbb{P}(PF, F) = \mathbb{P}(PF) \times \mathbb{P}(F)$$

On a alors :

$$\mathbb{P}(FP|F) + \mathbb{P}(PF|F) = \frac{\mathbb{P}(FP) \times \mathbb{P}(F)}{\mathbb{P}(F)} + \frac{\mathbb{P}(PF) \times \mathbb{P}(F)}{\mathbb{P}(F)}$$

$$\mathbb{P}(FP|F) + \mathbb{P}(PF|F) = \frac{2 \times \frac{1}{3} \times (\frac{2}{3})^2}{\frac{2}{3}}$$

D'où :

$$\mathbb{P}(FP|F) + \mathbb{P}(PF|F) = \frac{4}{9}$$

(c) Deux expressions équivalentes de $\mathbb{P}(X, Y)$:
On sait que, par définition :

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(Y)} \qquad \mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)}$$

i. en fonction de $\mathbb{P}(X)$ et $\mathbb{P}(Y|X)$:

$$\mathbb{P}(X,Y) = \mathbb{P}(X)\,\mathbb{P}(Y|X)$$

ii. en fonction de $\mathbb{P}(Y)$ et $\mathbb{P}(X|Y)$:

$$\mathbb{P}(X,Y) = \mathbb{P}(Y)\,\mathbb{P}(X|Y)$$

(d) En injectant $\mathbb{P}(X,Y) = \mathbb{P}(X)\,\mathbb{P}(Y|X)$ dans l'égalité : $\mathbb{P}(X|Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$, on obtient la formule du théorème de Bayes :

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X)\,\mathbb{P}(Y|X)}{\mathbb{P}(Y)}$$

(e) On pose :
U : l'évènement l'élève est étudiant à l'UdeM.
M : l'évènement l'élève est étudiant à McGill.

i. La probabilité que l'étudiant soit affilié à McGill est :
$\mathbb{P}(M) = 1 - \mathbb{P}(U) = 1 - 0.6 = 0.4$

ii. On pose maintenant :
B : l'évènement l'élève est billingue.
On utlise la formule de Bayes :

$$\mathbb{P}(M|B) = \frac{\mathbb{P}(M)\,\mathbb{P}(B|M)}{\mathbb{P}(B)}$$

De plus, la formule des probabilités totales nous donne le résultat suivant :

$$\mathbb{P}(B) = \mathbb{P}(M,B) + \mathbb{P}(U,B) = \mathbb{P}(M) \times \mathbb{P}(B|M) + \mathbb{P}(U) \times \mathbb{P}(B|U)$$

D'où :

$$\begin{aligned}
\mathbb{P}(M|B) &= \frac{\mathbb{P}(M)\,\mathbb{P}(B|M)}{\mathbb{P}(M) \times \mathbb{P}(B|M) + \mathbb{P}(U) \times \mathbb{P}(B|U)} \\
\mathbb{P}(M|B) &= \frac{0.3 \times 0.4}{0.4 \times 0.3 + 0.6 \times 0.6} \\
\mathbb{P}(M|B) &= \frac{1}{4}
\end{aligned}$$

2. **Bag of words (sac de mots) et modèle de sujet unique** [15 points]

On s'intéresse à un problème de classification où on veut prédire le sujet d'un document d'un certain corpus (ensemble de documents). Le sujet de chaque document peut être soit *sport*, soit *politique*. 1/3 des documents du corpus sont sur le *sport*, et 2/3 sont sur la *politique*.

On va utiliser un modèle très simple où on ignore l'ordre des mots apparaissant dans le document et on suppose que les mots dans un document sont indépendants les uns des autres, étant donné le sujet du document.

De plus, nous allons utiliser des statistiques très simples des documents: les probabilités qu'un mot choisi au hasard dans un document soit "goal", "kick", "congress", "vote", ou n'importe quel autre mot (dénoté par *other*). Nous appelons ces cinq catégories le vocabulaire ou dictionnaire pour les documents: $V = \{$*"goal"*, *"kick"*, *"congress"*, *"vote"*, *other*$\}$.

Soit les distributions suivantes des mots du vocabulaire, par sujet:

	$\mathbb{P}(\text{mot} \mid \text{sujet} = \textit{sport})$	$\mathbb{P}(\text{mot} \mid \text{sujet} = \textit{politique})$
mot = *"goal"*	1/100	5/1000
mot = *"kick"*	2/100	1/1000
mot = *"congress"*	1/1000	1/100
mot = *"vote"*	3/1000	4/100
mot = *other*	966/1000	944/1000

Table 1

Cette table nous dit par exemple que la probabilité qu'un mot choisi aléatoirement dans un document soit *"vote"* n'est que de 3/1000 si le sujet du document est le *sport*, mais est de 4/100 si le sujet est la *politique*.

(a) Quelle est la probabilité qu'un mot aléatoire dans un document soit "goal" étant donné que le sujet est la *politique* ?

(b) Quelle est l'espérance du nombre de fois où le mot "congress" apparait dans un document de 2000 mots dont le sujet est le *sport*?

(c) On tire aléatoirement un document du corpus. Quelle est la probabilité qu'un mot aléatoire de ce document soit "goal"?

(d) Supposons que l'on tire aléatoirement un mot d'un document et que ce mot est "kick". Quelle est la probabilité que le sujet du document soit le *sport*?

(e) Supposons que l'on tire aléatoirement deux mots d'un document et que le premier soit "kick". Quelle est la probabilité que le second mot soit "vote"?

(f) Pour en revenir à l'apprentissage, supposons que nous ne savons pas les probabilités conditionnelles étant donné chaque sujet ni les probabilités de chaque sujet (i.e. nous n'avons pas accès aux informations de la table 1 où aux proportions de chaque sujet), mais que nous avons un jeu de données de N documents où chaque document est annoté avec un des sujets *sport* ou *politique*. Comment estimeriez-vous les probabilitiés conditionelles (e.g., $\mathbb{P}(\text{mot} = \text{"}goal\text{"} \mid \text{sujet} = politique)$) et les probabilités des sujets (e.g., $\mathbb{P}(\text{sujet} = politique)$) à partir de ce jeu de données?

Réponses

(a) La probabilité qu'un mot aléatoire dans un document soit "goal" étant donné que le sujet est la *politique* est : $\frac{5}{1000}$.

(b) On sait que la probabilité que le mot "congress" apparaisse dans un document dont le sujet est le *sport* est $\frac{1}{1000}$. On peut donc s'attendre à obtenir le mot "congress" 1 fois tous les 1000 mots dans un document de type *sport*. Dans un document de 2000 mots dont le sujet est le *sport*, l'espérance du nombre de fois où le mot "congress" apparait est donc de 2.

(c) D'après la formule des probabilités totals :
$\mathbb{P}(\text{mot} = \text{"goal"}) =$
$\mathbb{P}(\text{sujet} = \text{"sport"}) \times \mathbb{P}(\text{mot} = \text{"goal"} \mid \text{sujet} = \text{"sport"}) +$
$\mathbb{P}(\text{sujet} = \text{"politique"}) \times \mathbb{P}(\text{mot} = \text{"goal"} \mid \text{sujet} = \text{"politique"})$
Donc :
$\mathbb{P}(\text{mot} = \text{"goal"}) = \frac{1}{3} \times \frac{1}{100} + \frac{2}{3} \times \frac{5}{1000}$

$\mathbb{P}(\text{mot} = \text{“goal”}) = \frac{1}{150}$

(d) On suppose que l'on tire aléatoirement un mot d'un document et que ce mot est “kick”. La probabilité que le sujet du document soit le sport est : $\mathbb{P}(\text{sujet} = \text{“sport”} \mid \text{mot} = \text{“kick”})$.
Cette probabilité est égale à :

$$\frac{\mathbb{P}(\text{mot} = \text{"kick"} \mid \text{sujet} = \text{"sport"}) \times \mathbb{P}(\text{sujet} = \text{"sport"})}{\mathbb{P}(\text{mot} = \text{"kick"})}$$

Or, d'après la formule des probabilités totals :
$\mathbb{P}(\text{mot} = \text{“kick”}) =$
$\mathbb{P}(\text{sujet} = \text{“sport”}) \times \mathbb{P}(\text{mot} = \text{“kick”} \mid \text{sujet} = \text{“sport”}) +$
$\mathbb{P}(\text{sujet} = \text{“politique”}) \times \mathbb{P}(\text{mot} = \text{“kick”} \mid \text{sujet} = \text{“politique”})$
Donc :

$$\mathbb{P}(\text{sujet} = \text{"sport"} \mid \text{mot} = \text{"kick"}) = \frac{\frac{2}{100} \times \frac{1}{3}}{\frac{1}{3} \times \frac{2}{100} + \frac{2}{3} \times \frac{1}{1000}}$$
$$\mathbb{P}(\text{sujet} = \text{"sport"} \mid \text{mot} = \text{"kick"}) = \frac{10}{11}$$

(e) On tire aléatoirement deux mots d'un document. Le premier est “kick”. On cherche la probabilité que le second mot soit “vote”.
D'après le théorème des probabilités totales, on sait que :

$$\mathbb{P}(B) = \mathbb{P}(A) \times \mathbb{P}(B|A) + \mathbb{P}(\overline{A}) \times \mathbb{P}(B|\overline{A})$$

On a alors :

$$\begin{aligned}\mathbb{P}(\text{mot ="vote"}|\text{mot="kick"}) &= \mathbb{P}(\text{"vote", "politique"}|\text{"kick"}) \\ &\quad + \mathbb{P}(\text{"vote", "sport"}|\text{"kick"}) \\ &= \mathbb{P}(\text{"vote"}|\text{"politique","kick"}) \times \mathbb{P}(\text{"politique"}|\text{"kick"}) \\ &\quad + \mathbb{P}(\text{"vote"}|\text{ "sport","kick"}) \times \mathbb{P}(\text{"sport"}|\text{"kick"})\end{aligned}$$

Or on sait que les mots sont indépendants les uns des autres dans un même document. Donc :

$$\begin{aligned}\mathbb{P}(\text{mot ="vote"}|\text{mot="kick"}) &= \mathbb{P}(\text{"vote"}|\text{"politique","kick"}) \times \mathbb{P}(\text{"politique"}|\text{"kick"}) \\ &\quad + \mathbb{P}(\text{"vote"}|\text{ "sport","kick"}) \times \mathbb{P}(\text{"sport"}|\text{"kick"}) \\ &= \mathbb{P}(\text{"vote"}|\text{"politique"}) \times \mathbb{P}(\text{"politique"}|\text{"kick"}) \\ &\quad + \mathbb{P}(\text{"vote"}|\text{ "sport"}) \times \mathbb{P}(\text{"sport"}|\text{"kick"})\end{aligned}$$

On a de plus :

$$\begin{aligned}
\mathbb{P}(\text{"vote"}|\text{"politique"}) &= \frac{4}{100} \\
\mathbb{P}(\text{"vote"}|\text{"sport"}) &= \frac{3}{1000} \\
\mathbb{P}(\text{"sport"}|\text{"kick"}) &= \frac{10}{11} \\
\mathbb{P}(\text{"politique"}|\text{"kick"}) &= \frac{\mathbb{P}(\text{"kick"}|\text{"politique"}) \times \mathbb{P}(\text{"politique"})}{\mathbb{P}(\text{"kick"})} \\
&= \frac{\frac{1}{1000} \times \frac{2}{3}}{\frac{1}{3} \times \frac{2}{100} + \frac{2}{3} \times \frac{1}{1000}} \\
&= \frac{1}{11}
\end{aligned}$$

On a donc :

$$\begin{aligned}
\mathbb{P}(\text{mot ="vote"}|\text{mot="kick"}) &= \frac{4}{100} \times \frac{1}{11} + \frac{3}{1000} \times \frac{10}{11} \\
&= \frac{7}{1100}
\end{aligned}$$

(f) On se place dans un problème d'apprentissage supervisé. Pour connaitre la probabilité des sujets, on calcule le nombre de documents de type sport, puis le nombre de documents de types politique. En divisant ensuite ces deux quantités par le nombre de comuents total N, on obtient $\mathbb{P}$(sujet = "sport") et $\mathbb{P}$(sujet = "politique").
Pour connaitre les probabilités conditionnelles, il suffit, pour chacun des suejts sport et politique, de calculer la fréquence du mot recherché. On procède ensuite au calcul de la moyenne empirique des fréquences calculés pour chacun des deux tyeps de document. On aura ainsi, pour un mot donné, une moyenne empirique pour le sujet politique et une moyenne empirique pour le sujet sport.
Pour mieux illustrer ces propos, considérons l'exemple suivant :
Si on souhaite connaître $\mathbb{P}$(mot = "goal" | sujet = "politique"), on considère tous les documents de type politique. Pour chacun de ces documents, on calcule la fréquence du mot "goal". Enfin, on effectue la moyenne empirique de ces fréquences. Si par exemple, il y a 2 documents de type politique, et que 1 mot sur 500 est "goal" dans le premier tandis que 2 mots sur 1500 sont "goal"

sur le deuxième, on a alors :
$\mathbb{P}(\text{mot} = \text{“goal”} \mid \text{sujet} = \text{“politique”}) = \frac{\frac{1}{500}+\frac{2}{1500}}{2} = \frac{1}{600}$

3. **Estimateur de maximum de vraisemblance** [10 points]

Soit la fonction de densité de probabilité suivante:

$$f_\theta(x) = 2\theta x e^{-\theta x^2}$$

où θ est un paramètre et x est un nombre réel positif.
Supposons que n points $D = \{x_1, \ldots, x_n\}$ sont tirés aléatoirement indépendemment selon $f_\theta(x)$.

(a) Soit $f_\theta(x_1, x_2, \ldots, x_n)$ la fonction de densité de probabilité jointe de n points indépendemment et identiquement distribué (i.i.d) selon $f_\theta(x)$. Exprimez $f_\theta(x_1, x_2, \ldots, x_n)$ en fonction de $f_\theta(x_1)$, $f_\theta(x_2), \ldots, f_\theta(x_n)$

(b) On définit l'estimateur du maximum de vraisemblance comme la valeur de θ qui maximise la vraisemblance de générer le jeu de donnée D à partir de la distribution $f_\theta(x)$. Formellement,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n),$$

Calculez l'estimateur du maximum de vraisemblance de θ.

Réponses

(a) Soit $f_\theta(x_1, x_2, \ldots, x_n)$ la fonction de densité de probabilité jointe de n points indépendemment et identiquement distribué (i.i.d) selon $f_\theta(x)$.
On a :

$$\begin{aligned} f_\theta(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f_\theta(x_i) \\ &= f_\theta(x_1) \times f_\theta(x_2) \times \ldots \times f_\theta(x_n) \\ &= \prod_{i=1}^{n} 2\theta x_i \exp(-\theta x_i^2) \\ &= (2\theta)^n \exp(-\sum_{i=1}^{n} \theta x_i^2) \prod_{i=1}^{n} x_i \end{aligned}$$

(b) On définit l'estimateur du maximum de vraisemblance tel que :

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n)$$

Calculons θ_{MLE} :

$$\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n) \\
&= \arg\max_{\theta \in \mathbb{R}} \quad \prod_{i=1}^{n} f_\theta(x_i) \\
&= \arg\max_{\theta \in \mathbb{R}} \quad (2\theta)^n \exp(-\sum_{i=1}^{n} \theta x_i^2) \prod_{i=1}^{n} x_i
\end{aligned}$$

Pour nous simplifier le calcul, on considère la log-vraisemblance :

$$\begin{aligned}
\ln \prod_{i=1}^{n} f_\theta(x_i) &= \ln((2\theta)^n \exp(-\sum_{i=1}^{n} \theta x_i^2) \prod_{i=1}^{n} x_i) \\
&= n \ln 2\theta - \sum_{i=1}^{n} \theta x_i^2 + \sum_{i=1}^{n} \ln x_i
\end{aligned}$$

On résoud l'équation :

$$\frac{\partial \ln(\prod_{i=1}^{n} f_\theta(x_i))}{\partial \theta} = 0$$

$$\frac{\partial (n \ln 2\theta - \sum_{i=1}^{n} \theta x_i^2 + \sum_{i=1}^{n} \ln x_i)}{\partial \theta} = 0$$

On a donc :

$$\begin{aligned}
\frac{2n}{2\theta} - \sum_{i=1}^{n} x_i^2 &= 0 \\
\frac{n}{\theta} &= \sum_{i=1}^{n} x_i^2 \\
\theta &= \frac{n}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

De plus, on a bien : $\frac{\partial^2 \ln(\prod_{i=1}^{n} f_\theta(x_i))}{\partial \theta^2} < 0$. En effet :

$$\frac{\partial^2 \ln(\prod_{i=1}^{n} f_\theta(x_i))}{\partial \theta^2} = -\frac{n}{\theta^2} - \sum_{i=1}^{n} x_i^2 < 0$$

Ainsi, le maximum de vraisemblance calculé précedemment est bien un maximum de vraisemblance global.

4. **Estimateur du maximum de vraisemblance et histograme** [10 points]

Soit le jeu de données $X_1, X_2, \cdots, X_n$, composé de n données i.i.d tirées d'une distribution de probabilité constante par morceaux. Il y a N morceaux de longueur égale entre 0 et 1 $(B_1, B_2, \cdots, B_N)$, où les constantes sont $\theta_1, \theta_2, \cdots, \theta_N$.

$$p(x; \theta_1, \cdots, \theta_N) = \begin{cases} \theta_j & \frac{j-1}{N} \leq x < \frac{j}{N} \text{ for } j \in \{1, 2, \cdots, N\} \\ 0 & \text{otherwise} \end{cases}$$

On définit μ_j pour $j \in \{1, 2, \cdots, N\}$ en tant que $\mu_j \coloneqq \sum_{i=1}^{n} \mathbb{1}(X_i \in B_j)$.

(a) Utilisant le fait que l'aire totale sous une distribution de probabilité est 1, exprimez θ_N en fonction de $\theta_1, \theta_2, \cdots, \theta_{N-1}$.

(b) Écrivez le log-vraisemblance du jeu de données en fonction de $\theta_1, \theta_2, \cdots, \theta_{N-1}$ et $\mu_1, \mu_2, \cdots, \mu_{N-1}$.

(c) Trouvez l'estimateur du maximum de vraisemblance de θ_j, $j \in \{1, 2, \cdots, N\}$.

Réponses

(a) Sachant que l'aire totale sous une distribution de probabilité est 1, et que les "morceaux" ont une longueur égale à $\frac{1}{N}$.
On a alors : $\frac{1}{N} \sum_{i=1}^{N} \theta_i = 1$.
D'où :

$$\frac{1}{N}\theta_N = 1 - \frac{1}{N} \sum_{i=1}^{N-1} \theta_i$$

$$\theta_N = N - \sum_{i=1}^{N-1} \theta_i$$

(b) Ecriture du log-vraisemblance :

$$\begin{aligned}\ln(L(p(x_1,...,x_n,\theta_1,...,\theta_N)) &= \sum_{i=1}^{n} \ln(\theta_{j_{\{X_i\in B_j\}}}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{N} \mathbb{1}_{\{X_i\in B_j\}} \ln(\theta_j) \\ &= \sum_{j=1}^{N}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\in B_j\}} \ln(\theta_j) \\ &= \sum_{j=1}^{N} \mu_j \ln(\theta_j)\end{aligned}$$

Pour obtenir un résultat en fonction de $\theta_1, \theta_2, \cdots, \theta_{N-1}$ et $\mu_1, \mu_2, \cdots, \mu_{N-1}$:

$$\begin{aligned}\ln(L(p(x_1,...,x_n,\theta_1,...,\theta_N)) &= \sum_{j=1}^{N-1} \mu_j \ln(\theta_j) + \mu_N \ln(\theta_N) \\ &= \sum_{j=1}^{N-1} \mu_j \ln(\theta_j) + \mu_N \ln(N - \sum_{i=1}^{N-1} \theta_i)\end{aligned}$$

De plus, sachant que $\sum_{i=1}^{N} \mu_i = n$, on a $\mu_N = n - \sum_{i=1}^{N-1} \mu_i$. D'où le résultat :

$$\ln(L(p(x_1,...,x_n,\theta_1,...,\theta_N)) = \sum_{j=1}^{N-1} \mu_j \ln(\theta_j) + (n - \sum_{i=1}^{N-1} \mu_i) \ln(N - \sum_{i=1}^{N-1} \theta_i)$$

(c) Calcul du maximum de vraisemblance noté θ_k.
Calculons $\frac{\partial}{\partial\theta_k} \ln(L(p(x_1,...,x_n,\theta_1,...,\theta_N)$:

$$\begin{aligned}\frac{\partial}{\partial\theta_k} \ln(L(p(x_1,...,x_n,\theta_1,...,\theta_N) &= \mu_k \frac{\partial}{\partial\theta_k} \ln\theta_k + \frac{(n - \sum_{i=1}^{N-1} \mu_i)}{N - \sum_{i=1}^{N-1} \theta_i} \times -1 \\ &= \frac{\mu_k}{\theta_k} - \frac{n - \sum_{i=1}^{N-1} \mu_i}{N - \sum_{i=1}^{N-1} \theta_i}\end{aligned}$$

Résolvons :

$$\frac{\mu_k}{\theta_k} - \frac{n - \sum_{i=1}^{N-1} \mu_i}{N - \sum_{i=1}^{N-1} \theta_i} = 0$$

$$\frac{\mu_k}{\theta_k} = \frac{n - \sum_{i=1}^{N-1} \mu_i}{N - \sum_{i=1}^{N-1} \theta_i}$$

$$\theta_k = \frac{N - \sum_{i=1}^{N-1} \theta_i}{n - \sum_{i=1}^{N-1} \mu_i} \mu_k$$

On a donc $\forall k \in \{1, 2, ..., N\}$, il existe $c \in \mathbb{R}$ vérifiant :

$$\theta_k = c\ \mu_k$$

Or :

$$\frac{1}{N} \sum_{i=1}^{N} \theta_i = \frac{1}{N} \sum_{i=1}^{N} c\ \mu_i = 1$$

Alors, sachant que $\frac{1}{N} \sum_{i=1}^{N} \mu_i = n$:

$$\frac{c\ n}{N} = 1$$

$$c = \frac{N}{n}$$

Donc :

$$\theta_k = \frac{N}{n} \mu_k$$

De plus, on a bien : $\frac{\partial^2 \ln(L(p(x_1,...,x_n,\theta_1,...,\theta_N)}{\partial \theta^2} < 0$. En effet :

$$\frac{\partial^2 \ln(L(p(x_1, ..., x_n, \theta_1, ..., \theta_N))}{\partial \theta^2} = -\frac{\mu_k}{\theta_k^2} - (n - \sum_{i=1}^{N-1} \mu_i) \frac{\theta_k}{(N - \sum_{i=1}^{N-1} \theta_i)^2} < 0$$

Ainsi, le maximum de vraisemblance calculé précedemment est bien un maximum de vraisemblance global.

5. **Méthodes à base d'histogrames** [10 points]

Soit le jeu de données $\{x_j\}_{j=1}^{n}$ où chaque point $x \in [0, 1]^d$. La fonction $f(x)$ représente la vraie distribution inconnue des données. Vous décidez d'utiliser une méthode à base d'histogramme pour estimer $f(x)$. Chaque dimension est divisée en m régions.

(a) Démontrez que pour un ensemble mesurable S, $\mathbb{E}[\mathbb{1}_{\{x \in S\}}] = \mathbb{P}(x \in S)$.

(b) En combinant le résultat de la question précédente avec la loi des grands nombres, démontrez que la probabilité estimée d'être dans la région i, telle que donnée par les méthodes à base d'histogramme, tend vers $\int_{V_i} f(x)dx$, la vraie probabilité d'être dans la région i, lorsque $n \to \infty$. V_i est le volume occupée par la région i.

(c) Soit le jeu de données MNIST, où chaque point vit en 784 dimensions. Nous divisons chaque dimensions en 2 régions. Combien de chiffres (en base 10) contient le nombre total de régions?

(d) Supposons l'existence d'un classificateur MNIST idéalisé basé sur un histogramme de $m = 2$ régions par dimension. La précision du classificateur augmente de $\epsilon = 5\%$ (à partir de 10% et jusqu'à un maximum de 100%) chaque fois que $k = 4$ nouveaux points sont ajoutés à chaque région. Quel est le plus petit nombre de points dont le classificateur aurait besoin pour atteindre une précision de 90% ?

(e) En supposant une distribution uniforme sur toutes les régions, quelle est la probabilité qu'une région en particulier est vide, en fonction de d, m et n?

Notez le contraste entre (b) et (e): même si pour une infinité de points de données, l'histogramme sera arbitrairement précis pour estimer la vraie distribution (b), en pratique le nombre d'échantillons requis pour obtenir même un seul point de données dans chaque région croît de façon exponentielle avec d.

Réponses

(a) Montrons que pour un ensemble mesurable S, on a $\mathbb{E}[\mathbb{1}_{\{x \in S\}}] = \mathbb{P}(x \in S)$.
On sait que par définition :

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{\{x \in S\}}] &= \int_{x \in [0,1]^d} \mathbb{1}_{\{x \in S\}} f(x)\, \mathrm{d}x \\
&= \int_{x \in S} 1 \times f(x)\, \mathrm{d}x + \int_{x \notin S} 0 \times f(x)\, \mathrm{d}x \\
&= \int_{x \in S} f(x)\, \mathrm{d}x \\
&= \mathbb{P}(x \in S) \qquad \text{par définition d'une loi de probabilité absolument continue}
\end{aligned}$$

(b) Loi des grands nombres :
La moyenne empirique calculée sur un échantillon, notée M, converge vers l'espérance lorsque la taille de l'échantillon tend vers l'infini.
On a alors :

$$\begin{aligned}\lim_{x\to+\infty} M(x \in i) &= \mathbb{E}[\mathbb{1}_{\{x\in i\}}] \\ &= \mathbb{P}(x \in i) \qquad \text{(question 1)} \\ &= \int_{x\in i} f(x)\,\mathrm{d}x \qquad \text{par définition} \\ &= \int_{V_i} f(x)\,\mathrm{d}x \qquad \text{avec } V_i \text{ est le volume occupée par la région } i\end{aligned}$$

(c) Le nombre total de régions est de 2^{784}. Ce nombre possède 237 chiffres en base 10 :

$$2^{784} = 2^{log_2(10)\frac{784}{log_2(10)}} = 10^{\frac{784}{log_2(10)}} \approx 10^{237}$$

(d) La précision du classificateur augmente de 5% à chaque fois que 4 nouveaux points sont ajoutés à chaque région. Donc, pour un gain de 5% de précision, il faut ajouter $4 \times 2^{784} = 2^{786}$ points. Or, on cherche ici un gain de 80% de précision, soit 16 fois un gain de 5%. Le plus petit nombre de point dont le classificateur aurait besoin pour atteindre une précision de 90% est donc de : $2^{786} \times 16 = 2^{790}$ points.

(e) On suppose une distribution uniforme sur toutes les régions. On cherche à calculer la probabilité qu'une région en particulier est vide :

$$\mathbb{P}(\forall\{x_j\}_{j=1}^n \notin i) = \bigcap_{j=1}^{n} \mathbb{P}(x_j \notin i)$$

En effet, on cherche la probabilité qu'aucun point $\{x_j\}_{j=1}^n$ n'appartienne à la région i. Sachant que les variables sont indépendantes et identiquement distribuées, on a alors :

$$\begin{aligned}\bigcap_{j=1}^{n} \mathbb{P}(x_j \notin i) &= \prod_{j=1}^{n} \mathbb{P}(x_j \notin i) \\ &= \mathbb{P}(x_1 \notin i) \times \mathbb{P}(x_2 \notin i) \times ... \times \mathbb{P}(x_n \notin i)\end{aligned}$$

De plus, on sait que la distribution est uniforme, donc:
$\mathbb{P}(x_1 \notin i) = \mathbb{P}(x_2 \notin i) = ... = \mathbb{P}(x_n \notin i)$. D'où :

$$\begin{aligned}\mathbb{P}(x_1 \notin i) \times \mathbb{P}(x_2 \notin i) \times ... \times \mathbb{P}(x_n \notin i) &= \mathbb{P}(x_1 \notin i)^n \\ &= (1 - \mathbb{P}(x_1 \notin i))^n\end{aligned}$$

Or la probabilité d'être dans une région i est égale à $\frac{1}{m^d}$ (puisqu'on considère une distribution uniforme sur toutes les régions) avec m^d le nombre total de régions. Donc :

$$\begin{aligned}\mathbb{P}(\forall \{x_j\}_{j=1}^n \notin i) &= (1 - \mathbb{P}(x_1 \notin i))^n \\ &= (1 - \frac{1}{m^d})^n\end{aligned}$$

Donc la probabilité qu'une région en particulier soit vide est égale à $(1 - \frac{1}{m^d})^n$.

6. **Mélange de Gaussiennes** [10 points]

Soit $\mu_0, \mu_1 \in \mathbb{R}^d$ et soit Σ_0, Σ_1 deux matrices $d \times d$ positives définies (i.e. symmétriques avec des valeurs propres strictement positives).
On définie maintenant les deux fonctions de densités de probabilités suivantes sur $\mathbb{R}^2$:

$$f_{\mu_0,\Sigma_0}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}$$

$$f_{\mu_1,\Sigma_1}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}$$

Ces fonctions de densité de probabilités correspondent à la distribution gaussienne multivariée de centre μ_0 et covariance Σ_0, notée $\mathcal{N}_d(\mu_0, \Sigma_0)$ et à la gaussienne multivariée de centre μ_1 et covariance Σ_1, notée $\mathcal{N}_d(\mu_1, \Sigma_1)$.

On lance maintenant une pièce équilibrée Y et on tire une variable aléatoire X dans $\mathbb{R}^d$, en suivant cette procédure : si la pièce atterit sur pile ($Y = 0$), on tire X selon $\mathcal{N}_d(\mu_0, \Sigma_0)$ et si la pièce atterit sur face ($Y = 1$), on tire X selon $\mathcal{N}_d(\mu_1, \Sigma_1)$.

(a) Calculez $\mathbb{P}(Y = 0|X = \mathbf{x})$, la probabilité que la pièce atterisse sur pile sachant $X = \mathbf{x} \in \mathbb{R}^2$, en fonction de μ_0, μ_1, Σ_0, Σ_1 et $\mathbf{x}$. Montrez toutes les étapes du calcul.

(b) Rappelez-vous que le classifieur optimale Bayes est $h_{Bayes}(\mathbf{x}) = \underset{y\in\{0,1\}}{\text{argmax}}\ \mathbb{P}(Y = y|X = \mathbf{x})$. Démontrez que si $\Sigma_0 = \Sigma_1$, le classifieur optimal Bayes est linéaire en $\mathbf{x}$.

Réponses

(a) Calcul de $\mathbb{P}(Y = 0|X = \mathbf{x})$:

$$\begin{aligned}\mathbb{P}(Y = 0|X = \mathbf{x}) &= \frac{\mathbb{P}(X = \mathbf{x}, Y = 0)}{\mathbb{P}(X = \mathbf{x})} \\ &= \frac{\mathbb{P}(X = \mathbf{x}|Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(Y = 0)\mathbb{P}(X = \mathbf{x}|Y = 0) + \mathbb{P}(Y = 1)\mathbb{P}(X = \mathbf{x}|Y = 1)}\end{aligned}$$

Cette expression est égale à :

$$\begin{aligned}&\frac{\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)} + \frac{1}{2} \times \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_1)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}} \\ &= \frac{\frac{1}{\sqrt{det(\Sigma_0)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}}{\frac{1}{\sqrt{det(\Sigma_0)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)} + \frac{1}{\sqrt{det(\Sigma_1)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}} \\ &= \frac{1}{1 + \frac{\frac{1}{\sqrt{det(\Sigma_1)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}}{\frac{1}{\sqrt{det(\Sigma_0)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}}} \\ &= \frac{1}{1 + \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}}{\sqrt{det(\Sigma_1)}} \times \frac{\sqrt{det(\Sigma_0)}}{e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}}}\end{aligned}$$

D'où le résultat final :

$$\mathbb{P}(Y = 0|X = \mathbf{x}) = \frac{1}{1 + \frac{\sqrt{det(\Sigma_0)}}{\sqrt{det(\Sigma_1)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)+\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}}$$

(b) On a d'après la question précédente :

$$\mathbb{P}(Y = 0|X = \mathbf{x}) = \frac{1}{1 + \frac{\sqrt{det(\Sigma_0)}}{\sqrt{det(\Sigma_1)}}e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)+\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}}$$

On s'intéresse à cette probabilité si $\Sigma_0 = \Sigma_1$.
Dans ce cas là, on a :

$$\mathbb{P}(Y = 0|X = \mathbf{x}) = \frac{1}{1 + e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)+\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}}$$

Intéressons nous au terme dans l'exponentielle :

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0) \\
&= -\frac{1}{2}(\mathbf{x}^T\Sigma_1^{-1} - \mu_1^T\Sigma_1^{-1})(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}^T\Sigma_0^{-1} - \mu_0^T\Sigma_0^{-1})(\mathbf{x}-\mu_0) \\
&= \frac{1}{2}(-\mathbf{x}^T\Sigma_1^{-1}\mathbf{x} + \mathbf{x}^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mathbf{x} - \mu_1^T\Sigma_1^{-1}\mu_1) \\
&\quad + \frac{1}{2}(\mathbf{x}^T\Sigma_0^{-1}\mathbf{x} - \mathbf{x}^T\Sigma_0^{-1}\mu_0 - \mu_0^T\Sigma_0^{-1}\mathbf{x} + \mu_0^T\Sigma_0^{-1}\mu_0)
\end{aligned}$$

En développant on obtient l'expression suivante :

$$\begin{aligned}
&\frac{1}{2}(-\mathbf{x}^T\Sigma_1^{-1}\mathbf{x} + \mathbf{x}^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mathbf{x} - \mu_1^T\Sigma_1^{-1}\mu_1 \\
&\qquad + \mathbf{x}^T\Sigma_0^{-1}\mathbf{x} - \mathbf{x}^T\Sigma_0^{-1}\mu_0 - \mu_0^T\Sigma_0^{-1}\mathbf{x} + \mu_0^T\Sigma_0^{-1}\mu_0)
\end{aligned}$$

Or, rappelons qu'on considère ici que $\Sigma_0 = \Sigma_1$. On obtient alors en simplifiant:

$$\begin{aligned}
&\frac{1}{2}(\mathbf{x}^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mathbf{x} - \mu_1^T\Sigma_1^{-1}\mu_1 - \mathbf{x}^T\Sigma_0^{-1}\mu_0 - \mu_0^T\Sigma_0^{-1}\mathbf{x} + \mu_0^T\Sigma_0^{-1}\mu_0) \\
&= \frac{1}{2}(2\mu_1\Sigma_1^{-1}\mathbf{x} - 2\mu_0\Sigma_0^{-1}\mathbf{x} + \mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1) \\
&= (\mu_1\Sigma_1^{-1} - \mu_0\Sigma_0^{-1})\mathbf{x} + \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1)
\end{aligned}$$

En posant :

$$\begin{aligned}
-\mathbf{w}^T &= \mu_1\Sigma_1^{-1} - \mu_0\Sigma_0^{-1} \\
-\mathbf{b} &= \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1)
\end{aligned}$$

On a alors :

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0) \\
&= -\mathbf{w}^T\mathbf{x} - \mathbf{b}
\end{aligned}$$

On obtient donc :

$$\begin{aligned}
\mathbb{P}(Y=0|X=\mathbf{x}) &= \frac{1}{1+e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)+\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}} \\
&= \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}-\mathbf{b}}} \\
&= \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+\mathbf{b})}} \\
&= \text{sigmoid}(\mathbf{w}^T\mathbf{x}+\mathbf{b})
\end{aligned}$$

Ainsi, avec :

$$\mathbb{P}(Y=1|X=\mathbf{x}) = 1-\text{sigmoid}(\mathbf{w}^T\mathbf{x}+\mathbf{b})$$

On a l'expression suivante :

$$\begin{aligned}
h_{Bayes}(\mathbf{x}) &= \underset{y\in\{0,1\}}{\text{argmax}}\ \mathbb{P}(Y=y|X=\mathbf{x}) \\
&= \underset{y\in\{0,1\}}{\text{argmax}}\ (\text{sigmoid}(\mathbf{w}^T\mathbf{x}+\mathbf{b}), 1-\text{sigmoid}(\mathbf{w}^T\mathbf{x}+\mathbf{b}))
\end{aligned}$$

Montrer que le classifieur optimal de Bayes est linéaire en $\mathbf{x}$ revient à montrer que la frontière de décision est linéaire en $\mathbf{x}$.
La frontière de décision se situe à la frontière où :

$$\mathbb{P}(Y=0|X=\mathbf{x}) = \mathbb{P}(Y=1|X=\mathbf{x})$$

.
Sachant qu'il n'y a que deux évènements possibles (obtenir pile ou obtenir face), on a sur la frontière de décision :

$$\mathbb{P}(Y=0|X=\mathbf{x}) = \mathbb{P}(Y=1|X=\mathbf{x}) = \frac{1}{2}$$

Donc, trouver la frontière de décision revient à résoudre :

$$\begin{aligned}
& \mathbb{P}(Y=0|X=\mathbf{x}) && = \frac{1}{2} \\
\iff\ & \text{sigmoid}(\mathbf{w}^T\mathbf{x}+\mathbf{b}) && = \frac{1}{2} \\
\iff\ & \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+\mathbf{b})}} && = \frac{1}{2} \\
\iff\ & 1+e^{-(\mathbf{w}^T\mathbf{x}+\mathbf{b})} && = 2 \\
\iff\ & e^{-(\mathbf{w}^T\mathbf{x}+\mathbf{b})} && = 1 \\
\iff\ & -(\mathbf{w}^T\mathbf{x}+\mathbf{b}) && = 0 \\
\iff\ & \mathbf{w}^T\mathbf{x}+\mathbf{b} && = 0
\end{aligned}$$

On a donc :

$$\mathbb{P}(Y = 0|X = \mathbf{x}) = \frac{1}{2} \iff \mathbf{w}^T\mathbf{x} + \mathbf{b} = 0$$

Ainsi, la frontière de décision est linéaire en $\mathbf{x}$ si $\Sigma_0 = \Sigma_1$.
Le classifieur optimal Bayes est donc linéaire en $\mathbf{x}$ si $\Sigma_0 = \Sigma_1$.