Master 2 informatique

Parcours IAAA, UE MIA

Feuille d'exercices numéro 2

Année 2024-25

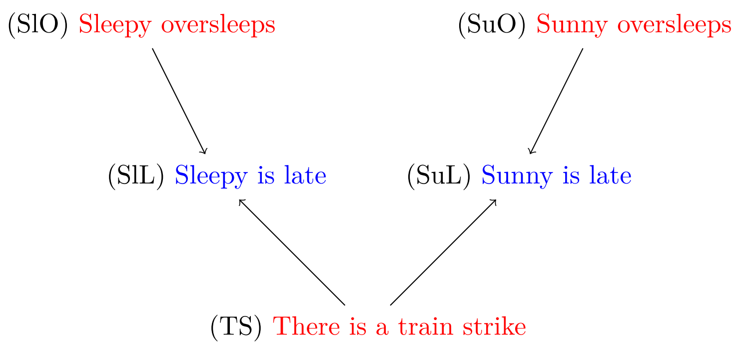
Statistique et probabilités.

1 Calcul probabiliste

Exercice 1 Propriétés élémentaires. ($\star$)

1. Prouvez que si A et B sont des événements indépendants, $P(A|B) = P(A)$.
2. On considère une fonction f de $\mathbf{R}$ vers $\mathbf{R}$. Quelles conditions doit-elle vérifier pour être une fonction de répartition valide ? Une densité de probabilité valide ?
3. Représentez graphiquement la fonction de masse, la densité de probabilité et la fonction de répartition pour une variable aléatoire X uniforme sur $[0, 1]$.
4. Même question pour X valant 1 avec probabilité .5, 2 avec probabilité .3 et 4 avec probabilité .2.
5. Un coffre A contient 100 pièces d'or. Un coffre B contient 60 pièces d'or et 40 pièces d'argent. Vous choisissez un coffre aléatoirement selon une loi uniforme et tirez une pièce aléatoirement selon une loi uniforme dans ce coffre. Si la pièce est en or, quelle est la probabilité que vous ayez choisi le coffre A ?

Exercice 2 Calcul probabiliste. ($\star\star$)



Considérons le modèle graphique représenté ci-dessus. SlO, SuO, SlL, SuL and TS sont des variables aléatoire binaires prenant leurs valeurs dans $\{0, 1\}$. Dans cet exercice nous allons essayer de déterminer ce qui peut être inféré sur les variables latentes (en rouge) à partir de l'observation des variables observées (en bleu).

Nous faisons l'hypothèse que si au moins un des deux évènements "Sleepy oversleeps" et "There is a train strike" a lieu, alors 'Sleepy is late' a lieu également (avec probabilité 1). De façon similaire, si au moins un des deux évènements 'Sunny oversleeps' et 'There is a train strike' a lieu, alors 'Sunny is late' a lieu également (avec probabilité 1). Nous pouvons l'écrire plus formellement, de la manière suivante :

$$P(SlL = 1|SlO = a, TS = b) = a \vee b,$$

et :

$$P(SuL = 1|SuO = a, TS = b) = a \vee b,$$

pour tout a, b dans $\{0, 1\}$. Le symbole $\vee$ représente le connecteur logique *ou* (inclusif) de $\{0, 1\}$ vers $\{0, 1\}$.

On note $l = P(SlO = 1)$, $u = P(SuO = 1)$ et $t = P(TS = 1)$.

1. Donner la factorisation de $P(SlL, SuL, SlO, SuO, TS)$ d'après le modèle graphique représenté ci-dessus.
2. La distribution de probabilité $P(SlL, SuL, SlO, SuO, TS)$ est-elle entièrement déterminée si les valeurs de l, u et t sont données ?
3. Calculer $P(TS = 1|SlL = 1)$ en fonction de l, u et t.
4. Calculer $P(SlO = 1|SlL = 1)$ en fonction de l, u et t.
5. Calculer $P(TS = 1|SlL = 1, SuL = 1)$ en fonction de l, u et t.
6. Calculer $P(SlO = 1|SlL = 1, SuL = 1)$ en fonction de l, u et t.
7. Supposer à présent que $l = 0.5$, $t = 0.1$ et que l'évenement 'Sleepy is late' a lieu. Quel évènement est alors le plus probable : 'There is a train strike' ou 'Sleepy overslept' ?
8. Même question si on suppose en plus que $u = 0.01$ et que l'évènement 'Sunny is late' est également observé.
9. Que se passe-t-il si on prend $l = 0.5$, $t = 0.1$ et $u = 0.2$?

2 Calcul de moments

Exercice 3 Preuve de propriétés vues en cours. ($\star\star$)

1. Montrez que $\text{Var}(X) = E[X^2] - E[X]^2$
2. Prouvez la loi de l'espérance totale.

3. Prouvez la loi de la variance totale.

Exercice 4 Espérance et variance d'estimateurs classiques. ($\star\star$)

Soit n un entier naturel et $X_1, X_2, \ldots, X_n$ des variables aléatoires réelles mutuellement indépendantes. Par souci de simplicité, on suppose que tous les moments des ces variables aléatoires existent. On note

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

et

$$\hat{V} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

1. Montrez que
$$\hat{V} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) - \hat{\mu}^2$$
2. Exprimez l'espérance de $\hat{\mu}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \ldots, X_n$ (prouvez votre résultat)

On suppose à présent que $X_1, X_2, \ldots, X_n$ suivent toute la même distribution (elles sont donc i.i.d. puisqu'on a déjà supposé qu'elles étaient indépendantes).

3. Répondez à nouveau à la question précédente dans ce cadre plus simple.
4. Exprimez la variance de $\hat{\mu}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \ldots, X_n$ (prouvez votre résultat)
5. Exprimez l'espérance de $\hat{V}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \ldots, X_n$ (prouvez votre résultat)
6. Exprimez la variance de $\hat{V}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \ldots, X_n$ (prouvez votre résultat)

Exercice 5 Loi de l'espérance totale. ($\star$)

Un téléphone kadok tient en moyenne $12h$ avec la batterie A, mais seulement $8h$ avec la batterie B. La batterie A se trouve dans 80% des téléphones kadok, le reste étant muni de la batterie B. Si vous achetez un téléphone kadok, combien d'heure vous attendez-vous à ce qu'il tienne ?

3 Application à des problèmes d'estimation

Exercice 6 Calculs de biais, variance, risque. ($\star$)

1. Calculer le biais et la variance de l'estimateur de la moyenne empirique pour un échantillon i.i.d.
2. Calculer le biais de l'estimateur de la variance suivant pour un échantillon i.i.d. : $\hat{V} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$
3. Proposez un estimateur non-biaisé de la variance pour un échantillon i.i.d.
4. Comparez le risque quadratique moyen des deux estimateurs de la variance.
5. Pouvez-vous donner un estimateur avec un risque plus faible que les deux considérés jusqu'ici ?

Exercice 7 Estimation par maximum de vraisemblance des paramètres d'une loi Gaussienne multivariée. ($\star$)

Supposons qu'on observe un échantillon i.i.d. $x_1, x_2, \ldots, x_n \in (\mathbf{R}^d)^n$ de loi gaussienne multivariée $\mathcal{N}(\mu^*, \Sigma^*)$, pour un vecteur $\mu \in \mathbf{R}^d$ et une matrice $\Sigma^* \in \mathbf{S}_d$, où $\mathbf{S}_d$ est l'ensemble des matrices à coefficients réels symétriques définies positives de taille d pas d. On cherche à estimer μ^* et Σ^* à partir de l'observation de $x_1, x_2, \ldots, x_n$.

On définit la *vraisemblance* d'un couple de paramètres $(\mu, \Sigma) \in \mathbf{R}^d \times \mathbf{S}_d$ comme :

$$\ell(\mu, \Sigma) := p(x_1, x_2, \ldots, x_n; \mu, \Sigma),$$

où p corresponds à la densité de probabilité de $x_1, x_2, \ldots, x_n$.

On considère l'estimateur du *maximum de vraisemblance* pour μ^*, Σ^* défini par

$$\hat{\mu}, \hat{\Sigma} \in \arg \max_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \ell(\mu, \Sigma).$$

1. Montrer que
$$\hat{\mu}, \hat{\Sigma} \in \arg \max_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \log(\ell(\mu, \Sigma)).$$
2. Donner une expression (la plus simple que vous pouvez) pour $\log(\ell(\mu, \Sigma)))$ en utilisant la formule donnant la densité d'une loi gaussienne multivariée non dégénérée.
3. On suppose que la matrice de covariance empirique $\frac{1}{n}(x_i - \mu)(x_i - \mu)^T$ est inversible et on admet que $\log(\ell(\mu, \Sigma)))$ est de classe C^1 et admet un unique maximum sur $\mathbf{R}^d \times \mathbf{S}_d$ en un point où son gradient s'annule. Calculer une expression explicite pour $(\hat{\mu}, \hat{\Sigma})$ en fonction de $x_1, x_2, \ldots, x_n$. Vous pouvez utiliser les identités de calcul différentiel matriciel suivantes

sans les démontrer :

$$\frac{\partial u^T M^{-1} v}{\partial M} = -(M^{-1})^T u v^T (M^{-1})^T,$$

$$\frac{\partial \det(M)}{\partial M} = \det(M)(M^{-1})^T,$$

où M est une matrice inversible et u et v sont des matrices colonnes de dimension compatible avec M (par exemple si M est de taille n par n, u et v sont de taille n par 1).

4. Montrer que le couple $(\hat{\mu}, \hat{\Sigma})$ ainsi obtenu forme une statistique suffisante pour le couple de paramètres (μ^*, Σ^*).

Exercice 8 Analyse des propriétés basiques d'un estimateur. ($\star\star$)

Supposons qu'on observe un échantillon $x_1, x_2, \ldots, x_n \in (\mathbf{R}^d)^n$, i.i.d. de distribution P. Soit $d : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$ une fonction mesurant la 'dissimilarité' entre deux points de $\mathbf{R}^d$. On suppose que d est symmétrique, c'est à dire que $d(x_1, x_2) = d(x_2, x_1)$ pour tout choix de x_1, x_2. Nous cherchons à estimer la dissimilarité moyenne entre deux points tirés aléatoirement et indépendamment suivant la distribution P :

$$\delta(P, d) := \mathbf{E}_{a,b \sim P \otimes P}[d(a, b)]$$

sur la base de $x_1, x_2, ..., x_n$ (la notation $a, b \sim P \otimes P$ signifie que a et b sont deux échantillons tirés indépendamment de la loi P). On suppose que $\mathbf{E}_{a,b \sim P \otimes P}[d(a, b)^2] < +\infty$.

Considérons l'estimateur :

$$\hat{\delta}(x_1, x_2, \ldots, x_n) := \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(x_i, x_j).$$

1. Quel est le biais de $\hat{\delta}$?
2. Quelle est la variance de $\hat{\delta}$? On l'exprimera en fonction de $\sigma_1^2 = \mathrm{Var}_{x_1 \sim P} \mathbf{E}_{x_2 \sim P}[d(x_1, x_2)]$ et $\sigma_2^2 = \mathrm{Var}_{(x_1, x_2) \sim P \otimes P}[d(x_1, x_2)]$.
3. Prouver l'inégalité de Markov : soit X est un variable aléatoire réelle positive, de moyenne finie μ et soit t un réel strictement positif, alors :

$$p(X \geq t) \leq \frac{\mu}{t}.$$

4. Utiliser l'inégalité de Markov pour prouver l'inégalité de Chebyshev : soit X est un variable aléatoire réelle de moyenne finie μ et de variance finie σ^2 et soit t un réel strictement positif,

alors :

$$p(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

5. Utiliser l'inégalité de Chebyshev et les résultats des deux premières questions pour montrer que $\hat{\delta}$ est un estimateur *faiblement consistant* de δ, c'est à dire qu'on a $\hat{\delta} \to_p \delta$, c'est à dire que pour tout $\epsilon > 0$:

$$\lim_{n \to +\infty} p(|\hat{\delta}(x_1, \dots x_n) - \delta| \geq \epsilon) = 0.$$

Exercice 9 Loi de l'espérance totale. ($\star$)

Un téléphone kadok tient en moyenne $12h$ avec la batterie A, mais seulement $8h$ avec la batterie B. La batterie A se trouve dans 80% des téléphones kadok, le reste étant muni de la batterie B. Si vous achetez un téléphone kadok, combien d'heure vous attendez-vous à ce qu'il tienne ?