Aix*Marseille université

Master 2 informatique

Parcours IAAA, UE MIA

Feuille d'exercices numéro 5

Année 2023-24

Statistiques.

Exercice 1 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 et la variance 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. Trouvez un estimateur non-biaisé de la variance pour un échantillon i.i.d.
4. Prouvez la décomposition biais-variance pour le risque quadratique moyen
5. Comparez le risque quadratique moyen des deux estimateurs de la variance.
6. Pouvez-vous donner un estimateur avec un risque plus faible que les deux considérés jusqu'ici ?
7. On s'intéresse à présent au problème d'estimer la variance de l'estimateur de la moyenne empirique. Que pouvons-nous déduire des questions précédentes à ce sujet ?

Exercice 2 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 3 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.

Exercice 4 Statistique asymptotique. ($\star$)

1. Montrer que la moyenne empirique est un estimateur fortement consistant de μ pour un échantillon $X_1, X_2, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$, c'est à dire que la moyenne empirique tend presque sûrement vers μ.
2. Montrer que l'estimateur usuel de la variance est fortement consistant.
3. Donnez une expression pour la distribution asymptotique de la norme L2 de la moyenne empirique de vecteurs aléatoires $X_1, X_2, \ldots, X_n$, i.i.d. de loi P. On suppose que les moments d'ordre 1 et 2 de P existent.

Exercice 5 Intervalles de confiance. ($\star\star$)

On considère des variables aléatoires rélles X_1, X_2,..., X_n i.i.d. de loi P et de variance :

$$\mathrm{Var}(P) = \sigma^2 < +\infty.$$

On suppose de plus que X_1, X_2,..., X_n prennent leur valeurs dans l'intervalle $[0, 1]$ et que la fonction de répartition de P est continue.

On définit un intervalle de confiance asymptotique pour $\mathbf{E}(P)$ de niveau $1 - \alpha$, pour $\alpha \in]0; 1[$, comme une région (i.e. un sous-ensemble) R_n de la droite réelle telle que :

$$\lim_{n\to+\infty} p\left(\mathbf{E}(P) \in R_n\right) \geq 1 - \alpha.$$

1. En vous appuyant sur le théorème de la limite centrale, donnez l'expression d'un intervalle de confiance asymptotique de niveau $1-\alpha$ pour $\mathbf{E}(P)$ utilisant l'estimateur de la moyenne empirique $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$ et prouvez que l'intervalle proposé est effectivement asymptotiquement de niveau $1-\alpha$.
2. En vous appuyant sur l'inégalité de concentration de Hoeffding (`https://fr.wikipedia.org/wiki/In%C3%A9galit%C3%A9_de_Hoeffding`), donnez l'expression d'un intervalle de confiance (non-asymptotique) de niveau $1-\alpha$ pour $\mathbf{E}(P)$ (toujours sur la base de $\hat{\mu}$) et prouvez que l'intervalle proposé est de niveau $1-\alpha$.
3. Comparez la largeur des intervalles obtenus par les deux méthodes et commentez.