1 – Estimation de densité

On considère un ensemble de données $D=\{x^{(1)},\dots,x^{(n)}\}$ avec $x\in\mathbb{R}^d$.

1. Estimateur de densité avec des fenêtres de Parzen à noyau Gaussien isotropique :

a) L'équation du noyau correspondant à une densité Gaussienne isotropique est :

$$K(x;x^{(i)})=N_{x^{(i)},\sigma^2}(x)=\frac{1}{(2\pi)^{d/2}\sigma^d}e^{-\frac{\|x-x^{(i)}\|^2}{2\sigma^2}}$$

b) Équation de la densité estimée par l'estimateur de densité de Parzen en un nouveau point x :

$$p(x)=\frac{1}{n}\sum_{i=1}^{n}N_{x^{(i)},\sigma^2}(x)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{(2\pi)^{d/2}\sigma^d}e^{-\frac{\|x-x^{(i)}\|^2}{2\sigma^2}}\right)=\frac{1}{n(2\pi)^{d/2}\sigma^d}\sum_{i=1}^{n}e^{-\frac{\|x-x^{(i)}\|^2}{2\sigma^2}}$$

c) Paramètres et hyper-paramètres :

Paramètre :

- Les données $x^{(i)}$ de l'ensemble d'entraînement à mémoriser ($n\times\mathbb{R}^d$).

Hyper-paramètre :

- L'écart-type $\sigma\in\mathbb{R}$ de la loi normale devra être fixé.

d) Prouver que la densité estimée vérifie les propriétés d'une densité de probabilité :

2. Estimation de densité paramétrique avec une densité Gaussienne isotropique :

a) Équation paramétrique de la densité estimée en un nouveau point x :

$$p(x) = N_{\mu,\sigma^2}(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

b) Différence entre les estimations de densité des numéros 1 et 2 :

Dans le cas de l'estimation par fenêtre de Parzen, il nous faut effectuer une moyenne sur n densités Gaussiennes et, pour chacune d'elles, le point $x^{(i)}$ sert de moyenne ($N_{x_{(i)},\sigma^2}$). Dans le cas présent, on utilise une seule densité Gaussienne pour laquelle il faudra apprendre les valeurs de moyenne μ et d'écart-type σ .

c) Paramètres et hyper-paramètres :

Paramètres :

- La moyenne $\mu \in \mathbb{R}^d$
- L'écart-type $\sigma \in \mathbb{R}$

Hyper-paramètre : Aucun

3. Estimation de densité paramétrique avec une densité Gaussienne diagonale:

a) Équation d'une densité Gaussienne diagonale, paramètres et hyper-paramètres :

Soit la matrice de covariance :

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{pmatrix}$$

L'équation paramétrique de la densité estimée en un nouveau point x est alors :

$$p(x) = N_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Paramètres :

- La moyenne $\mu \in \mathbb{R}^d$
- Les valeurs contenu sur la diagonale de la matrice de covariance Σ ; donc $d \times \mathbb{R}$.

Hyper-paramètre : Aucun

b) Indépendance des composantes d'un vecteur aléatoire suivant une distribution Gaussienne diagonale :

Soit $x \in \mathbb{R}^d$ un vecteur aléatoire qui suit une distribution Gaussienne diagonale $N_{\mu,\Sigma}(x)$.

Posons $x-\mu=k$ et $\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}=s$.

Puisque $\Sigma=\begin{pmatrix}\sigma_1^2 & 0 & ... & 0\\ 0 & \sigma_2^2 & ... & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & ... & \sigma_d^2\end{pmatrix}$, alors $\Sigma^{-1}=\begin{pmatrix}1/\sigma_1^2 & 0 & ... & 0\\ 0 & 1/\sigma_2^2 & ... & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & ... & 1/\sigma_d^2\end{pmatrix}$

$$p(x)=N_{\mu,\Sigma}(x)=\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}=s\times e^{-\frac{1}{2}k^t\Sigma^{-1}k}$$

$$=s\times e^{-\frac{1}{2}(k_1/\sigma_1^2,k_2/\sigma_2^2,...,k_d/\sigma_d^2)k}=s\times e^{-\frac{1}{2}\sum_{i=1}^{d}k_i^2/\sigma_i^2}$$

$$=s\prod_{i=1}^{d}e^{-\frac{1}{2}k_i^2/\sigma_i^2}=s\prod_{i=1}^{d}e^{-\frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)}$$

où $x^{(i)}$ est un vecteur de taille d avec la i^{ème} valeur de x en position i et des 0 partout ailleurs.

$$=\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\prod_{i=1}^{d}e^{-\frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)}$$

$$=\prod_{i=1}^{d}p(x^{(i)})$$

La densité estimée pour à partir d'un vecteur aléatoire $x\in\mathbb{R}^d$ qui suit une distribution Gaussienne diagonale est égale au produit des densités estimées des composantes de ce vecteur aléatoire. On peut en conclure que ces composantes sont des variables aléatoires indépendantes.

c) Minimisation du risque empirique :

d)

4. Sélection de modèle :

La première tâche à effectuer pour choisir le modèle le plus approprié est de sélectionner un ensemble de tous les modèles que nous pourrions effectuer. Ici, il s'agit des approches présentées aux sections 1.1, 1.2 et 1.3. Ensuite, pour chacun des modèles sélectionnées, on procède ainsi :

1. Soit A, le modèle courant.
2. Si le modèle contient des hyper-paramètres, répéter la procédure pour toutes les valeurs acceptables que ces hyper-paramètres peuvent prendre. Appelons cette configuration λ. Rappelons que l'approche des fenêtres de Parzen à noyau isotropique contient un hyper-paramètre (l'écart-type σ) tandis que les approches 2 et 3 n'en n'ont aucun ($\lambda=\emptyset$).
3. Mélanger les éléments de l'ensemble D s'ils étaient ordonnés.
4. Séparer l'ensemble d'entraînement D en deux sous-ensembles D_{train} et D_{valid}. L'ensemble d'entraînement devrait être plus grand que l'ensemble de validation.

5. Entraîner A avec la configuration d'hyper-paramètres choisis sur l'ensemble d'entraînement. On obtient alors : $\hat{f}_{(A_\lambda)} = A_\lambda(D_{train})$
6. Évaluer le résultat à l'aide de l'ensemble de validation.

$$e_{(A_\lambda)} = \hat{R}(\hat{f}_{(A_\lambda)}, D_{valid})$$

Une fois que ce travail a été effectuer pour tous les algorithmes et, pour chacun, sur chaque configurations d'hyper-paramètres λ jugés admissibles, on regarde les valeurs de $e_{(A_\lambda)}$ obtenues. Le modèle A_λ qui retourne une erreur de validation minimale $e^*_{(A_\lambda)}$ sera alors sélectionner.

Nous aurions également pu séparer D en 3 sous-ensembles plutôt que 2 et ainsi ajouter un ensemble de test. Une fois l'algorithme optimal trouvé à l'aide de l'ensemble d'entraînement et de l'ensemble de validation, nous aurions pu calculer un estimé non-biaisé de la performance de généralisation de l'algorithme choisit en calculant l'erreur empirique de validation sur l'ensemble de test; soit un ensemble de données dont le valeurs n'ont jamais été utilisées pour sélectionner l'algorithme dans les étapes précédantes. Cette étape nous permet de nous assurer (si nous avons assez de donner pour pouvoir effectuer un test concluent) que l'algorithme sélectionné est capable de généraliser sur de nouvelles données et qu'il n'est pas juste efficace sur les données de D_{train} et D_{valid} .

1 – Classifieurs de Bayes

1. Classifieur de Bayes obtenu avec des densités paramétriques Gaussiennes diagonales :