

TD 2

Exercice 1 – Exercice 1 : Classifieur bayésien

Soit \mathcal{X} un ensemble de description dans \mathbb{R}^d et \mathcal{Y} l'ensemble des labels $\{y_1, \dots, y_l\}$.

Q 1.1 Rappeler ce qu'est un classifieur bayésien.

Q 1.2 Exprimer l'erreur faite par le classifieur bayésien à un point \mathbf{x} . L'erreur est-elle minimale ?

Q 1.3 Soit $\lambda(y_j, y_i)$ le coût d'une erreur consistant à prédire le label y_j plutôt que y_i . Que valent les λ dans le cas de l'erreur 0-1 ? Donner quelques exemples de coûts asymétrique et des contextes d'utilisation.

Q 1.4 Quelle est l'expression du risque $R(y_i|\mathbf{x})$ de prédire y_i sachant \mathbf{x} en fonction de λ et des probabilités a posteriori ? Dans le cas 0-1 ?

Q 1.5 Donner l'expression du risque sur \mathcal{X} associé au classifieur f , $R(f)$.

Q 1.6 On se place dans le cas binaire. Exprimer le critère de décision en fonction de λ et des probabilités a posteriori, puis donner un critère de décision en fonction de λ , la distribution des classes et la vraisemblance.

Exercice 2 – Exercice 2 : Estimation de densité

Q 2.1 Donner l'estimation de la densité p_B d'une variable aléatoire X à l'intérieur d'une région d'intérêt B de volume V , en fonction d'un nombre k d'échantillons observés dans cette zone parmi n échantillons tirés.

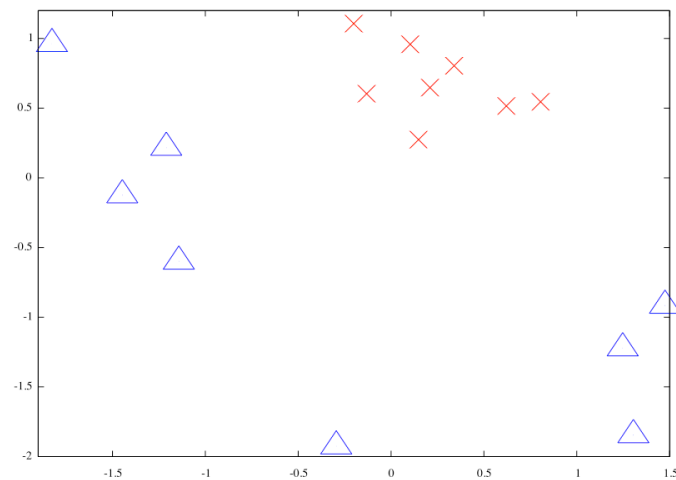
Q 2.2 Soit une variable aléatoire $X \in \mathcal{X}$. On souhaite estimer la densité p_X de cette variable à partir d'un ensemble d'observations \mathcal{X}_o . Décrire la manière de procéder pour réaliser cette estimation selon la méthode des histogrammes.

Q 2.3 Discuter des méthodes d'estimation de densité à noyaux

Exercice 3 – Exercice 3 : Classification selon voisinage

Q 3.1 Quelle différence entre les fenêtres de Parzen et les k -nn ? Que vérifie-t-on quand le nombre d'échantillons tend vers l'infini ?

Q 3.2 Sur l'exemple suivant, tracez la frontière de décision pour $k = 1$. Quel problème peut se poser pour des valeurs de k ?



Q 3.3 Ajouter un *outlier* en $(-0.5, -0.5)$. Comment évolue la frontière ?

Q 3.4 Et si $k = 3$? Que se passe-t-il quand k tend vers l'infini ?

Q 3.5 Soit \mathbf{x} un exemple à classifier, $(\mathbf{x}_i)_{i=1}^n$ une suite d'échantillons aléatoires et $(\mathbf{x}'_j)_{j=1}^n$ la suite extraite de l'ensemble précédent tel que \mathbf{x}'_j soit le plus proche voisin de \mathbf{x} à l'étape j parmi les $\{\mathbf{x}_i\}$. Montrer que la séquence $(\mathbf{x}'_i)_{i=1}^n$ converge vers \mathbf{x} .

Q 3.6 Exprimez le risque $r(\mathbf{x}, \mathbf{x}'_n)$, la probabilité de faire une erreur de classification sur \mathbf{x} à l'étape n en considérant le plus proche voisin \mathbf{x}'_n , en fonction des $q_k(\mathbf{x}) = P(y = k|\mathbf{x})$.

Q 3.7 Vers quoi converge $r(\mathbf{x}, \mathbf{x}'_n)$ quand le nombre d'échantillons tend vers l'infini ? Nous noterons $r(\mathbf{x})$ cette limite.

Q 3.8 Simplifier l'expression de $r(\mathbf{x})$.

Q 3.9 Montrer que $r(\mathbf{x}) \leq 2r_b(\mathbf{x})(1 - r_b(\mathbf{x}))$ dans le cas à 2 classes, avec $r_b(\mathbf{x})$ l'expression du risque bayésien pour x . Montrer que $r(\mathbf{x}) \leq r_b(\mathbf{x})(2 - \frac{K}{K-1}r_b(\mathbf{x}))$ dans le cas à K classes.

Indication : utiliser l'inégalité de Cauchy $|\sum_{i=1}^n u_i v_i|^2 \leq \sum_{i=1}^n |u_i|^2 \sum_{j=1}^n |v_j|^2$ en l'utilisant sur $K - 1$ q_i et en choisissant $v_j = 1$.