TD 2

Exercice 1 – Classifieur bayésien

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

Q 1.1 Rappelez ce qu'est un classifieur bayésien.

Q 1.2 Montrez qu'il fait une erreur 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 ? Donnez quelques exemples de coûts.

Q 1.4 Quelle est l'expression du risque $R(y_i|x)$ de prédire y_i sachant 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, $R(f)$.

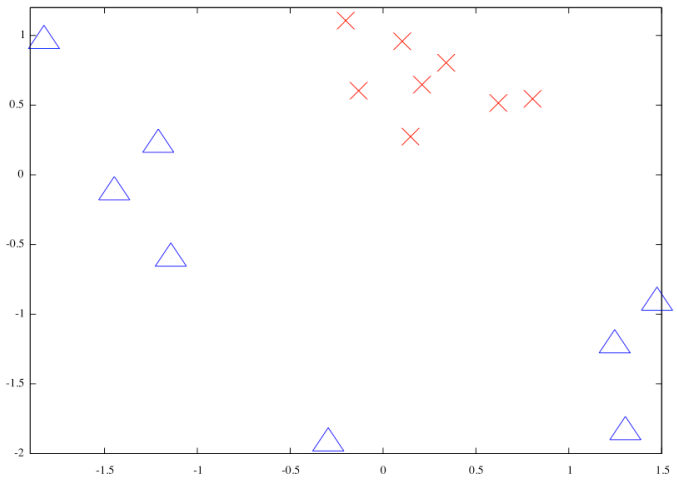
Q 1.6 On se place dans le cas binaire. Exprimez le critère de décision en fonction de λ et des probabilités a posteriori.

Exercice 2 – Estimation de densité

Q 2.1 Rappeler l'expression générale pour l'estimation de densité, $p(x)$ en fonction du nombre k d'échantillons qui tombent dans la région d'intérêt de volume V parmi les n échantillons tirés.

Q 2.2 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 2.3 Sur l'exemple suivant, tracez la frontière de décision pour $k = 1$. Quel problème peut se poser pour des valeurs de k ?



Q 2.4 Ajouter un *outlier* en $(-0.5, -0.5)$. Comment évolue la frontière ?

Q 2.5 Et si $k = 3$? Que se passe-t-il quand k tend vers l'infini ?

Q 2.6 Soit $f_1, ..., f_l$ les fonctions densités des différents labels, qu'on suppose strictement positive sur l'ensemble de définition. Soit un exemple aléatoire x, $(x_i)_{i=1}^n)$ une suite d'échantillons aléatoire et $(x'_i)_{i=1}^n$ tel que $x'_i \in \{x_1, ..., x_i\}$ soit le plus proche voisin de x à l'étape i. Montrez que la séquence (x'_i) converge vers x.

Q 2.7 Exprimez le risque $r(x, x'_n)$, la probabilité de faire une erreur de classification sur x à l'étape n en considérant le plus proche voisin x'_n, en fonction des $q_k(x) = P(y = k|x)$. Quel est l'expression du risque bayésien $r_b(x)$?

Q 2.8 Exprimez le risque $R(n)$ de mauvaise classification en fonction du nombre d'échantillons n en utilisant les $q_k(x)$.

Q 2.9 On se fixe à un point x donné. Simplifiez l'expression de $r(x, x)$ en utilisant $\left(\sum_j q_j(x)\right)^2$ et $\sum_i q_i(x)^2$.
et k la classe prédite par le classifieur bayésien.

Q 2.10 L'inégalité de Cauchy-Schwarz permet d'établir entre autre que pour k fixé (ici, le label prédit par le classifieur bayésien), $\left(\sum_{j\neq k} q_j(x)\right)^2 \leq (l-1)\sum_{j\neq k} q_j(x)^2$. En ré-arrangeant cette expression, montrez que $r(x) \leq 2r_B(x)$. Que venez vous de prouvez ?