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 $\{y_1, ..., y_l\}$.

Q 1.1 Rappeler ce qu'est un classifieur bayésien.

Rappeler $p(\mathbf{x})$ (l'evidence), $p(y)$ (le prior), $p(\mathbf{x}|y)$ (a priori, critere max de vraissemblance), $p(y|\mathbf{x})$ (critere max a posteriori)
classifieur bayésien $\Rightarrow \hat{y} = arg\max_y p(y|\mathbf{x})$
Evidence : $p(\mathbf{x}) = \sum_y p(\mathbf{x}|y)p(y)$

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

Erreur avec y_k la classe de x et $\hat{y}$ la prediction : $p(y_k \neq \hat{y}|x) = \sum_{y_i \neq \hat{y}} p(y_i|x) = 1 - p(\hat{y}|x)$.
Avec classifieur bayésien, erreur $= 1 - \max_y p(y|x) \Rightarrow$ risque minimal en chaque x (contradiction sinon)...

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.

Pour erreur 0-1 : $\lambda(y_j, y_i) = 1$ pour $j \neq i$, 0 sinon.
Exemple de coût asymetrique : alarme fraude bancaire ou on préfère le rappel à la précision (coûts moins importants pour les faux positifs que pour les faux négatifs)

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 ?

Le risque de prédire y_i pour x est : $R(y_i|x) = \sum_k \lambda(y_i, y_k)p(y_k|x)$ avec $p(y_k|x)$ la probabilité a posteriori que la vraie classe de x soit y_k

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

$R(f) = \int_{\mathcal{X}} R(f(x)|x)p(x)dx = \int_{\mathcal{X},\mathcal{Y}} \lambda(f(x), y)p(x, y)dxdy$: risque théorique
En pratique, on considère le risque empirique sur un ensemble d'entraînement $\mathcal{T} \subseteq \mathcal{X}$: $R(f) = \sum_{x_i \in \mathcal{T}} R(f(x^i))$, avec $R(f(x^i)) = \lambda(f(x^i), y^i)$ et y^i la classe observée pour x^i

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 vraissemblance.

$R(+|x) = \lambda_{+,+}p(+|x) + \lambda_{+,-}p(-|x) = \lambda_{+,+}p(+|x) + \lambda_{+,-}(1 - p(+|x))$,
$R(-|x) = \lambda_{-,+}p(+|x) + \lambda_{-,-}p(-|x) = \lambda_{-,+}p(+|x) + \lambda_{-,-}(1 - p(-|x))$

Si $R(+|x) < R(-|x)$ on prédit +, soit $(\lambda_{-,+} - \lambda_{+,+})p(+|x) > (\lambda_{+,-} - \lambda_{-,-})P(-|x)$.

Exercice 2 – Exercice 2 : Estimation de densité

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

La probabilité qu'un point x se trouve à l'intérieur d'une région $\mathcal{B}$ est donnée par $p(x \in \mathcal{B}) = \int\limits_{x' \in \mathcal{B}} p(x')dx'$
Si $\mathcal{B}$ est suffisamment petit, on peut supposer que tous les points $x \in \mathcal{B}$ possèdent une même densité de proba $p_{\mathcal{B}}$ et donc $p(x \in \mathcal{B}) \approx p_{\mathcal{B}} \int\limits_{x' \in \mathcal{B}} dx' = p_{\mathcal{B}} \times V$

La variable aléatoire Y correspond au nombre d'échantillons k dans $\mathcal{B}$ parmi n. On a $E(Y) = np(x \in \mathcal{B})$ (loi binomiale de paramètres n et $p(x \in \mathcal{B})$).
Quand le nombre d'échantillons n augmente, $k \to E(Y)$ (loi des grands nombres).
On a alors $k \approx np(x \in \mathcal{B}) \approx np_{\mathcal{B}}V$
Donc $p_{\mathcal{B}} \approx \frac{k}{nV}$.

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.

Decoupage de l'espace en boîtes régulières de volume V. (On peut détailler un peu pour qu'ils galèrent moins en TP)
Pour chaque boîte i on compte le nombre de données k_i dans la boîte i.
$p_X(x) = \frac{k_{b(x)}}{nV}$
On peut dessiner un histogramme pour loi normale par ex et montrer que quand on réduit la taille des boîtes on améliore l'estimation quand infinité de données mais on crée des trous quand $\mathcal{X}_o$ fini.

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

$p_X(x) = \sum_{x' \in \mathcal{X}_o} \frac{\phi(x,x')}{N}$
Si noyau créneau sur distance de x (1 si distance inférieure à ϵ, 0 sinon), alors équivalent à Parzen.
Sinon noyau gaussien par exemple : plus smooth, pas de problème d'absence de point dans le voisinage (mais plus lourd à calculer)

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 ?

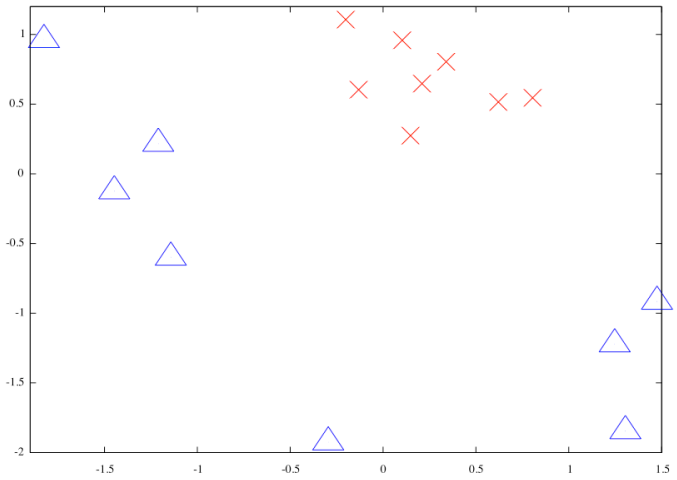
Fenêtres de Parzen : On définit un voisinage local en produisant une boîte $\mathcal{B}$ autour de la donnée à étiqueter.

On peut alors estimer : $p_{\mathcal{B}}(\mathbf{x}|y) \approx k_y/(V_{\mathcal{B}}n_y)$, avec k_y le nombre d'échantillons étiquetés y dans la boîte, n_y le nombre total d'échantillons étiquetés y dans l'ensemble de données observées et $V_{\mathcal{B}}$ le

volume de la boîte $\mathcal{B}$. Ce qui permet de calculer $p(y|x \in \mathcal{B})$ et prédire une classe pour l'objet x.
K-NN : Hypothèse que les k points les plus proches de x suivent la même distribution que x. Ensuite idem à Parzen
Parzen détermine un voisinage local alors que KNN s'intéresse aux k points les plus proches
Risque pour Parzen : personne dans le voisinage local
Risque pour KNN : points les plus proches trop éloignés pour que l'hypothèse de même densité soit vérifiée
Quand le nombre d'échantillons tend vers l'infini : on peut se permettre de choisir des tailles de fenêtres plus petites pour Parzen et nombres k de voisins qui tendent vers 1 puisque la probabilité des risques énoncés plus haut tend vers 0 (plus on a d'échantillons, moins on a de chances de n'avoir personne dans la fenêtre et plus on a de chance de prendre un voisin proche)
Lorsque le nombre d'échantillons est limité, approche de Watson-Nadaraga permet de pondérer knn selon la distance des voisins : Utilisation d'une fonction noyau déterminant une similarité ϕ entre deux points (ex fonction créneau qui revient à faire Parzen, noyau gaussien, polynomial, etc...) puis $E[y|x] = \sum_j y_j \frac{\phi(x, x_j)}{\sum_i \phi(x, x_i)}$
Peut être utilisé pour regression, lissage, classif, etc...

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 ?

$k = 1 \Rightarrow$ sur-apprentissage, frontière très peu régulière
k trop grand, perte d'info, sous-apprentissage, on s'intéresse à des points trop éloignés



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

avec k=1, change considérablement la frontiere, très instable.

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

k=3, frontière plus lisse, plus stable.
k infini, on tend vers classe majoritaire l'emporte partout

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 $\mathbf{x}$.

Soit ϵ aussi petit que voulu, soit $\mathcal{B}$ la boule de centre x_0 et de rayon ϵ.
Soit $\alpha = \int_{\mathcal{B}} p(x)dx > 0$, alors la proba que aucun $x_i \in \mathcal{B}$ pour $i = 1..n$ est $(1-\alpha)^n \to 0$ quand $n \to \infty$ (cas spécial du lemme de Borel-Cantelli, proba non nul donc ca va arriver à un moment dans une suite infinie d'expérience).
De plus, si x'_i appartient à $\mathcal{B}$, alors pour toutes les étapes suivantes, le plus proche voisin est également dans la boule : $\forall j > i, x'_j \in \mathcal{B}$ (x'_i est le plus proche voisin parmi les echantillons considérés, soit il le reste, soit l'autre est plus proche donc dans la boule).
La suite (x'_i) converge donc vers x_0 : pour ϵ donné, il existe toujours un n tel que $(x'_i)_{i>n} \in \mathcal{B}$ (définition de la convergence).

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 $q_k(\mathbf{x}) = P(y = k|\mathbf{x})$.

$r(x, x'_n) = p(y \neq y'|x, x'_n) = \sum_{i\in K} p(y \neq y_i, y' = y_i|x, x'_n) = \sum_{i\in K} P(y' = y_i|x'_n)P(y \neq y|x) = \sum_i q_i(x'_n)(1 - q_i(x))$

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.

Comme $\mathbf{x}'_n \to \mathbf{x}$, en remplaçant dans l'expression précédente en passant à la limite, $lim_\infty r(x, x'_n) = \sum_{i\in K} q_i(x)(1 - q_i(x))$

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

$r(\mathbf{x}) = \sum_{i\in K} q_i(\mathbf{x}) - \sum_{i\in K} q_i(\mathbf{x})^2 = 1 - \sum_{i\in K} q_i(\mathbf{x})^2$.

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. Monter 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 $v_j = 1$.

Soit z la classe prédite par le classifieur bayésien ($argmax q_i$). $r_b(x) = 1 - q_z = \sum_{i\neq z} q_i$ et $r(x) = 1 - \sum q_i^2$.
On a $(K-1)\sum_{i\neq z} q_i^2 \geq (\sum_{i\neq z} q_i)^2$, soit $-\sum_{i\neq z} q_i^2 \leq -\frac{1}{K-1}\sum_{i\neq 1} q_i)^2$, donc $r(x) \leq 1 - q_z^2 - \frac{1}{K-1}\sum_{i\neq z} q_i)^2$.
Alors : $r(x) \leq 1 - (1 - r_b(x))^2 - \frac{1}{K-1} r_b(x)^2 = 2r_b(x) - r_b(x)^2 - \frac{1}{K-1} r_b(x)^2 = 2r_b(x) - r_b(x)^2 \frac{K}{K-1} = r_b(x)(2 - \frac{K}{K-1} r_b(x))$