CORRECTIONS DE L'EXAMEN DE CLASSIFICATION DU VENDREDI 25 OCTOBRE 2019.

Exercice 1 : Classification Bayésienne (5 points)

On considère un problème de classification à deux classes ω_1 et ω_2 de densités uniformes sur les intervalles $[0, a]$ et $[0, 1]$ avec $a < 1$, c'est-à-dire

$$f(x|\omega_1) = \frac{1}{a}\mathcal{I}_{[0,a]}(x) \quad \text{et} \quad f(x|\omega_2) = \mathcal{I}_{[0,1]}(x) \tag{1}$$

où $\mathcal{I}_A$ est la fonction indicatrice sur l'ensemble A telle que $\mathcal{I}_A(x) = 1$ si $x \in A$ et $\mathcal{I}_A(x) = 0$ si $x \notin A$. On notera $P(\omega_1) = p \neq \frac{1}{2}$.

1. (1 pt) Déterminer les probabilités $P(\omega_1|x)$ et $P(\omega_2|x)$ et représenter les graphiquement pour $p = 1/4$ et $p = 3/4$, en distinguant avec soin les deux cas $x \in [0, a]$ et $x \in]a, 1]$.
   Réponse: En appliquant la formule de Bayes, on obtient

   $$P(\omega_1|x) = \frac{f(x|\omega_1)P(\omega_1)}{f(x)} = \frac{\frac{p}{a}\mathcal{I}_{[0,a]}(x)}{\frac{p}{a}\mathcal{I}_{[0,a]}(x) + (1-p)\mathcal{I}_{[0,1]}(x)} = \frac{p\mathcal{I}_{[0,a]}(x)}{p\mathcal{I}_{[0,a]}(x) + a(1-p)\mathcal{I}_{[0,1]}(x)}$$

   et

   $$P(\omega_2|x) = \frac{a(1-p)\mathcal{I}_{[0,1]}(x)}{p\mathcal{I}_{[0,a]}(x) + a(1-p)\mathcal{I}_{[0,1]}(x)}.$$

   Si on distingue les deux cas $x \in [0, a]$ et $x \in]a, 1]$, on obtient

   - $x \in [0, a]$

     $$P(\omega_1|x) = \frac{p}{p + a(1-p)} \quad \text{et} \quad P(\omega_2|x) = \frac{a(1-p)}{p + a(1-p)}$$

   - $x \in]a, 1]$

     $$P(\omega_1|x) = 0 \quad \text{et} \quad P(\omega_2|x) = 1.$$

2. (2 pts) Déterminer la règle de classification associée à ce problème avec la fonction de coût $0 - 1$ dans les deux cas $p = \frac{1}{4}$ et $p = \frac{3}{4}$.
   Réponse: le classifieur Bayésien affecte x à la classe ω_1 (ce que l'on notera $d^*(x) = \omega_1$) si

   $$P(\omega_1|x) > P(\omega_2|x)$$

   On a donc $d^*(x) = \omega_2, \ \forall x \in]a, 1]$. De plus, pour $x \in [0, a]$, on a

   $$d^*(x) = \begin{cases} \omega_1 \ \text{ si } \ p > a(1-p) \Leftrightarrow a < \frac{p}{1-p} \\ \omega_2 \ \text{ si } \ p < a(1-p) \Leftrightarrow a > \frac{p}{1-p} \end{cases} \tag{2}$$

   On remarquera que dans le cas $p < \frac{1}{2}$, on a $\frac{p}{1-p} < 1$ et donc les deux conditions $a < \frac{p}{1-p}$ et $a > \frac{p}{1-p}$ peuvent être satisfaites. Par contre, si $p > \frac{1}{2}$, on a $\frac{p}{1-p} > 1$ et donc on a nécessairement $p > a(1-p)$, soit $d^*(x) = \omega_1$. Dans les deux cas, $p = \frac{1}{4}$ et $p = \frac{3}{4}$, on obtient finalement

   - $p = \frac{1}{4}$

     $$d^*(x) = \omega_2, \ \forall x \in [a, 1] \tag{3}$$

     et pour $x \in [0, a]$

     $$d^*(x) = \begin{cases} \omega_1 \text{ si } a < \frac{1}{3} \\ \\ \omega_2 \text{ si } a > \frac{1}{3} \end{cases} \tag{4}$$

- $p = \frac{3}{4}$

$$d^*(x) = \begin{cases} \omega_1 \text{ si } x \in [0, a] \\ \omega_2 \text{ si } x \in]a, 1] \end{cases} \quad (5)$$

3. (2 pts) Déterminer la probabilité d'erreur associée à ce classifieur en fonction de a et p, en distinguant les deux cas $p < \frac{1}{2}$ et $p \geq \frac{1}{2}$.
Réponse: la probabilité d'erreur du classifieur est définie par

$$P_e = P[d^*(x) = \omega_1 | x \in \omega_2] P(x \in \omega_2) + P[d^*(x) = \omega_2 | x \in \omega_1] P(x \in \omega_1)$$

et se calcule comme suit

$$P_e = \int_{R_2} f(x|\omega_1) P(\omega_1) dx + \int_{R_1} f(x|\omega_2) P(\omega_2) dx$$

avec $R_1 = \{x | d^*(x) = \omega_1\}$ et $R_2 = \{x | d^*(x) = \omega_2\}$. On doit distinguer les cas $p < \frac{1}{2}$ et $p \geq \frac{1}{2}$:

- $p \geq \frac{1}{2}$. C'est le cas le plus simple qui correspond à $R_2 =]a, 1]$ et $R_1 = [0, a]$, d'où

$$P_e = \int_a^1 \frac{p}{a} \mathcal{I}_{[0,a]}(x) dx + \int_0^a q \mathcal{I}_{[0,1]}(x) dx = 0 + aq = a(1 - p).$$

- $p < \frac{1}{2}$. On doit distinguer les deux cas $a < \frac{p}{1-p}$ et $a > \frac{p}{1-p}$
  - Premier cas : $a < \frac{p}{1-p}$
    On a alors $R_2 =]a, 1]$ et $R_1 = [0, a]$, ce qui conduit au même résultat que pour $p \geq \frac{1}{2}$, soit
    $$P_e = a(1 - p).$$
  - Second cas : $a > \frac{p}{1-p}$
    On a alors $R_2 = [0, 1]$ et $R_1 = \emptyset$, d'où
    $$P_e = \int_0^1 \frac{p}{a} \mathcal{I}_{[0,a]}(x) dx = p.$$
    Ce résultat est compréhensible : comme on prend toujours la décision ω_2, la probabilité d'erreur est la probabilité de la classe ω_1.

Exercice 2 : Détection d'anomalies (4 points)

La méthode one-class SVM détermine un hyperplan séparateur par résolution du problème d'optimisation suivant

$$\begin{aligned} \underset{\boldsymbol{w}, \rho, \boldsymbol{\xi}_i}{\text{minimize}} \quad & \frac{1}{2} \|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu N} \sum_{i=1}^{N} \xi_i \\ \text{with} \quad & \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq \rho - \xi_i, \forall i = 1, ..., N. \end{aligned}$$

1. (1pt) Quelle est la règle de décision permettant de décider si un vecteur test $\boldsymbol{x}$ est une anomalie.
Réponse: On décide que le vecteur $\boldsymbol{x}$ est une anomalie si $\langle \boldsymbol{w}, \boldsymbol{x} \rangle < \rho$.

2. (1pt) Quel est le rôle des variables latentes ξ_i ?
Réponse: Le fait d'introduire des variable latentes ξ_i associées aux vecteurs $\boldsymbol{x}_i$ de la base d'apprentissage permet d'autoriser certains de ces vecteurs à ne pas respecter la contrainte $\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq \rho$. Plus précisément, cette contrainte est respectée lorsque $\xi = 0$ (ce qui signifie que le point $\boldsymbol{x}_i$ est à l'intérieur de la frontière de décision. Lorsque $\xi_i > 0$, le point $\boldsymbol{x}_i$ est à l'extérieur de la frontière de décision mais il y a peu de points vérifiant $\xi_i > 0$ en raison du terme $\sum_{i=1}^{N} \xi_i$ présent dans le terme à optimiser.

3. (1pt) Comment règle-t-on le paramètre ν et à quoi correspond-il ?
   Réponse: Ce paramètre correspond au nombre maximal de points de la base d'apprentissage situés à l'extérieur de la frontière de décision. Il peut être fixé par l'utilisateur comme on le fait pour une probabilité de fausse alarme (typiquement $\nu = 0.1$ ou $\nu = 0.05$ ou $\nu = 0.01$) ou optimisé par validation croisée.

4. (1pt) Dans le cas d'un noyau gaussien défini par $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$, comment règle-t-on le paramètre γ ?
   Réponse: Une règle simple est $\gamma = \frac{1}{2\sigma^2}$ avec σ la médiane des distances entre les données de la base d'apprentissage.

Questions sur l'article (11 points)

1. (1 pt) Expliquer la phrase "This is a supervised learning problem" utilisé dans l'introduction de la section 2.
   Réponse : on parle d'apprentissage supervisé lorsqu'on connait le nombre de classes (ici, ce sont les différentes catégories de documents) et qu'on a un nombre de données dans chaque classe que l'on va utiliser pour construire le classifieur. Dans le cas de cet article, les données sont des documents appartenant à chaque catégorie

2. (1 pt) Expliquer comment un document d est transformé en un vecteur de paramètres.
   Réponse : On compte le nombre de fois où chaque mot du dictionnaire w_i est présent dans le document, ce qui fournit un vecteur comme celui représenté dans la figure 1 dont les composantes sont notées DF(w_i). On effectue ensuite un changement de variables comme dans (1) et on divise chaque vecteur obtenu par sa norme (normalisation)

3. (2 pts) Quelle est la définition du gain en information utilisé par les auteurs pour déterminer les mots les plus discriminants ? Expliquer comment chaque terme de ce gain en information peut être calculé
   Réponse : D'après l'article de Yang et Petersen, le gain en information pour le mot w_k est défini par

$$G(w_k) = -\sum_{i=1}^{m} P(c_i)\ln[P(c_i)] + P(w_k)\sum_{i=1}^{m} P(c_i|w_k)\ln[P(c_i|w_k)] + P(\bar{w_k})\sum_{i=1}^{m} P(c_i|\bar{w_k})\ln[P(c_i|\bar{w_k})]$$

   où

   - $P(c_i)$ est le nombre de fois où la catégorie c_j est présente divisé par le nombre total de documents de la base d'apprentissage
   - $P(c_i|w_k) = \frac{P(w_k|c_i)P(c_i)}{P(w_k)}$ (règle de Bayes) avec
     - $P(w_k|c_i)$ est le nombre de documents de la catégorie c_i contenant le mot w_k divisé par le nombre de documents de la catégorie c_i.
     - $P(w_k) = \sum_i P(w_k|c_i)P(c_i)$ (théorème des probabilités totales)
     - $P(\bar{w_k}|c_i)$ est le nombre de documents de la catégorie c_i ne contenant pas le mot w_k divisé par le nombre de documents de la catégorie c_i.
     - $P(\bar{w_k}) = \sum_i P(\bar{w_k}|c_i)P(c_i)$

4. (1 pt) Expliquer pourquoi les vecteurs de données sont parcimonieux (sparse).
   Réponse : le nombre de mots possible est en général grand (supérieur à 10000) et seuls certains de ces mots sont présents dans un document d. Quand un mot n'est pas présent, la variable correspondante dans le vecteur de paramètres (feature vector) est nulle. En conséquence, le vecteur de paramètres (noté IDF dans l'article) contient en général beaucoup de 0, ce qui est la définition d'un vecteur parcimonieux.

5. (2 pts) Expliquer avec soin comment on détermine le vecteur $\overrightarrow{w}$ de la règle de décision (4) à partir des données d'apprentissage dans le cas de deux classes (i.e., ici de deux catégories).
Réponse : Comme expliqué dans le cours, le vecteur $\overrightarrow{w}$ est défini par

$$\overrightarrow{w} = \sum_{i=1}^{n} \alpha_i y_i \text{IDF}_i$$

où n est le nombre de données d'apprentissage (ici le nombre de documents), y_i est l'étiquette du vecteur IDF_i égale à $+1$ pour la catégorie $\#1$ et égale à -1 pour la catégorie $\#2$ et α_i est le $i$ème paramètre de Lagrange. Les paramètres de Lagrange sont déterminés en optimisant le critère (qui correspond à l'équation (9))

$$U(\boldsymbol{\alpha}) = -\frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{Y}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{Y}\boldsymbol{\alpha} + \sum_{i=1}^{n} \alpha_i$$

sous les contraintes $\alpha_i \geq 0$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$ (voir équation (10)).

6. (1 pt) Dans le cas d'utilisation d'un noyau RBF (radial basis function), comment les auteurs proposent-ils de déterminer γ ?
Réponse : les auteurs proposent de déterminer plusieurs classifieurs associés à différentes valeurs de γ. On retiendra la valeur de γ qui minimise la borne supérieure de la dimension de Vapnik définie dans (6), où R est le maximum des normes des vecteurs d'apprentissage et A est la norme du vecteur $\overrightarrow{w}$.

7. (1 pt) Rappeler le principe du classifieur Bayésien évoqué dans la section 4.1.
Réponse : le classifieur Bayésien affecte un vecteur de données IDF à la classe la plus probable a posteriori, c'est-à-dire qui maximise

$$P\left(C_i|\text{IDF}\right).$$

Ici le calcul de cette probabilité a posteriori utilise le fait qu'un document est composé de mots avec un certains nombre d'occurences, ce qui conduit à (18).

8. (1 pt) Expliquer le principe de l'algorithme de classification C4.5.
Réponse : le classifieur C4.5 est un arbre de décision avec un critère d'impureté égal à l'entropie.