Apprentissage statistique

Examen final - 2 heures

Les documents sont autorisés.

- Exercice 1 -

1. Rappeler la définition de la VC dimension d'une collection de sous-ensembles de $\mathbb{R}^d$.
2. Calculer la VC dimension de la collection $\mathcal{A}$ des orthants de $\mathbb{R}^2$:

$$\mathcal{A}_2 = \{\ ]-\infty, a_1] \times]-\infty, a_2] \quad : \quad a_1, a_2 \in \mathbb{R}^2\}$$

   On pourra argumenter par des figures.
3. Généraliser le calcul de la VC dimension au cas de la collection $\mathcal{A}_d$ des orthants de $\mathbb{R}^d$.
4. Soit :

$$\mathcal{G} = \left\{ \sum_{k=1}^{K} w_k \mathbb{I}\{A_k\} \ : \ K \geq 1, w_k > 0, A_k \in \mathcal{A}_2 \right\} .$$

   On considère la collection de sous-ensembles de $\mathbb{R}^2$ donnée par :

$$\mathcal{C} = \left\{ C = \{x \in \mathbb{R}^2 \ : \ g(x) > 1\} \quad : \quad g \in \mathcal{G} \right\} .$$

   Quelle est la VC dimension de $\mathcal{C}$? Justifier votre réponse. Indication : on pourra considérer des points sur une droite dans $\mathbb{R}^2$ de pente négative passant par l'origine.

- Exercice 2 -

On se place dans le cadre d'un problème de classification supervisée où l'échantillon est observé séquentiellement. Soit :

- une suite $(x_t, y_t)_{t\geq 1}$ d'observations $\mathbb{R}^d \times \{-1, +1\}$,
- $< \cdot, \cdot >$ le produit scalaire sur $\mathbb{R}^d$ et $\| \cdot \|$ la norme associée,
- des observations x_t sur la sphère unité : $\forall t$, $\|x_t\| = 1$,
- des observations linéairement séparables par un hyperplan contenant l'origine : il existe $w^* \in \mathbb{R}^d$ tel que $\|w^*\| = 1$ et $y_t \cdot < w^*, x_t > > 0$, $\forall t$.

On définit l'algorithme du perceptron sur la sphère unité de la manière suivante :

start $t = 1$, $w_1 = 0 \in \mathbb{R}^d$.

for $t = 1, \ldots, T$ **do**
$\widehat{y}_t = \text{sgn}(< w_t, x_t >)$
if $\widehat{y}_t = y_t$ **then** $w_{t+1} = w_t$
else $w_{t+1} = w_t + y_t x_t$
end for

1. Proposer une généralisation du perceptron décrit ci-dessus au cas d'observations linéairement séparables dans $\mathbb{R}^d$ où l'hyperplan séparateur ne contient pas nécessairement l'origine et où les x_t peuvent être en position quelconque.
2. On considère l'algorithme du perceptron sur la sphère unité. On pose pour $T > 1$ fixé :

$$\gamma_T = \min_{1\leq t\leq T} | < w^*, x_t > | \ .$$

   (a) Montrer que : $< w^*, w_{t+1} > \geq < w^*, w_t > + \gamma_T$ lorsque $\widehat{y}_t \neq y_t$.

   (b) Montrer que : $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$.

   (c) Supposons que le perceptron ait M_T points mal classés au bout de T itérations. Donner une borne inférieure pour $< w^*, w_{t+1} >$ en fonction de γ_T et M_T, et une borne supérieure pour $\|w_{t+1}\|$ en fonction de M_T.

   (d) Déduire de la question précédente une majoration de M_T dépendant uniquement de γ_T.

3. On considère à nouveau l'algorithme du perceptron sur la boule unité, mais on remplace l'évaluation de $\widehat{y}_t$ par le prédicteur dit à vastes marges :

$$\widetilde{y}_t = \begin{cases} +1 & \text{si} \quad < w_t, x_t > > \gamma_T/2 \\ -1 & \text{si} \quad < w_t, x_t > < -\gamma_T/2 \\ 0 & \text{sinon.} \end{cases}$$

   (a) Montrer que dans ce cas, on a :

$$\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2\|w_t\|} + \frac{\gamma_T}{2}$$

   (b) En déduire la majoration sur M_T dans ce cas.

4. Y a-t-il un avantage à considérer le perceptron à vastes marges ? On commentera notamment la capacité en généralisation de l'algorithme en termes statistiques.

- Exercice 3-

On rappelle le cadre et les résultats obtenus dans l'Exercice 4 de l'examen partiel. On considère le modèle de classification binaire $(X, Y) \sim P$ où P est une loi sur $\mathbb{R} \times \{0, 1\}$. On représente la loi conditionnelle de X sachant Y par les deux fonctions de répartition $F_y(x) = \mathbb{P}\{X \leq x \mid Y = y\}$ pour $x \in \mathbb{R}$, $y \in \{0, 1\}$. On note $p = \mathbb{P}\{Y = 1\}$. On considère la famille $\mathcal{G}_L$ des classifieurs linéaires sur $\mathbb{R}$ de la forme :

$$g_{(x_0,y_0)}(x) = \begin{cases} y_0 & \text{si } x \leq x_0 \\ 1 - y_0 & \text{sinon} , \end{cases}$$

avec $x_0 \in \mathbb{R}$ et $y_0 \in \{0, 1\}$. On admet :

- L'erreur de classification $L(x_0, y_0) = \mathbb{P}\{Y \neq g_{(x_0,y_0)}(X)\}$ pour les éléments de $\mathcal{G}_L$ s'exprime :

$$L(x_0, y_0) = \mathbb{I}\{y_0 = 0\}\,(pF_1(x_0) + (1-p)(1 - F_0(x_0))) + \mathbb{I}\{y_0 = 1\}\,(p(1 - F_1(x_0)) + (1-p)F_0(x_0))$$

- L'erreur de classification minimale $L_0 = \inf_{(x_0,y_0)} L(x_0, y_0)$ s'exprime :

$$L_0 = \frac{1}{2} - \sup_{x_0 \in \mathbb{R}} \left| pF_1(x_0) - (1-p)F_0(x_0) - p + \frac{1}{2} \right|$$

- On note l'erreur de Bayes $L^* = \inf_g L(g)$. On a : $L^* \leq L_0 \leq \min\{p, 1-p\}$.
- On suppose que $p = 1/2$. On a dans ce cas :

$$L_0 = \frac{1}{2} - \frac{1}{2} \sup_x |F_1(x) - F_0(x)| .$$

1. On considère la loi P décrite par :

   a) X suit une loi uniforme sur $[0, 1]$,
   b) $Y = \mathbb{I}\{X \notin]1/3 + \epsilon, 2/3 - \epsilon]\}$ où $0 < \epsilon < 1/6$.
   Calculer L^* et L_0. Peut-on dire que les classifieurs linéaires soient adaptés pour ce type de situation ?

2. On considère maintenant la construction d'un classifieur linéaire empirique. On suppose que $X_1, \ldots, X_n$ est un échantillon i.i.d. de loi P inconnue. On note $C(u, v)$ l'ensemble $\{]-\infty, u] \times \{v\}\} \cup \{[u, +\infty[\times \{1 - v\}\}$ et $(\hat{u}, \hat{v})$ le minimiseur du risque empirique et $\hat{g}$ le classifieur associé. Soit $L(\hat{g}) = \mathbb{P}\{(X, Y) \in C(\hat{u}, \hat{v})\} = P(C(\hat{u}, \hat{v}))$. Montrer que

$$L(\hat{g}) \leq L_0 + 2 \sup_{(u,v) \in \mathbb{R} \times \{0,1\}} | \, (P - \hat{P}_n)(C(u, v)) \, | \quad .$$

   où $\hat{P}_n = n^{-1} \sum_{i=1}^n \delta_{(X_i,Y_i)}$, avec $\delta_{(x,y)}$ est la mesure de Dirac en $(x, y) \in \mathbb{R} \times \{0, 1\}$.

3. Donner une borne exponentielle pour la quantité $\mathbb{P}\{L(\hat{g}) - L_0 \geq t\}$ pour tout $t > 0$ de la forme : $K \exp\{-c \cdot n(t/2)^2\}$ en précisant K et c. On pourra utiliser l'inégalité de Massart :

$$\mathbb{P}\left\{ \sqrt{n} \cdot \sup_{x \in \mathbb{R}} |F(x) - \hat{F}_n(x)| > \epsilon \right\} \leq 2 \exp\{-2\epsilon^2\}$$

   où F est une fonction de répartition, $\hat{F}_n$ est la fonction de répartition empirique et $\epsilon > 0$.

4. Déduire de l'inégalité précédente une borne supérieure pour $\mathbb{E}(L(\hat{g}) - L_0)$.

5. Commenter la propriété de consistance du minimiseur du risque empirique.