Introduction à l'apprentissage statistique

Examen final

Durée : 2h - documents autorisés

Exercice 1 -

1. Soit $\varphi : \mathbb{R} \to \mathbb{R}_+$ une fonction de perte. On définit la fonction suivante :

$$\forall u \in [0,1] \ , \quad H(u) = \inf_{t \in [-\infty, +\infty]} \{u\varphi(-t) + (1-u)\varphi(t)\} \ .$$

   Calculer $H_j(u) = H(u)$ dans le cas où $\varphi = \varphi_j$ correspondants aux trois cas suivants $(j = 1, 2, 3)$:
   (a) $\varphi_1(t) = \exp(t)$,
   (b) $\varphi_2(t) = \max\{0, 1+t\}$,
   (c) $\varphi_3(t) = (1+t)^2$.

2. Tracer les trois fonctions H_1, H_2, H_3 obtenues et proposer une étude rapide pour énoncer les propriétés qu'elles partagent.

3. On considère le cas $\varphi_1(t) = \exp(t)$. Montrer que :

$$\forall u \in [0,1] \ , \quad \left| \frac{1}{2} - u \right|^s \leq c^s (1 - H_1(u))$$

   pour des valeurs de $c > 0$ et $s \geq 1$ que l'on précisera.

4. On se place dans le cadre du modèle de données de classification binaire avec X dans $\mathbb{R}^d$ et Y dans $\{-1, +1\}$. On considère des règles de décision $f : \mathbb{R}^d \to \mathbb{R}$ et on définit le risque convexifié $A(f) = \mathbb{E}\big(\varphi(-Y \cdot f(X))\big)$. On note f^* l'élément minimisant $A(f)$. Quel est le lien entre $A(f^*)$ et H ?

5. On note $L(f) = \mathbb{P}\{Y \cdot f(X) < 0\}$. Utiliser la propriété sur H pour en déduire une majoration de $L(f) - L(f^*)$ dépendant de c, s et A sous des hypothèses sur φ qu'on précisera. Préciser les valeurs de c et s dans le cas exponentiel.

6. Commenter les cas de H_2 et H_3.

7. Quelle fonction de perte est préférable pour la classification binaire ?

Exercice 2 - On considère une classe $\mathcal{F}$ de fonctions $f : \mathbb{R}^d \to \mathbb{R}$ et on pose :

$$\hat{R}(\mathcal{F}, S_n) = \mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(x_i)\right)$$

où $S_n = \{x_1, ..., x_n\}$ est un ensemble de vecteurs de $\mathbb{R}^d$ fixés, et l'espérance est calculée relativement à la loi uniforme sur $\{-1, +1\}^n$ pour le vecteur aléatoire $(\varepsilon_1, \ldots, \varepsilon_n)$ de $\mathbb{R}^n$.

Rappels mathématiques.

- **Fonction lipschitzienne** - Une fonction $\Phi : \mathbb{R} \to \mathbb{R}$ est dite k-lipschitzienne, si elle satisfait pour tout t, t' :
$$|\Phi(t) - \Phi(t')| \leq k|t - t'| \ .$$
- **Définition du supremum** - Pour toute fonction $\Psi(h)$, on a : pour tout $\delta > 0$, il existe un élément h_0 tel que :
$$\Psi(h_0) \geq (1 - \delta) \sup_h \Psi(h) \ .$$

1. Soit Φ une fonction k-lipschitzienne. On propose d'établir une borne supérieure de $\hat{R}(\Phi \circ \mathcal{F}, S_n)$ en fonction de $\hat{R}(\mathcal{F}, S_n)$.
   (a) Calculer : $\mathbb{E}\left(\sup_{f\in\mathcal{F}}\big(u(f) + \varepsilon_n(\Phi \circ f)(x_n)\big)\right)$ où $u(f)$ est une quantité déterministe.
   (b) Montrer que : pour tout $\delta > 0$, on peut introduire f_1 et f_2 telles que
   $$(1-\delta)\mathbb{E}\left(\sup_{f\in\mathcal{F}} (u(f) + \varepsilon_n(\Phi \circ f)(x_n))\right) \leq \frac{1}{2}\left(u(f_1) + skf_1(x_n)\right) + \frac{1}{2}\left(u(f_2) - skf_2(x_n)\right)$$
   où $s \in \{-1, +1\}$.
   (c) En déduire que :
   $$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\big(u(f) + \varepsilon_n(\Phi \circ f)(x_n)\big)\right) \leq \mathbb{E}\left(\sup_{f\in\mathcal{F}}\big(u(f) + \varepsilon_n kf(x_n)\big)\right) \ .$$
   (d) Dériver une majoration de $\hat{R}(\Phi \circ \mathcal{F}, S_n)$ en utilisant un argument de récurrence.
2. On introduit la famille de fonctions : $\bar{\mathcal{F}} = \{z = (x, y) \mapsto -yf(x) \ : \ f \in \mathcal{F}\}$. Soit les échantillons IID $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ et $S_n = \{X_1, \ldots, X_n\}$. Montrer que :
   $$\mathbb{E}\hat{R}(\bar{\mathcal{F}}, D_n) = \mathbb{E}\hat{R}(\mathcal{F}, S_n)$$
3. On pose $\Phi(t) = (1 + t)_+ = \max\{0, 1 + t\}$. Quelle est la relation entre $\mathbb{E}\hat{R}(\Phi \circ \bar{\mathcal{F}}, D_n)$ et $\mathbb{E}\hat{R}(\mathcal{F}, S_n)$?
4. Donner les arguments principaux pour établir, avec grande probabilité, une borne supérieure sur l'erreur de classification pour une règle de décision définie par :
   $$\hat{f}_n = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} (1 - Y_i \cdot f(X_i))_+$$
   où $(X_1, Y_1), \ldots, (X_n, Y_n)$ est un n-échantillon IID.
5. Décrire un algorithme implémentant cet objectif.

Exercice 3 - On considère le problème d'apprentissage suivant :

- Modèle de données - Soit un triplet aléatoire (X, X', Y) où X, X' sont des vecteurs aléatoires IID sur $\mathbb{R}^d$ et Y est une variable aléatoire sur $\{-1, 0, +1\}$. On suppose $Y = f^*(X, X')$.
- Critère d'erreur - Pour toute règle de décision $h : \mathbb{R}^d \to \{-1, 0, +1\}$, on considère l'erreur :

$$L(h) = \mathbb{P}\{f^*(X, X') \neq 0,\ f^*(X, X')(h(X') - h(X)) \leq 0\}$$

1. Illustrer la pertinence de cette formulation. Que cherche-t-on à apprendre ici ?
2. Donner l'erreur minimale L^* pour ce modèle.
3. Justifier l'utilisation du critère convexifié :

$$A_n(h) = \sum_{i=1}^{n} \exp\left(-Y_i(h(X_i') - h(X_i))\right)$$

où les (X_i, X_i', Y_i) forment un échantillon IID.

4. On propose d'étudier des fonctions de décision h de la forme $h_T = \sum_{t=1}^{T} \alpha_t g_t$ où les α_t sont des poids réels et g_t des fonctions de décision à valeurs dans $\{-1, 0, +1\}$. Proposer un type d'algorithme s'appuyant sur un principe itératif pour construire (α_t, g_t) et utilisant un système de poids convexes $\pi_t(i)$ sur les points de l'échantillon $(i = 1, \ldots, n)$. On explicitera les poids $\pi_t(i)$.
5. On introduit :

$$\epsilon_t^+ = \sum_{i=1}^{n} \pi_t(i) \mathbb{I}\{Y_i \cdot (h(X_i') - h(X_i)) > 0\}$$

et

$$\epsilon_t^- = \sum_{i=1}^{n} \pi_t(i) \mathbb{I}\{Y_i \cdot (h(X_i') - h(X_i)) < 0\}\ .$$

Montrer que minimiser la dérivée directionnelle $\Delta(g)$ par rapport à g

$$\Delta(g) = \left.\frac{\partial A_n(h_{t-1} + \alpha g)}{\partial \alpha}\right|_{\alpha=0}$$

est équivalent à minimiser $\epsilon_t^- - \epsilon_t^+$.

6. Préciser le choix de α à chaque itération de l'algorithme.
7. Quels sont les paramètres de la méthode et comment les ajuster ?