Apprentissage statistique: TD3
Risk Minimization

Emile Contal
http://econtal.perso.math.cnrs.fr/teaching

30 janvier 2015

Exercice 1.

1. **(Classifieur de Bayes et risque quadratique)** Soient X et Y deux variables aléatoires sur l'espace de probabilité $(\Omega, \mathcal{F}, \Pr)$ avec X dans $\mathbb{R}^d$ et Y dans $\{0,1\}$. On définit μ et η deux mesures de probabilités telles que $\forall A \subseteq \mathcal{B}(\mathbb{R}^d)$, $\mu(A) = \Pr[X \in A]$ et $\forall x \in \mathbb{R}^d$, $\eta(x) = \mathbf{E}[Y \mid X = x]$.

   Un *classifieur* g est une fonction $g : \mathbb{R}^d \to \{0,1\}$. On définit le *risque* $L(g)$ d'un classifieur g:

   $$L(g) = \Pr[g(X) \neq Y]\,.$$

   Soit $g^\star$ le classifieur de Bayes, c'est à dire :

   $$g^\star(x) = \begin{cases} 1 & \text{si } \eta(x) > \frac{1}{2} \\ 0 & \text{sinon}\,. \end{cases}$$

   (a) Montrer que $g^\star = \operatorname{argmin}_{g:\mathbb{R}^d \to \{0,1\}} L(g)$.

   (b) Montrer que $\eta = \operatorname{argmin}_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}[(f(X) - Y)^2]$.

2. **(Risques pondérés)**

   (a) Soit $L_\omega(g) = \mathbf{E}[2\omega(Y)\mathbb{1}_{Y \neq g(X)}]$ le risque pondéré par ω tel que $\omega(0) + \omega(1) = 1$. Donner le classifieur de Bayes et son risque pour ce critère.

   (b) On considère des classifieurs g avec possibilité de rejet, $g : \mathbb{R}^d \to \{0, 1, \perp\}$. Soit $L_\omega(g)$ le risque associé pondéré par ω :

   $$L_\omega(g) = \Pr[Y \neq g(X), g(X) \neq \perp] + \omega \Pr[g(X) = \perp]\,.$$

   Donner le classifieur de Bayes et son risque pour ce critère.

3. **(Risques convexes)** On se place dans le cadre de la classification sur $\{-1, +1\}$. Soit ϕ une fonction convexe et $g : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ un classifieur. On définit $R_\phi(g)$ le risque convexe $\mathbf{E}\Big[\phi\big(Yg(X)\big)\Big]$. Donner le classifieur de risque optimal pour les fonctions ϕ suivantes :

(a) $\phi(x) = \exp(-x)$

(b) $\phi(x) = \max(0, 1-x)^2$

(c) $\phi(x) = \log(1 + e^{-x})$

Exercice 2.

1. **(Distance L1 entre les densités conditionnelles)** Soit $L^\star = L(g^\star)$ le risque du classifieur de Bayes.

   (a) Montrer que $L^\star = \frac{1}{2} - \frac{1}{2}\mathbf{E}[|2\eta(X) - 1|]$.

   (b) Soit f_0 (resp. f_1) la densité de X sachant que $Y = 0$ (resp. $Y = 1$). Fixons $E[Y] = \frac{1}{2}$. Montrer que $L^\star = \frac{1}{2} - \frac{1}{4}\int |f_1(x) - f_0(x)| \, \mathrm{d}x$.

2. **(Mélange de Gaussiennes)** Soit $f_0 \sim \mathcal{N}(m_0, \Sigma_0)$ et $f_1 \sim \mathcal{N}(m_1, \Sigma_1)$ deux densités Gaussiennes, et $p \in [0, 1]$, tels que f_i soit la densité de X sachant $Y = i$, et $\mathbf{E}[Y] = p$.

   (a) Donner le classifieur de Bayes pour ce problème.

   (b) Quelle est la forme de ce classifieur lorsque $\Sigma_0 = \Sigma_1$? De même lorsque $\Sigma_0 \neq \Sigma_1$.

Exercice 3 (bonus).

1. **(Risque empirique et règles linéaires)** NOTE : ajouter iid + densité de X

   Soit $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un *ensemble d'apprentissage*, où les couples (X_i, Y_i) sont des *observations*, c'est à dire des variables aléatoires tirées suivant la même loi que (X, Y). On définit le *risque empirique* $\widehat{L}_n(g)$ d'un classifieur g :

   $$\widehat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Y_i} \, ,$$

   et le *risque réel* $L(g)$:

   $$L(g) = \Pr[g(X) \neq Y] \, .$$

   Soit $\mathcal{H}$ l'ensemble des classifieurs linéaires dans $\mathbb{R}^d$. On remarque qu'un ensemble de d points suffisent à déterminer un hyperplan unique X-presque sûrement, donc deux classifieurs. Étant donné D_n, on peut donc définir $\mathcal{H}_n = \{h_1, \ldots, h_{2\binom{n}{d}}\}$, une famille de $2\binom{n}{d}$ classifieurs linéaires.

   Soient $\widehat{h}_n = \operatorname{argmin}_{h \in \mathcal{H}_n} \widehat{L}_n(h)$ et $h^\star = \operatorname{argmin}_{h \in \mathcal{H}} L(h)$. On cherche à montrer que $\mathbf{E}[L(\widehat{h}_n) - L(h^\star)]$ tend vers zéro.

   (a) Montrer que $\min_{h \in \mathcal{H}} \widehat{L}_n(h) \geq \widehat{L}_n(\widehat{h}_n) - \frac{d}{n}$.

(b) En déduire que pour tout $\frac{2d}{n} \leq \epsilon \leq 1$,

$$\Pr\left[L(\widehat{h}_n) - L(h^\star) > \epsilon\right] \leq \sum_{i=1}^{2\binom{n}{d}} \Pr\left[L(h_i) - \widehat{L}_n(h_i) > \frac{\epsilon}{2}\right] + \Pr\left[\widehat{L}_n(h^\star) - L(h^\star) + \frac{d}{n} > \frac{\epsilon}{2}\right].$$

(c) En appliquant l'inégalité de Hoeffding à une loi binômiale, montrer que,

$$\Pr\left[\widehat{L}_n(h^\star) - L(h^\star) > \frac{\epsilon}{2} - \frac{d}{n}\right] \leq e^{-2n\left(\frac{\epsilon}{2} - \frac{d}{n}\right)^2}.$$

(d) Grâce à la loi des espérances itérée, utiliser le même argument pour montrer que pour tout i,

$$\Pr\left[L(h_i) - \widehat{L}_n(h_i) > \frac{\epsilon}{2}\right] \leq e^{-2n\left(\frac{\epsilon}{2} - \frac{d}{n}\right)^2}.$$

(e) D'après l'inégalité de Markov sur $Z = \left(L(\widehat{h}_n) - L(h^\star)\right)^2$, et l'inégalité de Cauchy-Schwarz, en déduire une borne pour $\mathbf{E}[Z]$ en fonction de ϵ.

(f) Optimiser la borne sur ϵ pour montrer que,

$$\mathbf{E}[L(\widehat{h}_n) - L(h^\star)] \leq \sqrt{\frac{2\left((d+1)\log n + (2d+2)\right)}{n}}.$$

2. **(Cas linéairement séparable)** Lorsqu'il existe un classifieur de risque nul, montrer de même que,

$$\mathbf{E}[L(\widehat{h}_n)] \leq \frac{d\log n + 2}{n - d}.$$