Apprentissage statistique: TD4
Online-to-Batch & Probabilistic Framework

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

25 janvier 2015

Exercice 1.

Soient $(X_1, X_1), \dots, (X_T, X_T) \in (\mathcal{X}, \mathcal{Y})^T$ des données indépendantes et identiquement distribuées selon P. On se munit d'une perte $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ bornée par $M \geq 0$. On s'intéresse dans cet exercice à l'erreur de la prédiction moyenne après avoir appliqué un algorithme en ligne sur les données (par exemple EWA), qui ne connaît évidemment pas P.
Soient $h_1, \dots, h_{T+1} \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ les règles successives produites par l'algorithme en ligne. On rappelle la définition du regret:

$$R_T = \sum_{t=1}^{T} \ell\big(h_t(x_t), y_t\big) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell\big(h(x_t), y_t\big)\,.$$

On définit la perte espérée d'une prédiction par h suivant la distribution P :

$$R(h) = \mathbb{E}_P\big[\ell(h(x), y)\big]$$

1. Montrer que pour tout $\delta > 0$, l'inégalité suivante:

$$\frac{1}{T}\sum_{t=1}^{T} R(h_t) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(h_t(x_t), y_t) + M\sqrt{\frac{2\log\frac{1}{\delta}}{T}}\,,$$

   est vraie avec probabilité au moins $1 - \delta$.
   Indice : utiliser l'inégalité d'Azuma du TD1.

2. On considère maintenant que la perte ℓ est convexe en son premier argument.

   (a) Montrer qu'avec probabilité au moins $1 - \delta$:

$$R\Big(\frac{1}{T}\sum_{t=1}^{T} h_t\Big) \leq \frac{1}{T}\sum_{t=1}^{T} \ell\big(h_t(x_t), y_t\big) + M\sqrt{\frac{2\log\frac{1}{\delta}}{T}}\,.$$

(b) En déduire qu'avec probabilité au moins $1-\delta$:

$$R\Big(\frac{1}{T}\sum_{t=1}^{T} h_t\Big) \leq \inf_{h\in\mathcal{H}} R(h) + \frac{R_T}{T} + 2M\sqrt{\frac{2\log\frac{2}{\delta}}{T}}\,.$$

Indice: utiliser l'inégalité de Hoeffding pour comparer $\sum \ell(\cdot,\cdot)$ *et* $R(\cdot)$.

(c) En déduire une borne sur la perte espérée de la prédiction obtenue par moyennage des règles successives de l'algorithme EWA.

Exercice 2.

On se place dans le cadre du modèle de classification où (X,Y) suit une loi P sur $\mathbb{R}^d \times \{0,1\}$.

1. **(Risques pondérés)**

   (a) On considère comme classifieur des fonctions mesurables du type $g : \mathbb{R}^d \to \{0,1\}$ et des poids ω sur $\{0,1\}$ tels que $\omega(0)+\omega(1)=1$. Soit $L_\omega(g) = \mathbb{E}_P[2\omega(Y)\mathbb{I}_{Y\neq g(X)}]$ le risque pondéré par ω.
   Donner le classifieur optimal 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 $\omega \in [0,1]$:

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

   Donner le classifieur optimal et son risque pour ce critère.

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

1. **(Risque empirique et règles linéaires)** Soient $\{(X_1,Y_1),\ldots,(X_n,Y_n)\} = D_n$ un ensemble d'apprentissage ainsi que (X,Y) indépendants et tirés suivant P. On considère ici que X possède une densité dans $\mathbb{R}^d$. On définit le *risque empirique* $\widehat{L}_n(g)$ d'un classifieur g :

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

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 $\mathbb{E}[L(\widehat{h}_n) - L(h^\star)]$ 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) Soit $Z = \left(L(\widehat{h}_n) - L(h^\star)\right)^2$ montrer que pour $u \geq \left(\frac{2d}{n}\right)^2$,

$$\mathbb{E}[Z] \leq (2n^d + 1)e^{2d - nu/2} + u\ .$$

(f) Optimiser la borne sur u et utiliser l'inégalité de Cauchy-Schwarz pour montrer que lorsque $n \geq d$,

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

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

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