Data Mining 2: Modélisation statistique et apprentissage **G. Cohen**

Exercice 08 janvier 2018

ETheme 1 Apprentissage supervisé

1 Questions

1. Qu'est ce que l'apprentissage supervisé ? Nommer les cas spéciaux d'apprentissage supervisé selon le type des entrées/sorties (var. catégorielles ou continues)
2. Qu'est ce qu'une fonction de perte ? Qu'est ce que le risque fonctionnel ? Donner des exemples.
3. Qu'est ce que le risque empirique ? Qu'est ce que la minimisation du risque empirique ?
4. Qu'est ce que la généralisation ?
5. Qu'est ce que le sur-apprentissage ou apprentissage par coeur (overfitting) ?

2 Compromis optimisation-approximation-estimation

Dans ce problème, on considère l'espace des couples d'entrée sortie $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ générés par la distribution de probabilité $P(\mathbf{x}, y)$. On définit une fonction de perte $\ell(\hat{y}, y)$ (pa exemple, $\ell(\hat{y}, y) = |\hat{y} - y|^2$ comme en régression) pour mesurer l'écart entre la valeur prédite $\hat{y} = h(\mathbf{x})$ et la sortie réelle y. Le but est de trouver la fonction h^* qui minimise le risque espéré

$$R(h) = \int \ell(h(\mathbf{x}), y) dP(\mathbf{x}, y)$$

La distribution $P(\mathbf{x}, y)$ est généralement inconnue, on a à la place un échantillon $\mathcal{S}$ i.i.d de n exemples d'apprentissage $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$. On définit alors le risque empirique

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)$$

Le principe d'apprentissage vu en cours consiste à choisir en premier une famille $\mathcal{H}$ de fonctions (hypothèses) de prédiction candidates, puis de trouver la fonction $h_n = \arg\min_{h \in \mathcal{H}} R_n(h)$. Puisque l'hypothèse optimale h^* n'appartient pas forcément à la famille $\mathcal{H}$, on définit également $h^*_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(h)$. Par mesure de simplicité on fait l'hypothèse que $h^*, h^*_{\mathcal{H}}$ et h_n sont bien définie et unique. On peut alors décomposer l'excès de l'erreur comme

$$\mathbb{E}[R(h_n) - R(h^*)] = \mathbb{E}[R(h^*_{\mathcal{H}}) - R(h^*)] + \mathbb{E}[R(h_n) - R(h^*_{\mathcal{H}})] = \epsilon_{app} + \epsilon_{est} \quad (1)$$

où l'espérance est prise selon le choix aléatoire de l'ensemble d'apprentissage. L'erreur d'approximation ϵ_{app} mesure avec quelle proximité les hypothèses de $\mathcal{H}$ peuvent approximer la solution optimale h^*. L'erreur d'estimation ϵ_{est} mesure l'effet de la minimisation du risque empirique $R_n(h)$ au lieu du risque espéré (réel) $R(h)$.

Une faille de la décomposition de l'excès d'erreur ci-dessus est que l'on fait l'hypothèse que l'on trouve h_n qui minimise le risque empirique $R_n(h)$. Cependant, cette procédure est souvent une opération lourde en temps de calcul. On suppose que l'algorithme de minimisation retourne une approximation $\tilde{h}_n$ qui minimise une fonction objective à une tolérance prédéfinie $\rho \geq 0$

$$R_n(\tilde{h}_n) \quad < \quad R_n(h_n) + \rho$$

On peut alors décomposer l'excès d'erreur $\epsilon = \mathbb{E}[R(\tilde{h}_n) - R(h^*)]$ comme

$$\epsilon = \mathbb{E}[R(h^*_{\mathcal{H}}) - R(h^*)] + \mathbb{E}[R(h_n) - R(h^*_{\mathcal{H}}) + \mathbb{E}[R(\tilde{h}_n) - R(h_n)]] = \epsilon_{app} + \epsilon_{est} + \epsilon_{opt}$$

On appelle l'erreur additionnelle ϵ_{opt} l'erreur d'optimisation. Elle reflète l'impact de l'optimisation de l'approximation sur la performance en généralisation.

1. On vous demande dans cette question d'étudier comment change l'erreur d'approximation ϵ_{app}, l'erreur d'estimation ϵ_{est},l'erreur d'optimisation ϵ_{opt} et le temps de calcul T lorsque un des éléments suivants $\{\mathcal{H}, n, \rho\}$ augmente. (Augmenter $\mathcal{H}$ signifie que le nouvel ensemble $\mathcal{H}_{nouv.}$ contient l'ancien $\mathcal{H}_{anc.}$ ($\mathcal{H}_{anc.} \subset \mathcal{H}_{nouv.}$). Remplir la table 1 avec $\uparrow$ pour indiquer un accroissement, $\downarrow$ pour indiquer une diminution, et $\times$ pour indiquer non affecté. Expliquer brièvement votre réponse.

	$\mathcal{H}$	n	ρ
ϵ_{app}			
ϵ_{est}			
ϵ_{opt}			
T			

Figure 1: Tableau de variation