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

Exercice 19 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)

   - *L'apprentissage supervisé fait référence à l'apprentissage en présence d'un superviseur. Lorsque l'on essaye d'apprendre à classer des objets, le signal du superviseur est l'étiquette de la classe. Les objets, sont représentés par des vecteurs "$\mathbf{x}$" de variables ou "caractéristiques". On cherche à prédire l'attribut "y" de ces objets, qui est une autre variable. Une variable continue est un nombre réel. Une variable categorielle ou ordinale prend sa valeur parmi un ensemble fini de choix. Pour des variables d'entrée catégorielles la liste n'est pas ordonnée (e.g. le pays d'origine) alors que pour des variables d'entrée ordinales la liste est ordonnée (e.g. trois états clinique dans la progression d'une maladie.)*
   - *Si la variable de sortie y est continue, le problème est un problème de régression; si la variable de sortie y est catégorielle, le problème est un problème de classification.*

2. Qu'est ce qu'une fonction de perte ? Qu'est ce que le risque fonctionnel ? Donner des exemples.

   - *Une fonction de perte est une fonction qui mesure pour un $\mathbf{x}$ donné l'écart entre la sortie prédite $f(\mathbf{x})$ et la sortie réelle y : $\ell(f(\mathbf{x}), y)$.*
   - *Le risque fonctionnel pour une hypothèse h est l'espérance de la fonction de perte:*

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

     On parle également de risque réel, de risque vrai ou de risque en généralisation.

- *Des exemples de fonction de perte sont la perte quadratique utilisée souvent en régression* $(y-f(\mathbf{x}))^2$ *et la perte binaire* $0/1$ *utilisée en classification, qui vaut* 1 *dans le cas d'une erreur et* 0 *autrement.*

3. Qu'est ce que le risque empirique ? Qu'est ce que la minimisation du risque empirique ?

   - *Le risque empirique* R_n *est la perte moyenne sur un nombre fini d'exemples.*

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

   - *La minimisation du risque empirique correspond à la recherche d'une fonction* $f(\mathbf{x})$ *parmi une famille de fonctions qui minimisent le risque empirique.*

4. Qu'est ce que la généralisation ?
   C'est la capacité d'un système prédictif $f(\mathbf{x})$ *à faire de "bonnes" prédictions sur des exemples non rencontrés lors de la construction du modèle. C'est la capacité à étendre une compétence à des exemples non appris.*

5. Qu'est ce que le sur-apprentissage ou apprentissage par coeur (overfitting) ?
   C'est "coller" aux données d'apprentissage (erreur sur l'ensemble d'apprentissage nulle) mais c'est faire de mauvaises prédictions sur de nouveaux exemples.

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

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

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

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

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

$$\mathbb{E}[R(h_n) - R(h^*)] = \mathbb{E}[R(h^*_{\mathcal{H}}) - R(h^*)] + \mathbb{E}[R(h_n) - R(h^*_{\mathcal{H}})] = \epsilon_{app} + \epsilon_{est} \quad (2)$$

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 **??** 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}	$\downarrow$	$\times$	$\times$
ϵ_{est}	$\uparrow$	$\downarrow$	$\times$
ϵ_{opt}	$\times$	$\times$	$\uparrow$
T	$\uparrow$	$\uparrow$	$\downarrow$

Figure 1: Tableau de variation

- Effet sur ϵ_{app}: Lorsque $\mathcal{H}$ devient plus grand, $\mathbb{E}[R(f^*_{\mathcal{H}})]$, le risque du meilleur choix de $f \in \mathcal{H}$ décroit. Donc ϵ_{app} décroit.Lorsque n augmente, il n'y a pas d'effet parceque l'expression du vrai risque ne dépend pas de n.

- Effet ϵ_{est}: Lorsque $\mathcal{H}$ devient plus grand, $\mathbb{E}[R(f^*_{\mathcal{H}})]$, le risque du meilleur choix de $f \in \mathcal{H}$ décroit. Donc par définition ϵ_{est} s'accroit. Plus n augmente ,plus l'erreur diminue.

- Effet sur ϵ_{opt}: Lorsque ρ croit, la tolérance devient plus grande est une solution moins précise est acceptable. Donc l'erreur augmente.