Apprentissage Statistique en Grande Dimension-TD 1

Les exercices de ce premier chapitre sont en majorité consacrés à l'apprentissage supervisé.

Exercice 1. Dans les exemples suivants, dans quels cas pensez-vous qu'il est judicieux de choisir une méthode d'apprentissage flexible ?

1. Le nombre n d'observations est grand et le nombre de prédicteurs p est petit.
2. n est petit et p est grand.
3. La relation entre $\mathbf{x}$ et y semble vraiment non-linéaire.
4. La variance du terme d'erreur ε est grande (Considérer un modèle de régression $Y = f(\mathbf{X}) + \varepsilon$).

Dans les situations 1, 3 et 4, on pourra tenter d'apporter un début d'explication théorique.

Exercice 2 (Décomposition Biais/Variance)**.** On considère un modèle de régression

$$Y = f(\mathbf{X}) + \varepsilon$$

où ε est une variable aléatoire indépendante du couple $(\mathbf{X}, Y)$, centrée et de variance finie. Supposons que ℓ soit la fonction de perte définie par

$$\ell(y, y') = (y - y')^2.$$

Soit $\hat{f}$ une fonction de prévision (obtenue par l'apprentissage d'un échantillon observé) supposée déterministe (ou plus exactement construite sur un échantillon indépendant). On pose $\hat{Y} = \hat{f}(X)$.

1. Montrez que
$$\mathbb{E}[(Y - \hat{f}(X))^2] = \mathrm{Var}((\hat{f} - f)(X)) + \mathrm{Var}(\varepsilon) + \mathrm{Biais}^2$$
où $\mathrm{Biais} = \mathbb{E}[\hat{f}(X)] - \mathbb{E}[f(X)]$.
2. Préciser le calcul lorsque $f_\theta(x) = x\theta$ avec $x \in \mathbb{R}^n \times \mathbb{R}^p$ et $\theta \in \mathbb{R}^p$.
3. On regarde maintenant les choses conditionnellement à $\mathbf{X}$. Plus exactement, on suppose que pour tout $\mathbf{x} \in \mathbb{R}^d$, la réponse $Y(\mathbf{x})$ est réelle et de la forme $Y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$ (ε peut par exemple être vue comme l'incertitude d'un modèle déterministe). On considère maintenant $\hat{f}$ sous sa forme véritable, *i.e.* comme une variable aléatoire fonction de l'échantillon $\mathcal{D}_n$ observé).
   (a) Dans ce cas, que vaut $\mathbb{E}[(Y(\mathbf{x}) - \hat{f}(\mathbf{x}))^2]$?
   (b) Sauriez-vous indiquer intuitivement comment les quantités en jeu dépendent de la flexibilité de l'algorithme ?

4. Risque d'entraînement/Risque de Test ("Vrai Risque"). Rappelez la définition de ces deux quantités. Comment selon vous peut-on imaginer l'évolution du risque d'entraînement en fonction de la flexibilité du modèle ? Qu'en est-il du "vrai risque" ? Pour répondre à cette seconde question, placez-vous dans le cas précédent. (N.B. *Compromis Biais/Variance*)

Exercice 3. Dans les exemples suivants, dégagez les situations de régression ou de classification. Déterminez n et p.

1. On collecte les données de 500 entreprises. Pour chacune d'entre elles, on enregistre le chiffre d'affaires, le nombre d'employés, l'âge moyen des employés et le salaire moyen. On cherche ici à comprendre quels facteurs influent sur le salaire moyen par entreprise.
2. On considère un nouveau produit pour lequel on souhaite évaluer si sa mise en vente sera un succès ou un échec. Pour cela, on collecte les données de 20 produits similaires pour améliorer notre pronostic. Plus précisément, sont collectés : le prix de fabrication, le budget marketing, le prix le plus compétitif du marché ainsi que 10 autres variables. Par ailleurs, pour chaque produit, on sait également si sa mise en vente a été une réussite ou non.
3. On cherche ici à prédire le taux de change du Dollar pour la semaine suivante. Pour cela on collecte les données semaine par semaine sur l'année précédente de du taux de change du Dollar ainsi que des prix de 100 produits boursiers.

Exercice 4 (Validation Croisée)**.**

1. Le coût de calcul du prédicteur $\hat{f}_n$ peut être très long dans certaines situations (si par exemple, construit comme la solution non explicite d'un problème de minimisation). Ainsi si le nombre de classes K utilisé dans la validation croisée est important, cette procédure peut alors être très coûteuse. Selon vous, quel procédé peut-on envisager pour réduire le temps de calcul (si la structure de calcul le permet) ?
2. Supposons que l'échantillon est de taille n et que les ensembles $I_1, \dots, I_K$ sont de même cardinal r. Montrez que dans ce cas,

$$\mathbb{E}[\hat{R}_{CV}] = \mathbb{E}[\phi(\mathcal{D}_{n-r})]$$

où $\mathcal{D}_{n-r}$ est un échantillon de taille $n-r$, ϕ est une fonction de $(\mathcal{X} \times \mathcal{Y})^{n-r}$ vers $\mathbb{R}$ définie par :

$$\phi(\mathbf{d}_{n-r}) = \mathbb{E}[\ell(Y, \hat{f}_{\mathbf{d}_{n-r}}(\mathbf{X}))]$$

avec $\mathbf{d}_{n-r} = (\mathbf{x}_i, y_i)_{i=1}^{n-r}$ (déterministe).

3. On suppose que le modèle est de la forme $Y = f(\mathbf{X}) + \varepsilon$ (ε centrée indépendante de $\mathbf{X}$) et que $\ell(y, y') = (y - y')^2$. Explicitez la fonction ϕ.
4. On suppose dans cette question que le prédicteur est faiblement consistant. En déduire la limite de $\mathbb{E}[\hat{R}_{CV}]$ (on suppose que le nombre de folds K ne dépend pas de n).
5. Quels problèmes se poseraient-ils si l'on envisageait de calculer $\mathrm{Var}(\hat{R}_{CV})$? (Ceci explique en partie les limites théoriques de cette méthode.)

Exercice 5 (1-ppv)**.** Comme cela a été expliqué en cours, la manière la plus naturelle de mimer le prédicteur optimal de Bayes, consiste à approcher les probabilités conditionnelles $\mathbb{P}(Y = y|\mathbf{X} = x)$ (cas où $\mathcal{Y}$ discret) par une approximation locale de cette quantité. L'algorithme

des k-plus proches voisins en est une. On s'intéresse ici à sa mise en oeuvre sur un exemple simple de classification binaire puis nous focalisons sur la non-consistance du plus proche voisin ("1-ppv"). On pose $\mathcal{X} = [0,1]$ et $\mathcal{Y} = \{0,1\}$.

1. L'échantillon d'apprentissage est $(\mathbf{X}_1 = 0.8, Y_1 = 1)$, $(\mathbf{X}_2 = 0.4, Y_2 = 0)$, $(\mathbf{X}_3 = 0.7, Y_3 = 1)$. Donnez la valeur prédite pour toute nouvelle entrée $x \in \mathcal{X}$
   (a) par l'algorithme des 3-p.p.v.
   (b) par l'algorithme du p.p.v.
2. Dans cette question, on suppose que la loi $\mathbb{P} = \mathcal{L}(\mathbf{X}, Y)$ est la suivante : $\mathbf{X}$ suit la loi uniforme sur $[0,1]$ et, $Y = 1$ si $\mathbf{X} > 0.5$ et $Y = 0$ si $\mathbf{X} \leq 0.5$.
   (a) Donnez $\mathbb{P}(Y = 1 | \mathbf{X} = x)$ pour tout $x \in \mathcal{X}$. Qu'en pensez-vous ?
   (b) En déduire le prédicteur de Bayes (prédicteur optimal). Quel est son risque ?
   (c) Soit $\mathcal{D}_n$ un échantillon d'apprentissage (ou une base d'apprentissage). Soit E l'évènement : "tous les Y_i sont identiques". Que peut-on dire de l'algorithme du 1-ppv sur cet évènement ? Quelle est sa probabilité ?
   (d) En notant

   $$\mathcal{C}_i(\mathbf{X}) = \{x \in [0,1], \min_{k=1}^{n}(|\mathbf{X}_k - x|) = |\mathbf{X}_i - x|\} \quad \text{(Cellule de Voronoï, "}p.s.\text{ bien définie")}$$

   définir l'algorithme du plus proche voisin (noté à nouveau $\hat{f}$). Donnez une expression simple du risque associé.
   (e) Montrez que pour une suite de v.a. i.i.d. de loi $\mathcal{U}_{[0,1]}$, pour tout $r \in [0, 1/2]$,

   $$\mathbb{P}(\min_{k=1}^{n} |\mathbf{X}_k - \frac{1}{2}| > r) = (1 - 2r)^n.$$

   En déduire que

   $$\mathbb{E}[\min_{k=1}^{n} |\mathbf{X}_k - \frac{1}{2}|] = \frac{1}{2(n+1)}.$$

   (f) Expliquer alors pourquoi dans ce cas que l'algorithme du plus proche voisin est consistant et quantifier le risque de cet algorithme (en réalité, il faudrait travailler avec la 2ème statistique d'ordre mais c'est plus difficile).
3. Dans cette question, on suppose que la loi $\mathbb{P} = \mathcal{L}(\mathbf{X}, Y)$ est la suivante : $\mathbf{X}$ suit la loi uniforme sur $[0,1]$ et la loi conditionnelle de Y sachant $\mathbf{X} = x$ est donnée par

   $$\mathbb{P}(Y = 1 | \mathbf{X} = x) = \frac{2}{3} = 1 - \mathbb{P}(Y = 0 | \mathbf{X} = x).$$

   (a) Définir le prédicteur de Bayes dans ce cas. Quel est son risque [1] ?
   (b) On cherche maintenant à estimer le risque de l'algorithme du plus proche voisin pour ce modèle.

1. On rappelle ici que ce risque est "indépassable" au sens où aucun algorithme ne pourra prétendre avoir un risque inférieur au risque de Bayes. Celui-ci est inhérent au modèle.

i. Définir l'algorithme dans ce cas (en fonction de $\mathcal{D}_n$, par définition d'un prédicteur).

ii. Montrez que $\mathbf{X}$ et Y sont indépendants (On pourra calculer $\mathbb{P}(\mathbf{X} \in [a,b], Y = 1)$).

iii. Donnez une expression du risque de classification (à l'aide des cellules de Voronoï). On suppose la valeur de $\mathbf{X}$ connue. Donnez une expression du risque de classification (conditionnellement à $\mathbf{X} = x$).

iv. Montrez que le risque est égal à

$$2\mathbb{P}(Y = 1)\mathbb{P}(Y = 0).$$

v. Qu'en déduit-on quant à la consistance du plus proche voisin ?

Exercice 6 (Classification binaire (suite))**.** Soit un problème d'apprentissage où Y est à valeurs dans $\{-1, 1\}$ et $Y|\mathbf{X} = x$ suit une loi de Rademacher de paramètre $p(x)$ où $p(x) \in [0, 1]$.

1. Rappeler la définition du prédicteur optimal $f^\star$ en fonction de $p(x)$ lorsque la fonction de perte ℓ est définie par $\ell(y, y') = 1_{y \neq y'}$.
2. On considère la fonction de perte $\ell(y, y') = \log(1 + e^{-yy'})$. Déterminez le prédicteur optimal dans ce cas, *i.e.* déterminez la fonction f (si elle existe) minimisant $\mathbb{E}[\ell(Y, f(\mathbf{X})]$. Conclusion ?
3. On s'intéresse au problème de minimisation suivant : déterminez parmi les fonctions $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty, +\infty\}$, celle qui minimise $\mathbb{E}[\ell(Y, f(\mathbf{X})]$. Notons à nouveau $f^\star$ cette fonction. De quelle fonction usuelle s'agit-il ?

N.B. Cet exercice fait un lien avec la régression logistique. Dans le cadre des Modèles linéaires généralisés, on fait généralement l'hypothèse que la fonction p satisfait

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \langle \theta, x \rangle$$

où $\theta \in \mathbb{R}^p$. En d'autres termes, on fait une hypothèse de linéarité pour une transformation adéquate de la loi de Y sachant $\mathbf{X}$. Cette transformation est généralement appelée fonction de lien.

Exercice 7 (Risque asymétrique/Courbe ROC)**.** Comme cela a été expliqué en cours, les fonction de risque usuelles ne sont pas toujours adaptées au problème considéré. L'erreur de classification standard par exemple (en classification binaire) ne prend pas en compte l'importance des quantités en jeu. Par exemple, en médecine, il est usuellement plus grave de classer parmi les malades une personne saine plutôt qu'une personne saine parmi les malades. La courbe ROC (Receiver operating characteristic) est un outil venu du traitement du signal permettant de prendre en compte cette dissymétrie. On suppose que la sortie Y est égale à 1 si l'individu est malade et 0 sinon.

1. On note $\eta(x) = \mathbb{P}(Y = 1|\mathbf{X} = x)$ et on considère la règle de décision suivante : pour un seuil s fixé appartenant à $[0, 1]$,

$$f_s^\star(x) = \begin{cases} 1 & \text{si } \eta(x) > s \\ 0 & \text{sinon.} \end{cases} \tag{1}$$

Montrez que $f_s^\star$ minimise le risque associé à la fonction de perte :

$$\ell(y, y') = (1-s)1_{\{y=1,y'=0\}} + s1_{\{y=0,y'=1\}}.$$

Indication : On pourra remarquer que dans le cas $s = 1/2$, on retrouve le cas classique.

2. Dans cette partie, on se place dans le cas particulier de l'analyse discrimante linéaire :

$$\mathcal{L}(X|Y=i) = \mathcal{N}(m_i, \sigma_i^2), \quad i = 0, 1$$

et $\mathbb{P}(Y=1) = p$. On suppose également que $m_0 < m_1$.

(a) Représentez graphiquement cette situation.

(b) Pour s fixé, explicitez une règle de décision de la forme (1).

(c) On suppose dans cette question que $\sigma_1 = \sigma_2$. Dans ce cas, montrez que la règle peut s'écrire :

$$f(x) = \begin{cases} 1 & \text{si } x > \alpha_s \\ 0 & \text{sinon,} \end{cases}$$

où α_s est un seuil à définir.

(d) Notons Se et Sp les fonctions (Sensibilité et Spécificité) définies par

$$\text{Se}(\alpha) = \mathbb{P}(X > \alpha | Y = 1) \quad \text{et} \quad \text{Sp}(\alpha) = \mathbb{P}(X \leq \alpha | Y = 0).$$

A quoi correspondent ces quantités (vrai/faux,positif/négatif...) ? Reformulez ces quantités en termes de risque de 1ère espèce/2ème espèce (en supposant que H_0 ="Individu malade").

(e) Dans ce cadre (de la LDA), la courbe ROC est alors le graphe de

$$\{(1 - \text{Sp}(\alpha), \text{Se}(\alpha), \alpha \in \mathbb{R}\}.$$

Tracez cette courbe avec $m_1 = 1$ et $m_2 = 2$ ou $m_2 = 3$ avec dans chaque cas $\sigma_1 = \sigma_2 = 1$.

(f) Montrez que si F est la fonction de répartition de $\mathcal{L}(X|Y=0)$ et G, celle de $\mathcal{L}(X|Y=1)$, alors la courbe ROC est aussi le graphe de la fonction

$$\text{ROC}(t) = 1 - G(F^{-1}(t)), \quad t \in]0, 1[.$$

(g) Dans ce qui précède, les calculs sont effectués sous un point de vue "oracle", *i.e.* en travaillant sous l'hypothèse que la loi du couple (X, Y) est connue. A l'aide de ce qui précède, proposez un prédicteur basé sur un échantillon d'apprentissage $\mathcal{D}_n$ (On pourra s'appuyer sur des estimateurs usuels de la moyenne et de la variance). Ce prédicteur est-il consistant ?

(h) Pour un échantillon donné, tracez la courbe ROC correspondante (en remplaçant les fonctions de répartition par des fonctions de répartition empiriques).

N.B. La courbe ROC est un outil de mesure de la qualité de classification binaire assez usuel. Pour comparer différents classifieurs, on peut mesure l'aire sous la courbe appelée aire AUC.

3. Dans la partie précédente, on a fait des hypothèses fortes sur le modèle. En pratique, il n'est souvent pas envisageable d'utiliser cette approche. Sauriez-vous construire une approche non paramétrique de type K-plus proches voisins pour fabriquer un algorithme de décision avec *sensibilité s* ? (On s'appuiera sur une approximation appropriée de $\mathbb{P}(Y = 1|\mathbf{X} = x)$?