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. Si le nombre n d'observations est grand et le nombre de variables 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 (Règles de Bayes)**.**

1. Démontrez le point (i) du théorème 1.2.5.
2. Démontrez le point (iii) du théorème 1.2.5 dans le cadre multiclasses.

Exercice 3. On considère un modèle de régression sur $\mathbb{R}^p$ et on note ℓ la fonction de perte définie par

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

Notons $f^\star(X) = \mathbb{E}[Y|X]$.

1. Montrez que

$$Y = f^\star(X) + \varepsilon \quad \text{avec } \mathbb{E}[\varepsilon] = 0 \text{ et } \mathrm{Cov}\,(g(X), \varepsilon) = 0. \text{ pour toute fonction } g \text{ mesurable bornée.}$$

2. En déduire que pour une règle de décision $f : \mathbb{R}^p \to \mathbb{R}$, l'excès de risque satisfait

$$R_f - R_{f^\star} = \mathbb{E}[(f(X) - f^\star(X))^2].$$

3. En déduire que pour un prédicteur $\hat{f}_n$ (bâti sur un échantillon d'apprentissage indépendant du couple (X, Y), le risque moyen vaut

$$\mathbb{E}[R_{\hat{f}_n}] = \mathbb{E}[(\hat{f}_n(X) - f^\star(X))^2] + \mathbb{E}[\varepsilon^2].$$

4. On suppose maintenant que l'on souhaite calculer le risque en un point $\mathbf{x}$. Montrez que

$$\mathbb{E}[\ell(Y, \hat{f}_n(X))|X] = \Phi(X) \quad \text{où } \Phi(\mathbf{x}) = \mathrm{Var}((\hat{f}_n(\mathbf{x})) + \mathrm{Var}(\varepsilon) + (\mathbb{E}[\hat{f}_n(\mathbf{x})] - f^\star(\mathbf{x}))^2.$$

5. Sauriez-vous indiquer comment les quantités en jeu dans l'expression ci-dessus dépendent de la flexibilité de l'algorithme ? (On pourra considérer l'exemple des k-ppv).
6. Rappelez la définition du risque d'entraînement et du risque de test ces deux quantités (dans ce cadre). Expliquez dans le cadre du 1-ppv (quantitatif) pourquoi le risque d'entraînement tend vers 0.

Exercice 4. 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 du taux de change du Dollar ainsi que des prix de 100 produits boursiers.

Exercice 5 (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 $I_1, \ldots, I_K$ 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 6 (Données manquantes/Données "mal balancées")**.** 1. On dispose de N individus, p variables et $m < N$ colonnes où il manque une variable.
   — Que décidez-vous si N est grand devant p et m ?
   — Que décidez-vous si vous n'êtes pas dans la première situation ? Quels sont les points à considérer pour tenter de compléter le jeu de données de manière raisonnable ?

2. — On dispose de données dont l'effectif est mal équilibré. Quelles peuvent être les conséquences lorsque l'on fait de la prédiction ?
   — Comment pourriez-vous modifier la fonction de perte usuelle en classification pour pallier ce problème ?

Exercice 7 (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 $(X_1 = 0.8, Y_1 = 1)$, $(X_2 = 0.4, Y_2 = 0)$, $(X_3 = 0.7, Y_3 = 1)$. 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 $\mathbb{P} = \mathcal{L}(X, Y)$ est la suivante : X suit la loi uniforme sur $[0, 1]$ et, $Y = 1$ si $X > 0.5$ et $Y = 0$ si $X \leq 0.5$.
   (a) Donnez $\mathbb{P}(Y = 1|X = x)$ 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 introduisant la statistique d'ordre $(X^{(i)})_{i=1}^n$ associée à l'échantillon $(X_i)_{i=1}^n$, donner une expression simple de l'algorithme du plus proche voisin lorsque $\omega \in E^c$ (noté à nouveau $\hat{f}$). Montrez également que sur E^c,

$$R_{\hat{f}} = \left| \frac{X^{(i^\star)} + X^{(i^\star+1)}}{2} - \frac{1}{2} \right|.$$

Rappel : La statistique d'ordre associée à l'échantillon $(X_i)_{i=1}^n$ est la suite réordonnée des X_i : si $X_{k_1} < X_{k_2} < \ldots < X_{k_n}$, alors $X^{(i)} = X_{k_i}$.

(e) Notons $X_i^+ = (X_i - \frac{1}{2})1_{X_i > 1/2}$ et $X_i^- = (\frac{1}{2} - X_i)1_{X_i < 1/2}$. Montrez que

$$\mathbb{E}[R_{\hat{f}}] \leq \frac{1}{2}\left(\mathbb{E}[\min_{i=1}^n X_i^+] + \mathbb{E}[\min_{i=1}^n X_i^-]\right) + \mathbb{P}(E).$$

(f) Montrez que

$$\mathbb{E}[\min_{i=1}^n X_i^+] = \mathbb{E}[\min_{i=1}^n X_i^-].$$

(g) Montrez que pour tout $r \in [0, 1/2]$,

$$\mathbb{P}(\min_{i=1}^n X_i^+ > r) = \left(\frac{1}{2} - r\right)^n.$$

En déduire par le lemme de Wald que

$$\mathbb{E}[X_i^+] = \frac{2^{-n-1}}{n+1}.$$

(h) En déduire la consistance du 1-ppv dans ce cadre.

3. Dans cette question, on suppose que la loi $\mathbb{P} = \mathcal{L}(X, Y)$ est la suivante : X suit la loi uniforme sur $[0, 1]$ et la loi conditionnelle de Y sachant $X = x$ est donnée par

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

(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.

i. Définir l'algorithme dans ce cas (en fonction de $\mathcal{D}_n$, par définition d'un prédicteur). On pourra utiliser les *cellules de Voronoï* associées à l'échantillon :

$$\mathcal{C}_i = \{x \in [0, 1], \min_{k=1}^n (|X_k - x|) = |X_i - x|\}.$$

N.B. Les $\mathcal{C}_i$ ne forment pas exactement une partition car les frontières sont communes mais comme la loi uniforme est continue, on peut raisonner en faisant comme tel.

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

iii. Donnez une expression du risque de classification (à l'aide des cellules de Voronoï).

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 ?

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.

Exercice 8 (Théorème de Stone)**.** On s'intéresse à une méthode par partition (version simplifiée des k-ppv), *i.e.* dans le cas où $\mathcal{X}$ est un compact de $\mathbb{R}^p$. On note $(A_k^{(\varepsilon)})_{k=1}^{k_\varepsilon}$ une partition de $\mathcal{X}$ telle que $\sup_{k=1}^{k_\varepsilon} \mathrm{diam}(A_k^{(\varepsilon)}) \leq \varepsilon$. Notons $\mathcal{I}_k^\varepsilon = \{i \in \{1,\dots,n\}, X_i \in A_k^{(\varepsilon)}\}$ (cet ensemble est aléatoire). Dans le cas de la régression, le prédicteur est alors défini par :

$$\forall x \in A_k^{(\varepsilon)}, \quad \hat{f}(x) = \frac{1}{|\mathcal{I}_k^\varepsilon|} \sum_{i \in \mathcal{I}_k^\varepsilon} Y_i.$$

Discutez de la consistance de cette méthode (lorsque ε tend vers 0) (pour une démonstration précise et plus générale du théorème de Stone, voir le cours de S. Arlot accessible sur Moodle, on y trouvera également de nombreux exercices sur le sujet).

Exercice 9 (Classification binaire)**.** 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. 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 ?

   Indication : On pourra commencer par montrer que $R_f = \mathbb{E}[\Psi_f(X)]$ avec

   $$\Psi_f(x) = \log(1+e^{-f(x)})p(x) + \log(1+e^{f(x)})(1-p(x)).$$

2. 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 ce cadre, 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 10 (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 discriminante 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 $\sigma_0 = \sigma_1$. 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

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

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

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

Tracez cette courbe avec $m_0 = 1$ et $m_1 = 2$ ou $m_1 = 3$ avec dans chaque cas $\sigma_0 = \sigma_1 = 1$.

(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

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

(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 mesurer 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. Construisez une courbe ROC basée sur les effectifs de faux/vrais positifs/négatifs.
4. Dans la littérature, on utilise aussi le F_1-score comme mesure de la qualité d'un algorithme. Etudiez la construction du F_1-score et discutez l'intérêt de cette mesure. Même question pour le coefficient de corrélation de Matthews.

Exercice 11. Ce dernier exercice a vocation à jouer le rôle de mémo sur les méthodes non supervisées classiques. Parcourez le document ci-joint `http://www2.agroparistech.fr/IMG/pdf/ClassificationNonSupervisee-AgroParisTech.pdf` et listez les méthodes connues.