Apprentissage Statistique en Grande Dimension-TD 1 Corrigé

Exercice 1.

1. Si le nombre n d'observations est grand et le nombre de variables p est petit, alors on peut envisager un modèle flexible. Du point de vue de l'approximation locale, on peut considérer que si p est petit devant n, alors chaque point x de l'espace $\mathbb{R}^p$ dispose d'un nombre important de voisins de sorte que $\mathbb{P}(Y = y|X = x)$ ou $\mathbb{E}[Y|X = x]$ seront bien approximées par la moyenne de leurs réponses.
2. Dans le cas inverse, on ne peut envisager un modèle flexible.
3. Si la relation entre $\mathbf{x}$ et y semble vraiment non-linéaire ou plus précisément si l'on n'a pas de modélisation claire par un modèle paramétrique usuel, alors on doit si possible envisager un modèle flexible. Néanmoins, si le nombre n est également petit devant p, le problème devient statistiquement compliqué!! (Tout les situations ne sont pas "explicables"...)
4. Lorsque la variance du terme d'erreur ε est grande, on doit réduire la flexibilité du modèle. En particulier, si l'objectif est la **prédiction** de Y, trop de précision sur f ne sert à rien. Evidemment, dans la pratique, on ne sait pas à l'avance si les variations sont issues du bruit ou d'une forte irrégularité de f. C'est justement la mise en oeuvre de plusieurs algorithmes et la comparaison des erreurs qui permet de décider. Si l'objectif est de "filtrer le signal", *i.e.* d'identifier f, alors ça peut être envisagé si n est assez grand.

Exercice 2. *(i)* C'est une conséquence de la propriété de projection orthogonale. En effet,

$$\begin{aligned}\mathbb{E}[\ell(Y, f(X)] = \mathbb{E}[(Y - f(X))^2] &= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + 0 + \mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2] \\ &\geq \mathbb{E}[\ell(Y, \mathbb{E}[Y|X])].\end{aligned}$$

Il y a de plus égalité lorsque $f(x) = \mathbb{E}[Y|X = x]$ d'où le résultat.

(ii) On reprend les lignes de la preuve dans le cas binaire. On note $\eta(x, k) = \mathbb{P}(Y = k|X = x)$.

$$\begin{aligned}\mathbb{P}(Y \neq f(X)) = \sum_{k=1}^{K} \mathbb{P}(Y \neq k, f(X) = k) &= \sum_{k=1}^{K} \mathbb{E}[(1 - \eta(X, k))1_{f(X)=k}] \\ &\geq \sum_{k=1}^{K} \mathbb{E}[(1 - \max_{k=1}^{K} \eta(X, k))].\end{aligned}$$

Il reste à montrer que ce minorant est un minimum. Pour cela, on fixe

$$f^\star(x) = \text{Argmax}_{k=1}^{K} \mathbb{P}(Y = k|X = x) = \text{Argmax}_{k=1}^{K} \eta(x, k).$$

Par construction,

$$f^\star(x) = k \quad \Longleftrightarrow \quad \eta(x, k) = \max_{j=1}^{J} \eta(x, j).$$

En conséquence,

$$\sum_{k=1}^{K} \mathbb{E}[(1 - \eta(X, k))1_{f^\star(X)=k}] = \sum_{k=1}^{K} \mathbb{E}[(1 - \max_{k=1}^{K} \eta(X, k))].$$

ce qui termine la preuve.

N.B. Dans la plupart des preuves de ce type, on travaille comme si l'argmax était unique. Pour être parfaitement rigoureux, il faudrait aussi considérer le cas d'égalité. Si le max est atteint en au moins deux classes, il faut déterminer une règle arbitraire d'attribution de la classe (par exemple, si les classes sont numérotées, on peut choisir la classe par ordre croissant, on peut aussi tirer au sort...). La suite de la preuve s'en trouve inchangée.

Exercice 3. 1. On a

$$Y = f^\star(X) + \varepsilon \quad \text{où } \varepsilon = Y - \mathbb{E}[Y|X].$$

Or,

$$\mathbb{E}[Y - \mathbb{E}[Y|X]] = \mathbb{E}[Y] - \mathbb{E}[Y] = 0,$$

tandis que

$$\mathrm{Cov}\,(g(X), \varepsilon) = \mathbb{E}[g(X)\varepsilon] = \mathbb{E}[g(X)(Y - \mathbb{E}[Y|X])] = \mathbb{E}[g(X)Y] - \mathbb{E}[\mathbb{E}[g(X)Y|X]] = 0.$$

2.

$$R_f = \mathbb{E}[(Y - f(X))^2] = \mathbb{E}[(f^\star(X) + \varepsilon - f(X))^2] = \mathbb{E}[(f(X) - f^\star(X))^2] + \mathbb{E}[\varepsilon^2],$$

car $\mathbb{E}[(f(X) - f^\star(X))\varepsilon] = 0$ d'après la question précédente. Or

$$R_{f^\star} = \mathbb{E}[(Y - f^\star(X))^2] = \mathbb{E}[\varepsilon^2],$$

d'où le résultat.

3. Comme $\hat{f}_n$ est construit sur un échantillon indépendant $\mathcal{D}_n$, on a en conditionnant par $\mathcal{D}_n$,

$$\mathbb{E}[R_{\hat{f}_n} - R_{f^\star}] = \mathbb{E}[\Psi(\hat{f}_n)] \quad \text{où } \Psi(f) = R_f - R_{f^\star}.$$

Le résultat suit grâce à la question précédente.

4. Si on reprend les calculs précédents avec un conditionnement par rapport à X et une fonction f déterministe, on obtient

$$\mathbb{E}[\ell(Y, f(X))|X] = \mathbb{E}[(f(X) - f^\star(X))^2|X] + \mathbb{E}[\varepsilon^2|X] = (f(X) - f^\star(X))^2 + \mathbb{E}[\varepsilon^2].$$

Avec $\hat{f}_n$, le calcul est le même par indépendance. Il vient

$$\mathbb{E}[\ell(Y, \hat{f}_n(X))|X] = \Phi(X)$$

avec

$$\begin{aligned}\Phi(\mathbf{x}) &= \mathbb{E}[(\hat{f}_n(\mathbf{x}) - f^\star(\mathbf{x})^2] + \mathbb{E}[\varepsilon^2] \\ &= \mathrm{Var}((\hat{f}_n(\mathbf{x})) + \mathrm{Var}(\varepsilon) + (\mathbb{E}[\hat{f}_n(\mathbf{x})] - f^\star(\mathbf{x}))^2.\end{aligned}$$

5. Pour fixer les idées, considérons un algorithme à moyennisation locale, *i.e.* pour un échantillon $(X_i, Y_i)_{i=1}^n$ donné, notons $\mathcal{V}_{x,\delta} = \{i, |X_i - x| \leq \delta\}$ puis considérons le prédicteur de $Y(x)$ défini par

$$\hat{f}(x) = \frac{1}{\mathrm{Card}(\mathcal{V}_{x,\delta})} \sum_{i \in \mathcal{V}_{x,\delta}} Y_i. \tag{1}$$

Dans ce cas, plus δ est petit, moins il y a de points dans la boule de centre x et de rayon δ, donc plus il y a de variance [1]. En effet, on rappelle que si $(Y_i)_{i=1}^N$ désigne une suite de variable *i.i.d.*, alors

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N Y_i\right) = \frac{\mathrm{Var}(Y_1)}{N}.$$

Ainsi, plus N est petit, plus la variance est grande (Ici, la difficulté est que N est aléatoire car il s'agit de $\mathrm{Card}(\mathcal{V}_{x,\delta})$ mais l'idée reste correcte). A l'inverse, plus δ est petit, plus le biais (conditionnel) est petit car

$$\mathbb{E}[\hat{f}(x)] \approx \mathbb{E}[f(X_1)/X_1 \in B(x, \delta)]$$

qui est d'autant plus proche de $f(x)$ que δ est petit.

1. Remarquons que lorsque δ devient trop petit, alors, il peut ne plus y avoir de points dans $\mathcal{V}_{x,\delta}$, ce qui pose problème pour l'équation (1). Pour éviter ce problème, considérer le plus proche voisin lorsque $\mathcal{V}_{x,\delta}$ est vide.

6. Le risque d'entraînement R_{train} est défini (dans le cadre de la régression et des moindres carrés) par

$$R_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2,$$

où $(X_i, Y_i)_{i=1}^n$ désigne l'échantillon qui a servi à construire $\hat{f}$. Supposons que X_i est à densité dans $\mathbb{R}^p$ de sorte que $\mathbb{P}(X_i = X_j) = 0$ pour tous i, j tels que $i \neq j$. Dans ce cas, pour le 1-ppv, X_i est le plus proche voisin de $\ldots X_i$ de sorte que $\hat{f}(X_i) = Y_i$ et donc $R_{\text{train}} = 0$. Pour l'exemple ci-dessus où on fait la moyenne sur une boule de centre x et de rayon δ, on constate que lorsque $\delta \to 0$, $\text{Card}(\mathcal{V}_{X_i,\delta}) \to 1$. Ainsi, à nouveau, $\hat{f}(X_i) = Y_i$ pour δ assez petit, et on comprend donc qu'à n fixé,

$$\lim_{\delta \to 0} R_{\text{train}} = 0,$$

lorsque n est grand.
Une fois le prédicteur construit sur l'échantillon $(X_i, Y_i)_{i=1}^n$, le ("vrai") risque de test est lui défini, pour un échantillon indépendant $(\tilde{X}_i, \tilde{Y}_i)_{i=1}^N$ par

$$R_{\text{test}} = \frac{1}{N}\sum_{i=1}^{N}(\tilde{Y}_i - \hat{f}(\tilde{X}_i))^2 \approx \mathbb{E}[R_{\hat{f}}(X)] \quad \text{(Risque moyen)},$$

lorsque N est grand. Le comportement de R_{test} avec la flexibilité n'est pas clair car il dépend à la fois du biais et de la variance induites par $\hat{f}$. On peut simplement mentionner (à nouveau) que lorsque δ est trop petit, la variance induite par $\hat{f}$ risque de faire augmenter R_{test} tandis que lorsque δ est trop grand, R_{test} risque d'augmenter à cause du biais.

Exercice 4. Dans les exemples suivants, dégagez les situations de régression ou de classification. Déterminez n et p.

1. Régression (car la réponse est quantitative), $n = 500$, $p = 3$. 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. Classification (Réponse = Succès ou Echec), $n = 20$, $p = 13$.
3. Régression (Réponse = Taux de change). Si les taux de change semaine à semaine sont supposés indépendants, alors on a un modèle de régression standard avec $n = 365$, $p = 100$. Néanmoins, cette hypothèse est très discutable. Elle pourrait amener à considérer la suite des observations comme une série chronologique avec une dynamique d'évolution sous-jacente.

Exercice 5 (Validation Croisée)**.** 1. Le calcul de l'erreur de validation croisée est parallélisable.

2. Notons $r = \text{card}I_k$. On a

$$\mathbb{E}[\hat{R}_{CV}] = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r}\sum_{i\in I_k}\mathbb{E}[\ell(Y_i, \hat{f}_{-I_k}(X_i))].$$

Comme $\mathcal{D}_{-I_k}$ est indépendant de (X_i, Y_i), $i \in I_k$ et que la loi de $\mathcal{D}_{-I_k}$ et de (X_i, Y_i) ne dépend pas de k et i, on en déduit que $\mathbb{E}[\ell(Y_i, \hat{f}_{-I_k}(X_i))]$ ne dépend pas de k et i. De plus, comme $\mathcal{D}_{-I_k}$ a même loi que $\mathcal{D}_{n-r}$, il vient

$$\mathbb{E}[\ell(Y_i, \hat{f}_{-I_k}(X_i))] = \mathbb{E}[\ell(Y, \hat{f}_{\mathcal{D}_{n-r}}(X))]$$

avec $\mathcal{D}_{n-r}$ indépendant de (X, Y). On en déduit le résultat en écrivant

$$\int \mathbb{E}[\ell(Y, \hat{f}_{d_{n-r}}(X))]\mathbb{P}_{\mathcal{D}_{n-r}}(d_{n-r})$$

puis en remarquant que cette quantité ne dépend pas de k. L'erreur de validation croisée est donc un estimateur sans biais du risque moyen.

3. Lorsque $Y = f(X) + \varepsilon$, alors

$$\phi(d_{n-r}) = \mathbb{E}[(Y - \hat{f}_{d_{n-r}}(X))^2] = \mathbb{E}[(f - \hat{f}_{d_{n-r}}(X))^2] + \mathbb{E}[\varepsilon^2].$$

4. Un prédicteur est faiblement consistant si le risque moyen tend vers le risque optimal. Or,

$$\mathbb{E}[\ell(Y, \hat{f}_{\mathcal{D}_{n-r}}(X))] = \mathbb{E}[R_{\hat{f}_{\mathcal{D}_{n-r}}}]$$

ce qui implique que lorsque r est fixé et que n tend vers $+\infty$,

$$\lim_{n\to+\infty} \mathbb{E}[\hat{R}_{CV}] = \frac{1}{K}\sum_{k=1}^{K} \lim_{r\to+\infty} \mathbb{E}[R_{\hat{f}_{\mathcal{D}_{n-r}}}] = \lim_{r\to+\infty} \mathbb{E}[R_{\hat{f}_{\mathcal{D}_{n-r}}}] = \inf_f R_f.$$

5. Notons $Z_k = \frac{1}{|I_k|}\sum_{i\in I_k} \ell(Y_i, \hat{f}_{-I_k}(X_i))$. Comm les (X_i, Y_i) sont indépendants et de même loi,

$$\mathrm{Var}(Z_k) = \frac{1}{|I_k|}\mathrm{Var}\left(\ell(Y, \hat{f}_{-I_k}(X))\right) = \frac{1}{r}\mathrm{Var}\left(\ell(Y, \hat{f}_{n-r}(X))\right).$$

En revanche, on a à calculer

$$\mathrm{Var}(\frac{1}{K}\sum_{k=1}^{K} Z_k) = \frac{1}{K^2}\left(\sum_{k=1}^{K}\mathrm{Var}(Z_k) + 2\sum_{i<j}\mathrm{Cov}(Z_i, Z_j)\right).$$

La difficulté ici est le calcul des termes de covariance. Si la validation croisée a pour effet d'augmenter la moyennisation, elle a aussi pour conséquence d'induire de la covariance. Intuitivement, on peut penser que si r est petit devant n, alors ces termes de covariance sont petits devant $\mathrm{Var}(Z_k)$ mais la quantification de son impact est un problème difficile.

Exercice 6 (Données manquantes/Données "mal balancées"). 1. On dispose de N individus, p variables et $m < N$ lignes où il manque une variable.

(a) Si N est grand devant p et m, alors on peut éliminer les m lignes où il manque une variable (Omission)

(b) Dans le cas contraire, on peut être tenté de compléter le tableau par une valeur estimée (Imputation). Le plus simple est de remplacer par la médiane (dans le cas quantitatif) ou par la classe la plus représentée dans le cas où les variables sont catégorielles. Néanmoins, on sent bien que cette méthode peut poser problème. On peut aussi envisager de remplacer la valeur/modalité manquante par la plus probable localement, *i.e.* en calculant les k-plus proches voisins (sur les autres variables) et en remplaçant par la moyenne des k-plus proches voisins ou la classe la plus probable sur les k-plus proches voisins (k pouvant être "appris"). A nouveau, ce type de choix a pour impact de biaiser le problème d'apprentissage puisqu'on commence déjà à apprendre pour remplir le tableau. Dans ce second cas, si la variable manquante a sur les k voisins les plus proches une variance assez faible, on peut considérer que ce choix est raisonnable (A l'inverse,...). Pour compléter cette discussion, on peut aussi noter que l'on a aussi possibilité d'éliminer la colonne s'il y a beaucoup de valeurs manquantes sur cette même colonne (variable) et/ou si elle semble de seconde importance (par exemple, via une ACP, on peut constater que la variable n'amène pas beaucoup d'information). Enfin, si l'on voit que la variable joue un rôle fondamental et qu'il n'est pas raisonnable de préestimer sa valeur, alors, on peut considérer judicieux d'éliminer l'individu.

Remarque : Il s'agit seulement ici de quelques réflexions sur les données manquantes. Il existe une théorie précise sur le sujet que nous ne développerons pas ici (mais qu'il est bien sûr utile de consulter si les données manquantes sont un point important de votre analyse de données). Par ailleurs, ces réflexions sont plutôt à prendre en compte lorsque les données sont "aléatoirement manquantes", *i.e.* que l'absence de données manquantes est due par exemple à un oubli de mesure à un problème de report,.... Dans de nombreux cas, l'absence d'une donnée peut ne pas être aléatoire, comme par exemple, une réponse à laquelle un individu a choisi volontairement de ne pas répondre lors d'une enquête. Dans ce cas, le traitement de cette absence de donnée doit bien entendu être adapté.

2. (a) Considérons l'exemple suivant : on a 100 individus dont 70 de type A et 30 de type B. Si l'on utilise l'erreur de classification comme fonction de perte et que l'on estime le risque associé de manière usuelle (soit erreur test soit validation croisée), alors il se peut que l'algorithme sélectionné prédise tout le temps A, ce qui génère une erreur de classification de 0.3. En effet, dans des cas "fortement aléatoires", ce prédicteur peut être le meilleur choix mais évidemment, il n'est que la conséquence du déséquilibre du jeu de données.

   (b) Pour éviter ce problème, on peut commencer par regarder le *rappel* : la probabilité d'être classé B (resp. classé A) sachant qu'on est de type B (resp. de type A). Pour ce qui est de la classe B, le rappel est nul dans l'exemple ci-dessus alors qu'il est égal à 1 pour la classe A.

   Pour un problème de détection de fraude ou de détection de cancer par exemple, le point important est d'avoir un rappel élevé sur la classe problématique, souvent sous-représentée. Elle correspond donc souvent à la classe B. Ci-dessus, on a mis en avant une manière de détecter un problème lié au déséquilibre de classes. La question suivante est : quelles modifications doit-on apporter à la procédure d'apprentissage pour éviter ce problème ? Il y a plusieurs réponses :
   — On peut tenter de rééquilibrer le jeu de données en réduisant la partie sur-représentée ou en injectant artificiellement des individus de type B . Attention à cette deuxième méthode, si on injecte des individus en retirant dans la classe sous-représentant ou même en "interpolant" (ex : SMOTE), on injecte de la dépendance dans le système. Cette méthode est donc à utiliser avec parcimonie. Néanmoins, comme dans tout problème d'apprentissage, il n'est pas interdit d'essayer : tant que la procédure d'apprentissage est réalisée avec un tirage au sort des échantillons d'apprentissage/validation/test et que les erreurs test et rappels test sont meilleures que dans le cas initial, on peut conserver cette approche.
   — On peut modifier la fonction de coût en attribuant plus de poids aux erreurs de "première espèce", *i.e.* celles où l'on prévoit A alors qu'on est B. Par exemple, si

   $$\ell(y, y') = (1-\alpha)1_{\{y=A, y'=B\}} + \alpha 1_{\{y=B, y'=A\}},$$

   alors le cas $\alpha = 1/2$ correspond au cas classique. Si $\alpha = 1$, on ne compte que les erreurs de première espèce et dans ce cas, le meilleur choix consiste à toujours prévoir B ce qui n'est pas satisfaisant non plus ! Par contre, on peut augmenter progressivement la valeur de α et voir l'effet sur le rappel (et sur la précision). C'est assez proche du principe de la courbe ROC.

Exercice 7 (1-ppv)**.** 1. L'échantillon d'apprentissage est $(X_1 = 0.8, Y_1 = 1)$, $(X_2 = 0.4, Y_2 = 0)$, $(X_3 = 0.7, Y_3 = 1)$.

   (a) L'algorithme des 3-p.p.v. donne : $\hat{f}(x) = 1$ pour tout $x \in [0, 1]$.

   (b) L'algorithme du p.p.v. donne : $\hat{f}(x) = 1$ si $x > 0.55$, 0 sinon.

2. (a)

$$\mathbb{P}(Y = 1 | X = x) = \begin{cases} 1 & \text{si } x > 0.5 \\ 0 & \text{si } x \leq 0.5. \end{cases}$$

   (b) Le prédicteur de Bayes dans ce cas est

$$f^\star(x) = \begin{cases} 1 & \text{si } x > 0.5 \\ 0 & \text{si } x \leq 0.5. \end{cases}$$

   Son risque est nul.

   (c)

$$E = E_0 \cup E_1$$

   où $E_j = \cap_{i=1}^n \{Y_i = j\}$, $j = 0, 1$. Notons $\hat{f}$ le 1-ppv. Si $\omega \in E_j$, alors $\hat{f}(x) = j$ pour tout $x \in [0, 1]$. L'indépendance entre les X_i implique

$$\mathbb{P}(E_0) = \mathbb{P}(E_1) = 2^{-n} \Longrightarrow \mathbb{P}(E) = 2^{1-n}.$$

(d) Notons $(X^{(i)})_{i=1}^n$ l'échantillon $(X_i)_{i=1}^n$ réordonné dans l'ordre croissant. Notons

$$i^\star := \max\{i \geq 1, X^{(i)} \leq \frac{1}{2}\}.$$

$i^\star$ est bien défini sur E^c. L'algorithme du 1-ppv donne :

$$\hat{f}(x) = \begin{cases} 0 & \text{si } x \leq \frac{X^{(i^\star)}+X^{(i^\star+1)}}{2} \\ 1 & \text{si } x > \frac{X^{(i^\star)}+X^{(i^\star+1)}}{2}. \end{cases}$$

Pour un point $x \in [0,1]$, l'algorithme se trompe dans les deux situations suivantes :
- $\frac{X^{(i^\star)}+X^{(i^\star+1)}}{2} < 1/2$ et $x \in [\frac{X^{(i^\star)}+X^{(i^\star+1)}}{2}, 1/2]$,
- $\frac{X^{(i^\star)}+X^{(i^\star+1)}}{2} > 1/2$ et $x \in [1/2, \frac{X^{(i^\star)}+X^{(i^\star+1)}}{2}]$.

En utilisant que X suit la loi uniforme, on en déduit que le risque (aléatoire) associé satisfait sur E^c

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

(e) On remarque que

$$\left| \frac{X^{(i^\star)} + X^{(i^\star+1)}}{2} - \frac{1}{2} \right| \leq \frac{1}{2} \left(\left| X^{(i^\star)} - \frac{1}{2} \right| + \left| X^{(i^\star+1)} - \frac{1}{2} \right| \right) \leq \frac{1}{2} \left(\min_{i=1}^n X_i^+ + \min_{i=1}^n X_i^- \right).$$

Le résultat suit en décomposant sur E et E^c et en remarquant que sur E, $R_{\hat{f}} \leq 1$.

(f) En utilisant la symétrie de la loi uniforme (par rapport à 1/2), on remarque facilement que X_i^+ et X_i^- ont même loi. Plus précisément, pour toute fonction f mesurable bornée,

$$\mathbb{E}[f(X_i^+)] = \frac{1}{2}f(0) + \int_{\frac{1}{2}}^1 f(u - \frac{1}{2})du = \frac{1}{2}f(0) + \int_0^{\frac{1}{2}} f(\frac{1}{2} - v)dv = \mathbb{E}[f(X_i^-)].$$

Ainsi, comme les X_i^+ et les X_i^- sont deux suites de variables indépendantes, il vient que $\min_{i=1}^n X_i^+$ et $\min_{i=1}^n X_i^-$ ont même loi et donc même espérance.

(g) Pour tout $r > 0$,

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

Comme $\mathbb{P}(\min_{i=1}^n X_i^+ > r) = 0$ pour tout $r \geq 1/2$, on déduit du lemme de Wald que

$$\mathbb{E}[\min_{i=1}^n X_i^+] = \int_0^{\frac{1}{2}} \left(\frac{1}{2} - r \right)^n dr = \frac{2^{-n-1}}{n+1}.$$

(h) De ce qui précède, on déduit que

$$\mathbb{E}[R_{\hat{f}}] \leq \frac{2^{-n}}{n+1} + 2^{1-n} \xrightarrow{n \to +\infty} 0.$$

Cela montre bien la consistance du 1-ppv.

3.

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

(a) Le prédicteur de Bayes dans ce cas est la fonction $f^\star$ définie par $f^\star(x) = 1$ pour tout $x \in [0,1]$. Son risque est égal à :

$$R_{f^\star} = \mathbb{P}(Y \neq 1) = \int_0^1 \mathbb{P}(Y = 0 | \mathbf{X} = x) \mathbb{P}_X(dx) = \frac{1}{3}.$$

(b) Risque de l'algorithme du plus proche voisin pour ce modèle ?

i. Notons $\hat{f}$ l'algorithme du 1-ppv et pour tout i, $\mathcal{C}_i$, l'ensemble des points de $[0,1]$ dont X_i est le plus proche (cellule de Voronoï associée à X_i). Remarquons que cet ensemble est bien défini pour presque tout $x \in [0,1]$ car X est à densité. On a :

$$\hat{f}(x) = Y_i \quad \forall x \in \mathcal{C}_i.$$

ii. La loi conditionnelle de Y sachant X ne dépend pas de X. On en déduit facilement que X et Y sont indépendants. En effet, pour tous $a < b \in [0,1]$,

$$\mathbb{P}(X \in [a,b], Y = 1) = \int_a^b \mathbb{P}(Y = 1 | X = x) \mathbb{P}_X(dx) = \frac{2}{3}(b-a) = \mathbb{P}(Y=1)\mathbb{P}(X \in [a,b]).$$

iii. Le risque $R_{\hat{f}}$ satisfait

$$R_{\hat{f}} = \mathbb{P}(Y \neq \hat{f}(X) | \mathcal{D}_n) = \sum_{i=1}^n \mathbb{P}(X \in \mathcal{C}_i, Y_i \neq Y | \mathcal{D}_n).$$

Ainsi,

$$\mathbb{E}[R_{\hat{f}}] = \sum_{i=1}^n \mathbb{P}(X_i \text{ ppv de } X, Y_i \neq Y),$$

où (X,Y) est indépendant de la suite $(X_i, Y_i)_{i=1}^n$. Au vu de la question précédente, on a en fait 4 variables indépendantes. En particulier,

$$\mathbb{E}[R_{\hat{f}}] = \sum_{i=1}^n \mathbb{P}(X_i \text{ ppv de } X)\mathbb{P}(Y_i \neq Y).$$

Comme $\mathbb{P}(Y_i \neq Y) = \mathbb{P}(Y_1 \neq Y)$, on a en fait

$$\mathbb{E}[R_{\hat{f}}] = \mathbb{P}(Y_1 \neq Y) \sum_{i=1}^n \mathbb{P}(X_i \text{ ppv de } X) = \mathbb{P}(Y_1 \neq Y)$$

car la somme est égale à 1.

iv.
$$\mathbb{E}[R_{\hat{f}}] = \mathbb{P}(Y_1 \neq Y) = \mathbb{P}(Y_1 = 1, Y = 0) + \mathbb{P}(Y_1 = 0, Y = 1) = 2\mathbb{P}(Y_1 = 1)\mathbb{P}(Y = 0)$$

car Y_1 et Y sont indépendants.

v. On en déduit que

$$\mathbb{E}[R_{\hat{f}}] = 2 \times \frac{2}{3} \times \frac{1}{3} = \frac{4}{9}.$$

Cette valeur est indépendante de n est strictement supérieure à $\frac{1}{3}$. Le 1-ppv n'est pas consistant dans ce cas.

Exercice 8 (Théorème de Stone)**.** On illustre ici le théorème de Stone sur une méthode par partitionnement. Avec les notations introduites en TD,

$$\hat{f}_{n,\varepsilon}(x) = \frac{1}{N(n,\varepsilon)} \sum_{i=1}^n Y_i 1_{X_i \in A_k^\varepsilon} \quad \forall x \in A_k^\varepsilon.$$

Si $\mathbb{E}[|Y|] < +\infty$ et $\mathbb{P}(X \in A_k^\varepsilon) > 0$, on déduit de la loi forte des grands nombres (en multipliant et en divisant par n) que

$$\hat{f}_{n,\varepsilon}(x) \xrightarrow{n \to +\infty} \mathbb{E}[Y | X \in A_k^\varepsilon].$$

Supposons maintenant que pour tout $x \in \mathcal{X}$, $\mathcal{L}(Y|X=x)$ a une densité $p_{Y/X=x}$ par rapport à la mesure de Lebesgue et que $(x,y) \mapsto p_{Y/X=x}(y)$ est continue sur $\mathcal{X} \times \mathbb{R}$. Dans ce cas,

$$\mathbb{E}[Y|X \in A_k^\varepsilon] = \frac{1}{\mathbb{P}(X \in A_k^\varepsilon)} \int_{A_k^\varepsilon} \mathbb{E}[Y|X=x]\mathbb{P}_X(dx) = \frac{1}{\mathbb{P}(X \in A_k^\varepsilon)} \int_{A_k^\varepsilon} \int_{\mathbb{R}} y p_{Y/X=x}(y) dy \mathbb{P}_X(dx).$$

Pour tout $x_0 \in A_k^\varepsilon$, lorsque ε est petit,

$$\int_{A_k^\varepsilon} p_{Y/X=x}(y)\mathbb{P}_X(dx) \approx p_{Y|X=x_0}(y)\mathbb{P}(X \in A_k^\varepsilon)$$

de sorte que par Fubini,

$$\begin{aligned}
\int_{A_k^\varepsilon} \int_{\mathbb{R}} y p_{Y/X=x}(y) dy \mathbb{P}_X(dx) &= \int_{\mathbb{R}} y \left(\int_{A_k^\varepsilon} p_{Y/X=x}(y)\mathbb{P}_X(dx) \right) dy \\
&\approx \mathbb{P}(X \in A_k^\varepsilon) \int_{\mathbb{R}} y p_{Y/X=x_0}(y) dy \\
&= \mathbb{P}(X \in A_k^\varepsilon)\mathbb{E}[Y|X=x_0].
\end{aligned}$$

On en déduit (avec ce raisonnement pas tout à fait rigoureux) que pour ε petit et pour tout $x_0 \in A_k^\varepsilon$,

$$\mathbb{E}[Y|X \in A_k^\varepsilon] \approx \mathbb{E}[Y|X=x_0].$$

Il vient

$$\lim_{\varepsilon \to 0} \lim_{n \to +\infty} \hat{f}_{n,\varepsilon}(x_0) = f^*(x_0) = \mathbb{E}[Y|X=x_0].$$

Sous des hypothèses assez générales (moments suffisants pour Y), il vient (par des arguments de type "convergence dominée"),

$$\lim_{\varepsilon \to 0} \lim_{n \to +\infty} R_{\hat{f}_{n,\varepsilon}} = R_{f^*} \quad p.s.$$

d'où une "potentielle" consistance forte en considérant une suite $(\hat{f}_{n,\varepsilon_n})_{n \geq 1}$ avec ε_n tendant vers 0 "à la bonne vitesse" : cette condition est similaire à celles des k-ppv où l'on demande que k_n tende vers $+\infty$ et que k_n/n tende vers 0 mais nous ne souhaitons pas la détailler ici.

Exercice 9 (Classification binaire (suite))**.** 1. Soit f une règle de prévision.

$$R_f = \mathbb{E}[\log(1+e^{-Yf(X)})] = \mathbb{E}[\log(1+e^{-f(X)})1_{Y=1}] + \mathbb{E}[\log(1+e^{f(X)})1_{Y=-1}]$$

. On a

$$\mathbb{E}[\log(1+e^{-f(X)})1_{Y=1}|X=x] = \log(1+e^{-f(x)})\mathbb{P}(Y=1|X=x) = \log(1+e^{-f(x)})p(x).$$

Ainsi,

$$R_f = \mathbb{E}[\log(1+e^{-f(X)})p(X)] + \mathbb{E}[\log(1+e^{f(X)})(1-p(X))].$$

En notant ϕ la fonction définie par

$$\phi(p) = \min\{\log(1+e^{-1})p + \log(1+e^1)(1-p), \log(1+e^1)p + \log(1+e^{-1})(1-p)\},$$

on en déduit immédiatement que pour toute règle de prévision satisfait

$$R_f \geq \mathbb{E}[\phi(p(X))].$$

Il nous faut donc maintenant prouver que la quantité à droite de l'inégalité peut être atteinte. Pour ce faire, résolvons l'inéquation

$$\log(1+e^{-1})p + \log(1+e)(1-p) < \log(1+e)p + \log(1+e^{-1})(1-p).$$

Celle-ci est équivalente à

$$2\log\left(\frac{1+e}{1+e^{-1}}\right)p > \log\left(\frac{1+e}{1+e^{-1}}\right),$$

i.e. à $p > 1/2$. Ainsi, si l'on considère le prédicteur de Bayes, on remarque alors que par construction,

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

Ainsi,

$$R_{f^\star} = \inf_f \mathbb{E}[\ell(Y, f(X))].$$

N.B. Ici, il faut bien voir que même si la notation ne le fait pas apparaître, $R_{f^\star}$ dépend évidemment de ℓ. Cette quantité est donc différente de celle de l'erreur de classification standard.

En conclusion, cette fonction de perte a le même minimiseur qui est le prédicteur "naturellement optimal". On peut donc envisager dans la suite de sélectionner un algorithme à l'aide de cette fonction de perte.

2. Considérons maintenant le cas où l'on cherche à résoudre le problème d'optimisation pour les fonctions f de $\mathcal{C}$ vers $\mathbb{R}$. Si l'on reprend le calcul précédent, on constate que dans ce cas,

$$R_f \geq \mathbb{E}[F_{p(X)}(f(X))]$$

où

$$F_p(a) = \log(1+e^{-a})p + \log(1+e^a)(1-p).$$

On a

$$R_f \geq \mathbb{E}[\inf_{a\in\mathbb{R}} F_{p(X)}(a)].$$

Déterminons maintenant $f^\star$. Pour cela, calculons

$$F_p'(a) = \frac{-pe^{-a}}{1+e^{-a}} + \frac{(1-p)e^a}{1+e^a} = \frac{1}{1+e^a}\left(-p + e^a(1-p)\right),$$

Ainsi, lorsque $p \in]0,1[$:

$$F_p'(a) \geq 0 \iff -p + e^a(1-p) \geq 0 \iff a \geq \log\left(\frac{p}{1-p}\right).$$

F_p admet donc un minimum global en $\log\left(\frac{p}{1-p}\right)$. Lorsque $p = 0$, l'infimum de F_p est atteint en $-\infty$ et lorsque $p = 1$, celui-ci est atteint en $+\infty$. On pose donc

$$f^\star(x) = \begin{cases} \log\left(\frac{p(x)}{1-p(x)}\right) & \text{si } p(x) \in]0,1[\\ -\infty & \text{si } p(x) = 0 \\ +\infty & \text{si } p(x) = 1 \end{cases}$$

Par construction,

$$F_{p(X)}(f^\star(X)) = \inf_{a\in\mathbb{R}} F_{p(x)}(a),$$

de sorte que

$$R_{f^\star} = \inf R_f.$$

Remarque : On constate donc que si l'on est capable de déterminer $f^\star$, alors on est capable de trouver $x \mapsto p(x)$.

Exercice 10 (Risque asymétrique/Courbe ROC)**.** 1. On considère la fonction de perte asymétrique :

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

Soit f une règle de prévision :

$$\mathbb{E}[\ell(Y, f(X))] = (1-s)\mathbb{E}[\eta(X)1_{f(X)=0}] + s\mathbb{E}[(1-\eta(X))1_{f(X)=1}].$$

On a donc

$$\mathbb{E}[\ell(Y, f(X))] \geq \mathbb{E}[\min\{(1-s)\eta(X), s(1-\eta(X)\}].$$

Par ailleurs,

$$R_{f_s^\star} = \mathbb{E}[\ell(Y, f_s^\star(X))] = (1-s)\mathbb{E}[\eta(X)1_{\{\eta(X)\leq s\}}] + s\mathbb{E}[(1-\eta(X))1_{\{\eta(X)>s\}}].$$

Or, on remarque pour $p \in [0,1]$, $(1-s)p \leq s(1-p)$ si et seulement si $p \leq s$. Cela implique donc que

$$R_{f_s^\star} = \mathbb{E}[\min\{(1-s)\eta(X), s(1-\eta(X)\}].$$

Le cas $s = 1/2$ correspond bien au cas usuel.

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) Représentation graphique : 2 gaussiennes avec des moyennes différentes...

(b) Il nous faut ici calculer $\eta(x)$. On doit inverser par la formule de Bayes. Formellement (car $\{X=x\}$ est de mesure nulle),

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

De manière plus rigoureuse,

$$\eta(x) = \mathbb{P}(Y=1|X=x) = \frac{pf_{m_1,\sigma_1^2}(x)}{pf_{m_1,\sigma_1^2}(x) + (1-p)f_{m_0,\sigma_0^2}(x)}.$$

La règle de Bayes s'en déduit alors en attribuant 1 si cette quantité est plus grande que s et 0 sinon.

(c) Si $\sigma_0 = \sigma_1$, alors

$$\eta(x) \geq s \Longleftrightarrow 1 \geq s\left(1 + \frac{1-p}{p}e^{\frac{(m_0-m_1)}{\sigma_0^2}x + \frac{m_1^2-m_0^2}{2\sigma_0^2}}\right).$$

On résout cette inéquation pour obtenir

$$\eta(x) \geq s \Longleftrightarrow x > \alpha_s := \frac{m_1+m_0}{2} + \frac{\sigma_0^2}{m_1-m_0}\log\left(\frac{s(1-p)}{p(1-s)}\right).$$

La règle de décision recherchée est alors :

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

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

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

Si l'on dit qu'un malade est positif et qu'un individu sain est négatif, alors

$$\text{Se}(\alpha) = \mathbb{P}(\text{classé positif} \mid \text{positif})$$

et

$$\text{Sp}(\alpha) = \mathbb{P}(\text{classé négatif} \mid \text{négatif}).$$

Le risque de première espèce correspond à $1 - \text{Se}(\alpha)$ tandis que le risque de seconde espèce correspond à $1 - \text{Sp}(\alpha)$.

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

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

Pour tracer cette courbe, remarquons que si F désigne la fonction de répartition de $\mathcal{N}(m_0, 1)$ et G celle de $\mathcal{N}(m_1, 1)$, alors

$$1 - \mathrm{Sp}(\alpha) = 1 - F(\alpha)$$

tandis que

$$\mathrm{Se}(\alpha) = 1 - G(\alpha).$$

Ainsi, si l'on pose $u = 1 - F(\alpha)$, alors $\mathrm{Se}(\alpha) = 1 - G(F^{-1}(1-u))$. On trace donc la courbe $u \mapsto G(F^{-1}(1-u))$. Remarquons que par les propriétés de translation de la loi normale, $F(t) = \phi(t - m_0)$ (de sorte que $F^{-1}(v) = m_0 + \phi^{-1}(v)$) et $G(t) = \phi(t - m_1)$. Ainsi,

$$1 - G(F^{-1}(1-u)) = 1 - \phi(m_0 - m_1 + \phi^{-1}(1-u)).$$

On obtient

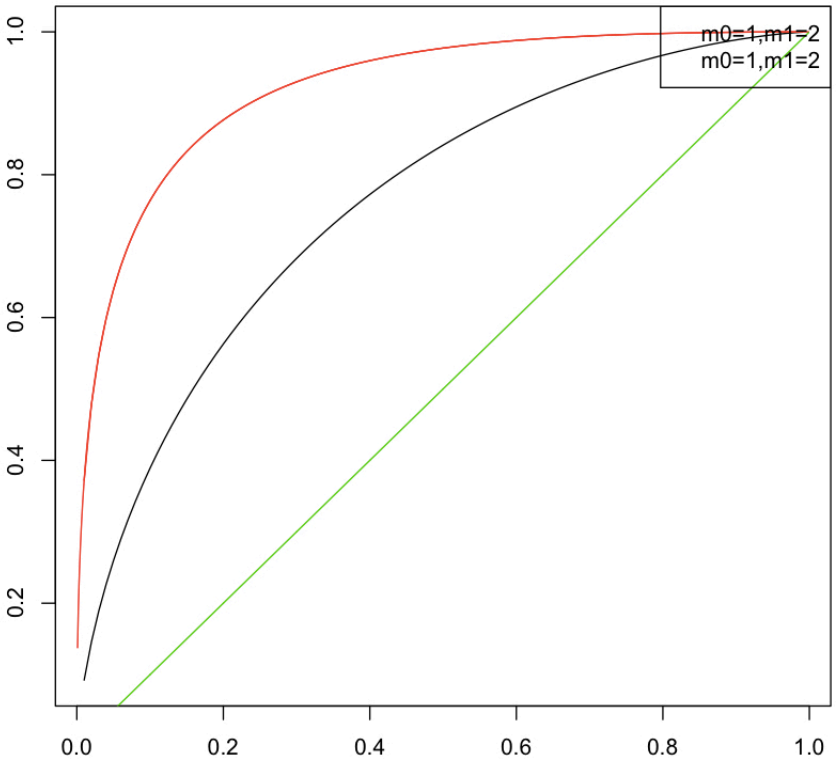

Figure : Courbes ROC

L'aire AUC est naturellement plus grande lorsque les moyennes sont plus éloignées.

(f) Si on construit maintenant un vrai algorithme d'apprentissage, alors on peut remplacer α_s par $\hat{\alpha}_s$ défini de la manière suivante. On suppose pour simplifier que p est connu et que les n_1 premières observations sont des malades et les n_2 suivantes des individus sains. Dans ce cas, il est naturel d'estimer m_0 par $\hat{m}_0 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ et m_1 par $\hat{m}_1 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} X_i$. On définit alors

$$\hat{\alpha}_s := \frac{\hat{m}_1 + \hat{m}_0}{2} + \log\left(\frac{s(1-p)}{p(1-s)}\right).$$

Comme l'estimateur de la moyenne est fortement consistant, on en déduit que $\hat{\alpha}_s \to \alpha_s$ lorsque n_1 et n_2 tendent vers $+\infty$. Par convergence dominée, on en déduit que le prédicteur associé est consistant (au sens faible mais aussi au sens fort). Dans le cas général où les variances connues, ce type de procédé est généralisable (laissé en exercice).

(g) A faire en TP.

3. Commentaire sur la courbe ROC en pratique lorsque le modèle n'est pas connu. Prenons par exemple le cas des k-plus proches voisins. Dans ce cas, pour chaque valeur de s, on peut fabriquer un algorithme de prévision $\hat{f}_s$. Plus précisément, si on note $\hat{\eta}(x)$ l'estimation de $\mathbb{P}(Y = 1|X = x)$ obtenue par vote majoritaire (par exemple en 3-ppv avec 2 voisins de classe 1 et 1 de classe 0, $\hat{\eta}(x) = 2/3$), alors on peut poser
$$\hat{f}_s(x) = 1_{\hat{\eta}(x)>s}.$$
Sur un échantillon test/validation, ce dernier génère un quadruplet $(TP(s), FP(s), FN(s), TN(s))$ qui à son tour génère :
$$\hat{\text{Se}}(s) = \frac{TP(s)}{TP(s) + FN(s)}$$
et
$$\hat{\text{Sp}}(s) = \frac{TN(s)}{TN(s) + FP(s)}.$$
Dans ce cas, construire la courbe ROC de l'échantillon consiste à faire varier s et à tracer le graphe associé.

Compléments de vocabulaire :

1. Rappel : Probabilité d'être classé positif sachant qu'on est positif (en pratique $\frac{TP}{TP+FN}$) (même chose que la sensibilité). Plus le rappel est élevé, plus le prédicteur classe des individus positifs.
2. Précision : Probabilité d'être positif sachant qu'on est classé positif (en pratique $\frac{TP}{TP+FP}$) (alors que la sensibilité est le même objet appliqué aux individus négatifs). Plus la précision est élevée, moins le modèle se trompe sur les positifs.

   Précision et Rappel peuvent bien sûr être envisagés sur chaque classe (tout dépend du contexte).
3. F_1-score : c'est la moyenne harmonique du rappel et de la précision :
$$\frac{2 \times \text{Rappel} \times \text{Précision}}{\text{Rappel} + \text{Précision}} = \frac{2}{\frac{1}{\text{Rappel}} + \frac{1}{\text{Précision}}}.$$
C'est l'inverse de la moyenne arithmétique des inverses. Il s'agit donc aussi d'un nombre entre 0 et 1 qui vaut 1 lorsque Rappel et Précision sont maximaux et 0 lorsque Rappel ou Précision sont nulles. Cela donne un compromis entre rappel et précision. Notons qu'à nouveau elle peut être utilisée pour chaque classe. Dans le cadre multiclasse, elle l'est d'ailleurs par défaut dans `sklearn` (voir `sklearn.metrics.f1_score`). Cette "métrique" est souvent utilisée en pratique.

   On trouvera bien sûr dans la littérature d'autres scores de performance tels que le score de Fowlkes-Mallows qui est construit comme la moyenne géométrique de la précision et du rappel (donc $\sqrt{\text{Précision} \times \text{Rappel}}$).