Contrôle Terminal

Le contrôle est composé de deux exercices indépendants et est noté sur 24 points (13 points pour le premier exercice et 11 points pour le second). Si aucun ajustement des notes n'est nécessaire, la note finale sera le minimum entre votre nombre de points et 20. Si un ajustement est nécessaire, votre note finale ne sera pas inférieure à celle calculée avec la règle précédente.

Ici log est le logarithme népérien.

Exercice 1

Soient $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ des variables aléatoires indépendantes et identiquement distribuées.

Le but de l'apprentissage supervisé est d'apprendre une fonction $\hat{f}$ à partir de $(X_1, Y_1), \dots, (X_n, Y_n)$ telle que $\hat{f}(X) \approx Y$ (en espérance ou avec forte probabilité) où le sens de $\approx$ dépend de la nature du problème (i.e. de la fonction de perte choisie).

Comme nous l'avons vu en classe, il est possible de résoudre ce problème de manière directe en cherchant directement à travailler sur la loi de Y sachant X, par exemple en considérant l'estimateur minimisant l'erreur empirique.

Let but de cet exercice est d'étudier une nouvelle méthode générique pour résoudre ce problème : les méthodes génératives.

L'idée générale est d'estimer la loi de Y puis la loi de X sachant Y avant d'en déduire la loi de Y sachant X via la formule de Bayes.

Pour le reste de l'exercice, nous fixons $\mathcal{X} = \mathbb{R}^d$ et $\mathcal{Y} = \{-1, 1\}$.

1) (1 point) Quel est le nom de ce type de problème où nous essayons de "prédire" une variable catégorielle ?

2) (1 point) Écrire le problème de minimisation du risque empirique pour ce problème où la fonction de perte choisie est la fonction de perte du $0-1$: $l(y, z) = \mathbb{1}_{y \neq z}$ et où la fonction recherchée appartient à un ensemble $\mathcal{F}$.

Pour le reste du problème, nous supposons que X est une variable à densité par rapport à la mesure de Lebesgue. Nous notons f_X cette densité. Pour simplifier l'analyse, nous supposons que $f_X(x) > 0, \forall x \in \mathcal{X}$. De plus, comme Y vit dans un ensemble fini $\{-1, 1\}$, sa loi est uniquement déterminée par $\pi_1 = \mathbb{P}(Y = 1)$. Nous notons également $\pi_{-1} = \mathbb{P}(Y = -1)$, et nous avons donc $\pi_{-1} = 1 - \pi_1$. Encore une fois, pour simplifier l'analyse, nous supposons que $0 < \pi_1 < 1$.

3) (1 point) En développant de deux manières différentes la loi du couple $\mathbb{P}(X = x, Y = c)$ avec $x \in \mathcal{X}$ et $c \in \{-1, 1\}$, prouver la formule de Bayes :

$$\forall x \in \mathcal{X}, \forall c \in \{-1, 1\}, \quad \mathbb{P}(Y = c | X = x) = \frac{\pi_c f_{X|Y=c}(x)}{\pi_1 f_{X|Y=1}(x) + \pi_{-1} f_{X|Y=-1}(x)}$$

où $f_{X|Y=c}$ fait référence à la densité (par rapport à la mesure de Lebesgue) de X conditionnellement à l'événement $(Y = c)$.

Pour pouvoir appliquer cette formule, nous allons devoir estimer quatre quantités : deux scalaires, π_1 et π_{-1} et deux densités de probabilités, $f_{X|Y=1}$ et $f_{X|Y=-1}$.

Les trois prochaines questions portent sur l'estimation de π_1 et de π_{-1}. Comme $\pi_1 = 1 - \pi_{-1}$, nous nous intéresserons exclusivement à l'estimation de π_1.

4) (1 point) Montrer que l'estimateur du maximum de vraisemblance de π_1 en utilisant l'échantillon $(Y_1, \dots, Y_n)$ donne

$$\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i = 1} \, .$$

On pourra remarquer que $\mathbb{P}(Y = c) = \pi_1^{\mathbb{1}_{c=1}} (1 - \pi_1)^{1 - \mathbb{1}_{c=1}}$ et penser à passer par le log.

Rappel : Pour un modèle probabiliste $(\mathbb{P}_\theta, \theta \in \Theta)$ et avec un échantillon $(Y_1, \dots, Y_n)$, le maximum de vraisemblance est défini par

$$\hat{\theta} \in \text{argmax}_{\theta \in \Theta} \mathbb{P}_\theta(Y_1, \dots, Y_n) \, .$$

On prendra donc le temps d'identifier dans notre cas ce qui paramétrise la loi de Y.

5) (3 points) Étudier le biais et la variance de cet estimateur. Montrer que cet estimateur est sans biais et calculer explicitement sa variance.

6) (1 point) En utilisant l'inégalité de Hoeffding, montrer que cet estimateur est fortement concentré autour de sa moyenne avec haute probabilité.

Rappel de l'inégalité de Hoeffding : Soient $Z_1, \dots, Z_n$ des variables indépendantes bornées telles que $a \leq Z_i \leq b$ pour tout i. Alors, pour tout $t > 0$, nous avons :

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i (b-a)^2}\right).$$

Pour le reste du problème, nous définissons $\hat{\pi}_{-1} = 1 - \hat{\pi}_1$.

Les trois questions suivantes portent sur l'estimation des densités conditionnelles $f_{X|Y=1}$ et $f_{X|Y=-1}$ via des estimateurs à noyaux.

Attention : Il n'y a pas nécessairement de lien avec les méthodes à noyaux vues en classe. Jusqu'à la fin de cet exercice, le terme de noyau fait référence aux notions définies ci-dessous. En dehors de cet exercice, le sens premier d'un noyau redevient celui vu en classe.

7) (1 point) Un estimateur à noyaux d'une densité de probabilité f dans $\mathbb{R}^d$ est défini par :

$$\hat{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - Z_i}{h}\right),$$

où $Z_1, \dots, Z_n$ sont sensées être des variables iid dont la loi a pour densité f, $K : \mathbb{R}^d \to [0, +\infty)$ est une fonction dont l'intégrale sur $\mathbb{R}^d$ vaut 1 appelée noyau et $h > 0$ est un paramètre appelé bande passante. Montrer que $\hat{f}_h$ est bien une densité de probabilité en vérifiant qu'elle est positive et qu'elle s'intègre à 1.

Pour un noyau K fixé et une bande passante $h > 0$, ce principe nous conduit à considérer l'estimateur suivant pour $f_{X|Y=c}$:

$$\forall c \in \{-1, 1\}, \forall x \in \mathbb{R}^d, \quad \hat{f}_{X|Y=c} = \frac{1}{h^d \sum_{i=1}^{n} \mathbb{1}_{Y_i=c}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \mathbb{1}_{Y_i=c}\,.$$

8) (2 point) Voici trois exemples classiques de noyaux en dimension d :

— **Noyau uniforme** : $K(u) = \frac{1}{V_d} \mathbb{1}_{\|u\| \leq 1}$ où V_d est le volume de la boule unité en dimension d.

— **Noyau gaussien** : $K(u) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|u\|^2}{2}}$.

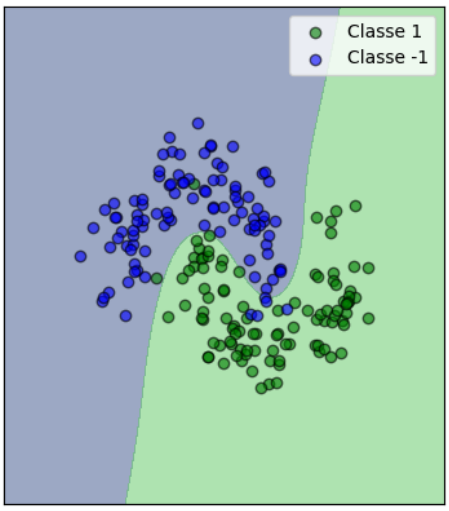
— **Noyau d'Épanechnikov** : $K(u) = \frac{d+2}{2V_d}(1 - \|u\|^2)\mathbb{1}_{\|u\| \leq 1}$.

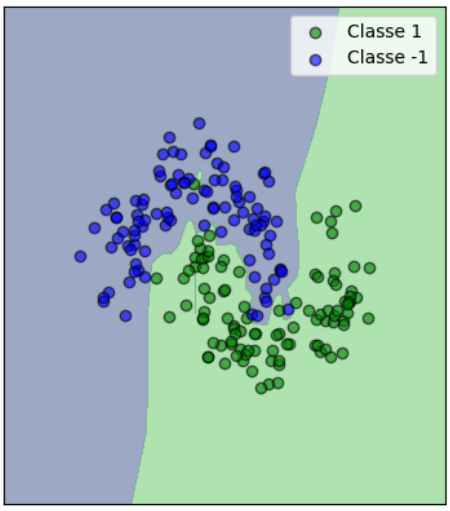
Dessiner grossièrement ces noyaux lorsque $d = 1$ et expliquer l'effet du choix du noyau et du paramètre h sur l'estimation. Que se passe-t-il si h est trop grand ou trop petit ? On pourra donner une justification en parlant du biais et de la variance de la procédure considérée.

9) (2 points) Nous fixons K le noyau Gaussien et nous définissons $\hat{f}(x) \in \operatorname{argmax}_{c \in \{-1,1\}} \hat{p}(c|x)$ où $\hat{p}(c|x)$ est un estimateur de la loi conditionnelle $\mathbb{P}(Y = c|X = x)$ défini par :

$$\hat{p}(c|x) = \frac{\hat{\pi_c} \hat{f}_{X|Y=c}(x)}{\hat{\pi_1} \hat{f}_{X|Y=1}(x) + \hat{\pi}_{-1} \hat{f}_{X|Y=-1}(x)}\,.$$

Il est important de voir que $\hat{f}$ dépend implicitement de h, la bande passante. Il s'agit du seul hyperparamètre laissé libre. Les trois figures suivantes représentent trois frontières de classification pour trois valeurs de h différentes. Pour chaque figure, indiquer en justifiant si nous sommes dans une situation de sur-apprentissage, de sous-apprentissage, ou dans un bon compromis.





Exercice 2

Soient $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$. Dans le cas où $\mathcal{X} = \mathbb{R}^d$, la régression logistique consiste à prédire $y = \text{sgn}(\hat{\theta}^T x)$ pour de nouveaux xs où sgn est la fonction de signe (prenant les valeurs $+ - 1$) et où $\hat{\theta}$ est solution de

$$\hat{\theta} \in \text{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i \theta^T x_i}) + \lambda \|\theta\|^2$$

où $\lambda \geq 0$ contrôle le niveau de régularisation. Le but de cet exercice est d'étudier la version "à noyaux" de cette méthode.

Étant donné $\mathcal{H}$ un espace de Hilbert et $\phi : \mathcal{X} \mapsto \mathcal{H}$, nous allons considérer le noyau $k : x_1, x_2 \mapsto \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}$.

1) (2 points) Prouver que ce noyau est bien symétrique positif, c'est-à-dire que k est symétrique et que pour tout $n \geq 1$, pour tous $x_1, \ldots, x_n \in \mathcal{X}$ et pour tous $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$.

2) (2 point) Justifier pourquoi nous pouvons nous restreindre à choisir un noyau symétrique positif sous cette forme.

La variante "à noyaux" de la régression logistique consiste à prédire $y = \text{sgn}(\langle \hat{\theta}, \phi(x) \rangle_{\mathcal{H}})$ où $\hat{\theta}$ est solution de

$$\hat{\theta} \in \text{argmin}_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i \langle \theta, \phi(x_i) \rangle_{\mathcal{H}}}) + \lambda \|\theta\|_{\mathcal{H}}^2 \ .$$

Ce nouveau problème est potentiellement en dimension infinie. Le reste de l'exercice montre comment se ramener à un problème de dimension finie et comment résoudre ce nouveau problème.

3) (2 points) Supposons $\lambda > 0$, montrer que ce problème admet une unique solution $\hat{\theta}$ et que de plus $\hat{\theta} \in \text{Vect}(\phi(x_1), \ldots \phi(x_n))$. On pourra au choix utiliser directement un résultat vu en classe ou penser à décomposer θ sur $\text{Vect}(\phi(x_1), \ldots \phi(x_n))$ et sur son orthogonal et exploiter le fait que $\lambda > 0$.

Il est donc possible d'effectuer le changement de variables $\theta = \alpha_1 \phi(x_1) + \cdots + \alpha_n \phi(x_n)$. La fonction de prédictions a alors la forme suivante :

$$f(x) = \text{sgn}\left(\alpha_1 \langle \phi(x_1), \phi(x) \rangle_{\mathcal{H}} + \cdots + \alpha_n \langle \phi(x_n), \phi(x) \rangle_{\mathcal{H}}\right) \ .$$

4) (2 points) Sur le jeu de données $\{(x_1, y_1) = ((0, 1), 1), (x_2, y_2) = ((-1, 1), -1), (x_3, y_3) = ((-1, -1), -1)\}$, donner l'expression de f et représenter à peu près la frontière de décision (i.e. l'ensemble des points classifiés en 1 et en -1) lorsque $\alpha_1 = 1$, $\alpha_2 = \alpha_3 = -1$ pour le noyau Gaussien $(x_1, x_2 \mapsto e^{-\|x_1 - x_2\|^2})$ et pour le noyau linéaire $(x_1, x_2 \mapsto x_1^T x_2)$.

5) (1 point) Montrer qu'en réécrivant le problème en fonction de $\alpha_1, \ldots, \alpha_n$, ce dernier se réécrit

$$\hat{\alpha} \in \text{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i (K\alpha)_i}) + \lambda \alpha^T K \alpha \ ,$$

où $K = (k(x_i, x_j))_{i,j} \in S_n^+(\mathbb{R})$, et α est utilisé pour noter le vecteur $(\alpha_1, \ldots, \alpha_n)$.

6) (2 points) Sans se soucier de garantir la convergence, quel algorithme d'optimisation peut être utilisé pour résoudre le nouveau problème en $\alpha_1, \ldots, \alpha_n$? Vous donnerez les formules des différentes quantités qui seront utiles pour le déroulement de l'algorithme.