TD 1
Bases de probabilités et statistique

Exercice 1

Un sondage a receuilli des informations sur le prix (au kilogramme) de certains fruits et legumes. Le tableau suivant donne les effectifs pour chaque paire type / classe de prix.

	pomme	poire	courgette	aubergine
2 €	12	24	54	23
3 €	45	26	72	16
4 €	34	63	34	33

Par exemple, dans le sondage il y a 12 pommes qui coutent 2 € (au kilogramme).

1. Calculer la probabilité que le fruit/légume du sondage soit une pomme sachant que son prix est 2 €.
2. Calculer la probabilité que le fruit/légume du sondage coute 3 € ou plus sachant que c'est une aubergine.
3. Calculer la probabilité que le fruit/légume du sondage soit un fruit sachant qu'il coute 3 € ou plus.
4. Calculer l'espérance conditionelle d'un fruit/légume du sondage sachant que c'est une poire (le prix moyen d'une poire dans le sondage).
5. Calculer l'espérance conditionelle d'un fruit/légume du sondage sachant que c'est un légume (le prix moyen d'un légume dans le sondage).
6. Calculer la variance conditionelle d'un fruit/légume du sondage sachant que c'est une courgette.

Exercice 2

On considère un couple de variables aléatoires (X, Y) sur $[0,1]^2$ dont la densité de probabilité est la fonction $f_{X,Y} : [0,1]^2 \to \mathbb{R}^+$ définie par, pour $x, y \in [0,1]^2$,

$$f_{X,Y}(x,y) = \frac{1}{c}\exp(-|x-y|),$$

où $c > 0$ est une constante (ne dépendant pas de x, y).

1. Calculer la constante c.
2. Pour $x \in [0,1]$, calculer la densité de X en x, c'est à dire la fonction $f_X : [0,1] \to \mathbb{R}^+$.
3. Calculer la fonction de densité conditionelle de Y sachant X, c'est à dire la fonction $f_{Y|X} : [0,1]^2 \to \mathbb{R}^+$.
4. Calculer l'espérance de Y sachant que X vaut 1.
5. Calculer la variance de Y sachant que X vaut 1.
6. Calculer la probabilité que $Y \geq 1/2$ sachant que X vaut 1.

Elément de cours : la loi de Poisson La loi de Poisson de paramètre $\lambda > 0$, notée $\mathcal{P}(\lambda)$, est une loi de probabilité sur $\mathbb{N} = \{0, 1, 2, \ldots\}$. Pour une variable aléatoire X telle que $X \sim \mathcal{P}(\lambda)$ on a

$$\mathbb{P}[X = k] = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}$$

et

$$\mathbb{E}[X] = \lambda.$$

Exercice 3

Soient $X_1, \ldots, X_n$ iid selon une loi de Poisson $\mathcal{P}(\lambda)$, où λ est inconnu.

1. Certains des exercices ont étés fournis par Adrien Mazoyer.

1. Pour $x_1, \dots, x_n \in \mathbb{N}^n$, calculer la probabilité que $(X_1, \dots, X_n) = (x_1, \dots, x_n)$ en fonction de λ.
2. Le maximum de vraissemblance consiste à maximiser la probabilité précédente en fonction de λ, lorsque $(x_1, \dots, x_n)$ est fixé et égal à $(X_1, \dots, X_n)$ (ce dernier vecteur est appelé vecteur des observations). Calculer l'estimateur $\widehat{\lambda}_{\mathrm{ML}}$ du maximum de vraisemblance de λ.

Exercice 4

Soient $X_1, \dots, X_n$ iid selon une loi Uniforme $\mathcal{U}(0, t)$, où $t \geq 0$ est inconnu.

1. Pour $(x_1, \dots, x_n) \in [0, +\infty[^n$, calculer la valeur de la fonction densité de probabilité de $(X_1, \dots, X_n)$ évaluée en $(x_1, \dots, x_n)$, en fonction de t.
2. Le maximum de vraissemblance consiste à maximiser la densité de probabilité précédente en fonction de t, lorsque $(x_1, \dots, x_n)$ est fixé et égal à $(X_1, \dots, X_n)$ (ce dernier vecteur est appelé vecteur des observations).Calculer l'estimateur $\widehat{t}_{\mathrm{ML}}$ du maximum de vraisemblance de t.
3. On considère maintenant que les variables sont iid selon une loi Uniforme $\mathcal{U}(t, t+1)$. Montrer alors que, quel que soit $t \in \mathbb{R}$, presque surement (avec probabilité 1), $\max(X_1, \dots, X_n) \leq \min(X_1, \dots, X_n) + 1$.
4. Calculer la valeur de la fonction densité de probabilité de $(X_1, \dots, X_n)$ évaluée en $(x_1, \dots, x_n)$, en fonction de t, lorsque $\max(x_1, \dots, x_n) \leq \min(x_1, \dots, x_n) + 1$.
5. Trouver l'ensemble des t qui maximisent la vraissemblance (il peut y en avoir plusieurs).

Exercice 5

Une maladie se propage dans une population, avec un taux de contamination de 1 personne pour 1000. Un nouveau test de dépistage de cette maladie est proposé avec les taux de détection suivants. Une personne malade obtiendra bien un test positif avec probabilité 99%. Une personne saine en revanche pourra obtenir un résultat positif avec probabilité 0.2%.

Calculez la probabilité qu'une personne soit effectivement malade si son test est positif.

TD 2
Régression logistique

Exercice 1

On cherche à construire un classifieur qui prend en entrée $x \in [-1,1]$ et qui le classifie en la classe 0 ou la classe 1. Ce classifieur va être paramétré par $\alpha, \beta \in \mathbb{R}$. La classe associée à l'entrée x sera 1 si $C_{\alpha,\beta}(x) \geq 1/2$ et 0 sinon, en définissant

$$C_{\alpha,\beta}(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

1. Tracer approximativement la courbe de la fonction $t \mapsto C_{0,1}(t) = e^t/(1+e^t)$ $(t \in \mathbb{R})$.
2. Montrer que $e^t/(1+e^t) \geq 1/2 \iff t \geq 0$.
3. Pour $\alpha = 0$ et $\beta = 1$, quels x sont classés en 0 et quels x sont classés en 1.
4. Même question pour $\alpha = 1$ et $\beta = -1$.
5. Pour un $x \neq 0$ et $\alpha \in \mathbb{R}$ fixés, lorsque β devient assez grand $(\beta \to +\infty)$, comment est classé x.
6. Même question lorsque $\beta \to -\infty$.

Exercice 2

Maintenant, on cherche à construire un classifieur qui prend en entrée $x = (x_1, x_2) \in [-1,1]^2$ et qui le classifie en la classe 0 ou la classe 1. Ce classifieur va être paramétré par $\alpha, \beta_1, \beta_2 \in \mathbb{R}$. La classe associée à l'entrée x sera 1 si $C_{\alpha,\beta_1,\beta_2}(x) \geq 1/2$ et 0 sinon, en définissant

$$C_{\alpha,\beta_1,\beta_2}(x) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2}}.$$

1. On prend $\alpha = 0, \beta_1 = 1, \beta_2 = 0$. Représenter graphiquement quels x sont classés 0 et quels x sont classés 1.
2. Même question avec $\alpha = 1, \beta_1 = 2, \beta_2 = -1$.
3. En général, pour α, β_1, β_2 quelconques tels que $(\beta_1, \beta_2) \neq (0,0)$, déterminer quels x sont classés 0 et quels x sont classés 1. La réponse à cette question permet de dire que l'on étudie un *classifieur linéaire*.

Exercice 3

On propose maintenant un modèle probabiliste qui correspond à ce classifieur. On fixe une dimension $d \in \mathbb{N}$, $d \geq 2$. On considère $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{R}^d$. On considère un couple de variables aléatoires $(X,Y) \in [-1,1]^d \times \{0,1\}$ où X suit la loi uniforme sur $[-1,1]$ et, pour tout $x \in [-1,1]^d$,

$$\mathbb{P}(Y=1|X=x) = 1 - \mathbb{P}(Y=0|X=x) = C_\beta(x) = \frac{e^{\beta^\top x}}{1+e^{\beta^\top x}}.$$

1. On considère un premier classifieur $f : [-1,1]^d \to \mathbb{R}$ qui attribue la classe 1 à x si $C_\beta(x) \geq 1/2$ et la classe 0 sinon. Prouver que le risque de classification $\mathbb{P}(f(X) \neq Y)$ s'écrit

$$\mathbb{P}(f(X) \neq Y) = \frac{1}{2^d} \int_{[-1,1]^d} \min\left(C_\beta(x), 1 - C_\beta(x)\right) dx.$$

   On pourra d'abord calculer $\mathbb{P}(f(X) \neq Y|X=x)$ pour tout $x \in [-1,1]^d$ et ensuite utiliser la formule de l'espérance totale

$$\mathbb{P}(f(X) \neq Y) = \mathbb{E}\left(\mathbb{P}(f(X) \neq Y|X)\right).$$

   On pourra aussi utiliser (sans démonstration)

$$\mathbb{P}(f(X) \neq Y|X=x) = \mathbb{P}(f(x) \neq Y|X=x)$$

et, que si $0 \leq t \leq 1/2$ alors $t = \min(t, 1-t)$.

2. *Question bonus plus difficile.* Pour $\alpha \in \mathbb{R}$ fixé, prouver que lorsque $||\beta|| \to +\infty$,

$$\mathbb{P}(f(X) \neq Y) \to 0.$$

3. On considère le risque de classification $\mathbb{P}(g(X) \neq Y)$ où $g : [-1,1]^d \to \mathbb{R}$ est un autre classifieur qui attribue la classe 1 à x si $C_\gamma(x) \geq 1/2$ et la classe 0 sinon. Ici $\gamma \in [-1,1]^d$ est un vecteur différent de β. Prouver que

$$\mathbb{P}(g(X) \neq Y) = \frac{1}{2^d} \int_{[-1,1]^d} \left[\mathbf{1}_{g(x)=f(x)} \min\left(C_\beta(x), 1 - C_\beta(x)\right) + \mathbf{1}_{g(x)\neq f(x)} \max\left(C_\beta(x), 1 - C_\beta(x)\right)\right] dx.$$

4. Prouver que

$$\mathbb{P}(f(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y).$$

Interpreter ce résultat.

Exercice 4

On considère n couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid et de même loi que (X, Y) dans l'exercice précédent. On fixe $x_1, \ldots, x_n \in [-1,1]^d$.

1. On fixe $y_1, \ldots, y_n \in \{0,1\}$. Calculer

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n\right).$$

en fonction de β. On admettra que

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n\right) = \mathbb{P}\left(Y_1 = y_1 | X_1 = x_1\right) \times \cdots \times \mathbb{P}\left(Y_n = y_n | X_n = x_n\right).$$

Indication : on pourra montrer que

$$\mathbb{P}\left(Y_i = y_i | X_i = x_i\right) = C_\beta(x_i)^{y_i}\left(1 - C_\beta(x_i)\right)^{1-y_i}.$$

2. On note $\mathcal{L}(\beta) = -\log(P(\beta))$ où $P(\beta)$ est la probabilité à calculer dans la question précédente. On note $M(\beta)$ sa matrice Hessienne calculée en β. La matrice M est donc de taille $d \times d$ et son élément i, j est égal à $\partial^2 \mathcal{L}(\beta)/\partial\beta_i\partial\beta_j$. Montrer que

$$M(\beta) = \sum_{i=1}^n \frac{e^{x_i^\top \beta}}{\left(1 + e^{x_i^\top \beta}\right)^2} x_i x_i^\top.$$

3. On pose

$$c(\beta) = \left(\min_{i=1,\ldots,n} \frac{e^{x_i^\top \beta}}{\left(1 + e^{x_i^\top \beta}\right)^2}\right) \lambda_{\min}(X^\top X)$$

où $\lambda_{\min}(X^\top X)$ est la plus petite valeur propre de la matrice $X^\top X$. Montrer que pour tout vecteur v avec $||v|| = 1$, on a

$$v^\top M v \geq c(\beta)||v||^2.$$

4. *Question bonus plus difficile.* On suppose que

$$\lim_{||\beta|| \to +\infty} \mathcal{L}(\beta) = +\infty.$$

Montrer que le problème d'optimisation

$$\min_{\beta \in \mathbb{R}^d} \mathcal{L}(\beta)$$

admet un unique minimiseur. *On dira alors que l'estimateur du maximum de vraissemblance est unique dans le modèle de régression logistique.*

TD 3
Arbres de classification

Ici log est le logarithme neperien.

Exercice 1

On considère un jeu de données $v_1, \dots, v_N \in \{1, 2, \dots, k\}$ (les nombres symbolisent k classes). On définit alors *l'entropie empirique* comme

$$E(v_1, \dots, v_N) = -\sum_{\ell=1}^{k} \hat{p}_\ell \log(\hat{p}_\ell)$$

en définissant

$$\hat{p}_\ell = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{v_i=\ell}$$

et en utilisant la convention $0 * \log(0) = 0$.

L'entropie empirique représente la variabilité des données.

1) Montrer que l'entropie est ≥ 0 et qu'elle est égale à 0 si et seulement si $v_1 = \cdots = v_N$.

2) Calculer $E(1, 2, 3, 2, 3, 2, 2, 2, 1, 2)$

3) Calculer $E(1, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 2)$.

Exercice 2

On a un jeu de données défini par ce tableau.

i	1	2	3	4	5	6	7	8	9	10
$x^{(i)}$	$(0.1, 0.1)$	$(0.2, 0.2)$	$(0.3, 0.8)$	$(0.4, 0.4)$	$(0.5, 0.7)$	$(0.6, 0.3)$	$(0.7, 0.5)$	$(0.8, 0.9)$	$(0.9, 0.6)$	$(1, 1)$
y_i	1	2	1	1	1	2	2	2	2	2

On note $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$ pour $i = 1, \dots, 10$.

1) Calculer $E(y_1, \dots, y_{10})$.

2) On cherche à séparer les données $(y_1, \dots, y_{10})$ en 2 groupes. Le groupe 1 sera celui des y_i pour lesquels $x_1^{(i)} \leq t$ avec $t = 0.45$ et le groupe 2 sera celui des autres y_i. On note $u_1, \dots, u_k$ les éléments du groupe 1 et $v_1, \dots, v_\ell$ les éléments du groupe 2 (on a $k + \ell = 10$). Calculer

$$\frac{k}{10} E(u_1, \dots, u_k) + \frac{\ell}{10} E(v_1, \dots, v_\ell)$$

qui est l'entropie empirique moyenne après séparation en 2 groupes. Faire un dessin qui représente cela.

3) On cherche à nouveau à séparer les données $(y_1, \dots, y_{10})$ en 2 groupes. Cette fois le groupe 1 sera celui des y_i pour lesquels $x_2^{(i)} \leq s$ avec $s = 0.45$ et le groupe 2 sera celui des autres y_i. On note $u_1, \dots, u_k$ les éléments du groupe 1 et $v_1, \dots, v_\ell$ les éléments du groupe 2 (on a $k + \ell = 10$). Calculer

$$\frac{k}{10} E(u_1, \dots, u_k) + \frac{\ell}{10} E(v_1, \dots, v_\ell).$$

Faire un dessin qui représente cela.

4) Quelle séparation en deux groupes préférez-vous dans une optique de classification supervisée, et pourquoi ?

5) La construction d'un arbre de classification se fait selon le principe des questions précédentes. Nous allons illustrer cela en quelques étapes ici (faire un dessin à chaque étape).

— Faire la séparation de la question 2, mais cette fois, trouver la valeure de t qui minimise la quantité

$$\frac{k}{10} E(u_1, \dots, u_k) + \frac{\ell}{10} E(v_1, \dots, v_\ell)$$

que l'on notera e_1.

— Ensuite, faire la même chose mais selon la question 3 en trouvant la valeur de s qui minimise la quantité

$$\frac{k}{10}E(u_1, \ldots, u_k) + \frac{\ell}{10}E(v_1, \ldots, v_\ell)$$

que l'on notera e_2.

— Garder celle des deux séparations qui correspond à la plus petite valeure entre e_1 et e_2. Cela revient à diviser le carré $[0,1]^2$ en 2 rectangles.

— Dans chacun des deux rectangles faire la même chose que toutes les étapes d'avant (si il y a encore deux classes représentées). A la fin, on a divisé le carré $[0,1]^2$ en 3 ou 4 rectangles. Cela correspond aux premières étapes de construction d'un arbre de classification. Dans chacun des rectangles, on classifie un nouveau x selon la classe qui est majoritaire dans le rectangle, parmi les 10 données d'apprentissage.