Université de Caen Examen du 20 avril 2018. Durée 2h00 **Analyse de données**
UFR des Sciences Par Faïcel Chamroukhi **M1 Statistique & Analyse Décisionnelle/Informatique**

Consignes :
- Sont interdits : Documents, calculettes, téléphones, écouteurs, ordinateurs, tablettes.
- Il est interdit de composer avec un crayon.
- Votre feuille double d'examen doit porter, à l'emplacement réservé, vos nom, prénom, et signature.
- Cette zone réservée doit être cachée par collage.
- Vos feuilles intercalaires doivent être toutes numérotées.
- Le barème est donné à titre indicatif.

Exercice 1 *(4 pts)* Soit (X, Y) un couple de variables aléatoires réelles et soit $((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon de n observations. Chacune des situations présentées dans la Figure 1 représente le nuage de données d'un échantillon de taille $n = 500$. Pour chaque situation, donner une valeur approchée du cœfficient de corrélation linéaire empirique r et justifier votre réponse.

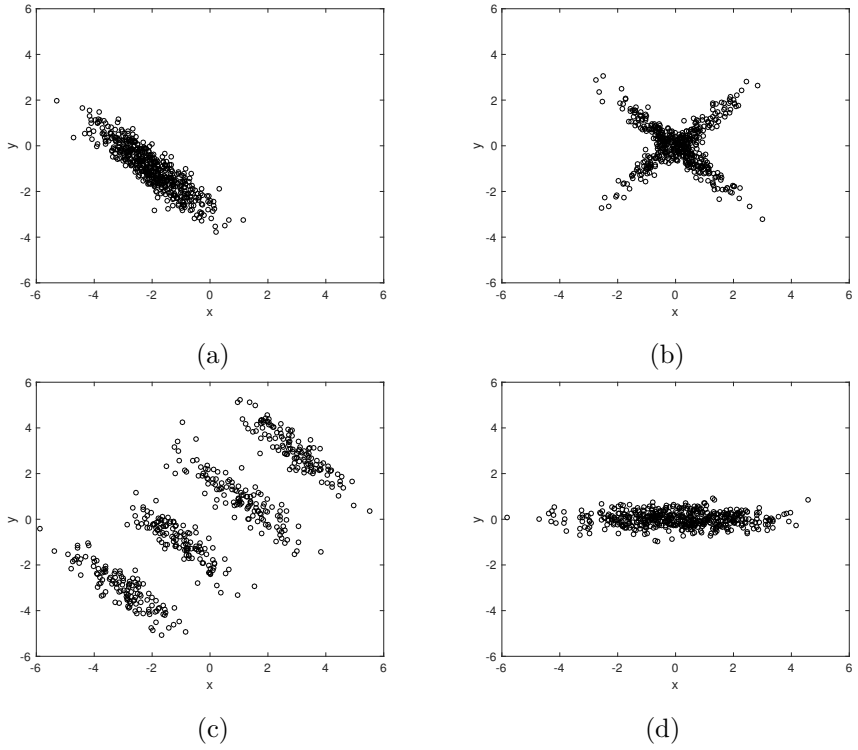


FIGURE 1 – Nuages de données ($\circ$)

Exercice 2 *(4 pts)* On considère un échantillon aléatoire indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ du couple $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ est un vecteur composé de $p = 2$ prédicteurs réels ($\boldsymbol{X} \in \mathbb{R}^2$) et $Y \in [\![0, 1]\!]$ une variable à prédire. On dispose d'un échantillon observé d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et on cherche à prédire les classes de nouvelles observations $(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_m)$ sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère l'analyse discriminante où la densité de la classe $k \in [\![0, 1]\!]$ est définie par : $f(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}) = \frac{1}{2\pi |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right)$. La prédiction s'effectue par la règle du maximum a posteriori (MAP) qui consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori :

$$\widehat{y}_i = \arg \max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}),$$

$\boldsymbol{\theta}$ étant le vecteur paramètre du modèle. On note par $\pi_k = \mathbb{P}(Y_i = k)$, la probabilité a priori de la classe k. On suppose que $\pi_0 = 0.5$, $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 = \mathbf{g}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$ et $\boldsymbol{\Sigma}_1 = \lambda \boldsymbol{\Sigma}_0$, où $\lambda > 1$ et $\mathbf{I}$ est la matrice identité.

1. Montrer que $Y = 1$ si et seulement si $||\boldsymbol{x} - \mathbf{g}||_2^2 \geq r$, en déterminant r en fonction de λ.
2. A quoi correspond la frontière de décision dans ce cas ?

Exercice 3 *(4 pts)* On considère le cadre de l'exercice précédent et on se propose maintenant d'utiliser un autre modèle de prédiction, celui de régression logistique non-linéaire suivant défini par

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp(\phi(\boldsymbol{x}; \boldsymbol{\theta}))}{1 + \exp(\phi(\boldsymbol{x}; \boldsymbol{\theta}))} \tag{1}$$

où $\phi(\boldsymbol{x}; \boldsymbol{\theta})$ étant une transformation non-linéaire (polynomiale) de $\boldsymbol{x}$ et $\boldsymbol{\theta}$ le vecteur paramètre du modèle. La prédiction avec ce modèle consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori, i.e. :

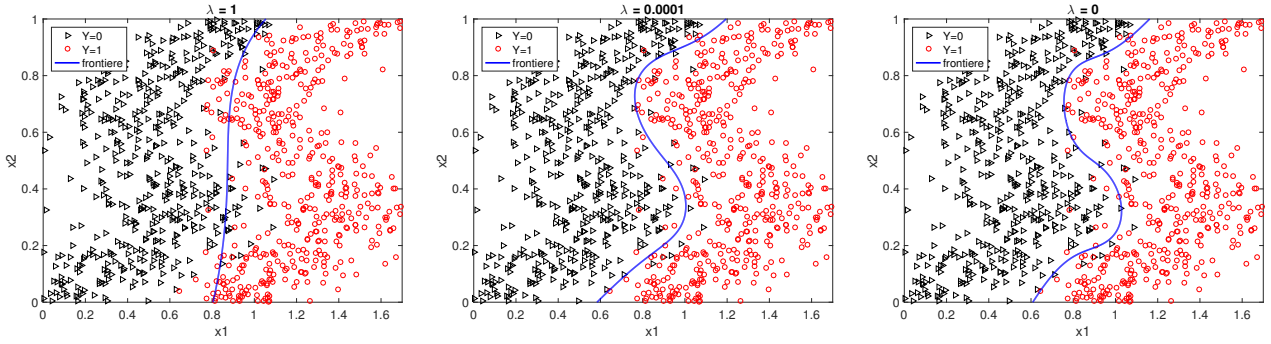
$$\widehat{y}_i = \arg \max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k|\boldsymbol{X}_i = \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}),$$

où $\widehat{\boldsymbol{\theta}}$ étant le vecteur paramètre du modèle appris en minimisant le risque quadratique régularisé suivant :

$$\ell_\lambda(\boldsymbol{\theta}) = -\ln L(\boldsymbol{\theta}) + \lambda\Omega(\boldsymbol{\theta}) \tag{2}$$

où $\ln L(\boldsymbol{\theta})$ est la fonction de log-vraisemblance classique et $\Omega(\boldsymbol{\theta})$ une fonction de pénalité sur $\boldsymbol{\theta}$ de niveau λ dont l'objectif est d'assurer un compromis entre complexité du modèle et ajustement aux données.

1. Rappeler le nom et le principe de l'algorithme d'estimation des paramètres en régression logistique.
2. Rappeler la décomposition biais-variance du risque quadratique.
3. Chacune des figures 2-(a, b, c) montre un échantillon d'apprentissage et un modèle de prédiction estimé $\widehat{\boldsymbol{\theta}}$ en minimisant (2), représenté par la frontière de décision, pour la même régularisation $\Omega(\boldsymbol{\theta})$ et différentes valeurs de λ. Discuter la qualité prédictive de chacun des trois modèles.



(a) (b) (c)

FIGURE 2 – Nuage de données (○,▷) et modèle de prédiction (—)

Exercice 4 *(8 pts)* On considère un échantillon aléatoire indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ du couple $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ est un vecteur de p prédicteurs binaires ($\boldsymbol{X} \in [\![0,1]\!]^p$) et $Y \in [\![0,1]\!]$ une variable à prédire représentant la classe de $\boldsymbol{X}$. On dispose d'un échantillon observé d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et on cherche à prédire les classes de nouvelles observations $(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_m)$ sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère la classifieur Bayésien naïf, une méthode probabiliste de prédiction qui repose sur l'hypothèse simplificatrice forte suivante :

$\mathcal{H}$: Pour chaque classe Y_i, les variables $X_{ij}, j = 1, \ldots, p$ de chaque individu $\boldsymbol{X}_i$, sont indépendantes.

La prédiction s'effectue par la règle du maximum a posteriori (MAP) qui consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori :

$$\widehat{y}_i = \arg\max_{k\in[\![0,1]\!]} \mathbb{P}(Y_i = k|\boldsymbol{X}_i = \boldsymbol{x}_i; \widehat{\boldsymbol{\Psi}}), \tag{3}$$

$\widehat{\boldsymbol{\Psi}}$ étant l'estimateur du vecteur paramètre $\boldsymbol{\Psi}$ du modèle estimé sur l'ensemble d'apprentissage.

1. Les variables binaires X_{ij} de chaque vecteur $\boldsymbol{X}_i$ de la même classe $Y_i = k$ sont de loi Bernoulli de paramètre $0 < \theta_{kj} < 1$, i.e. :

   $$\mathbb{P}(X_{ij} = x_{ij}|Y_i = k; \theta_{kj}) = {\theta_{kj}}^{x_{ij}}(1-\theta_{kj})^{1-x_{ij}} \text{ où } \theta_{kj} = \mathbb{P}(X_{ij} = 1|Y_i = k; \theta_{kj}).$$

   En déduire d'après $\mathcal{H}$ la loi conditionnelle du vecteur $\boldsymbol{X}_i$, notée $\mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i|Y_i = k; \boldsymbol{\theta}_k)$.

2. On note par $\alpha = \mathbb{P}(Y_i = 1)$, la probabilité a priori de la classe 1. On a alors $\boldsymbol{\Psi} = (\alpha, \boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_0^\top)^\top$ que l'on cherche à estimer la méthode du maximum de vraisemblance. Donner l'expression de la fonction de vraisemblance

   $$L(\boldsymbol{\Psi}) = \mathbb{P}\big((\boldsymbol{X}_1 = \boldsymbol{x}_1, Y_1 = y_1), \ldots, (\boldsymbol{X}_n = \boldsymbol{x}_n, Y_n = y_n); \boldsymbol{\Psi}\big). \tag{4}$$

3. En déduire celle de la fonction de log-vraisemblance $\ln L(\boldsymbol{\Psi})$.

4. Montrer en maximisant la log-vraisemblance que les estimateurs du maximum de vraisemblance sont donnés par :

   $$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i; \quad \widehat{\theta}_{1j} = \frac{\sum_{i=1}^n Y_i X_{ij}}{\sum_{i=1}^n Y_i}; \quad \widehat{\theta}_{0j} = \frac{\sum_{i=1}^n (1-Y_i) X_{ij}}{\sum_{i=1}^n (1-Y_i)}.$$

   soit en formulation vectorielle pour les $\boldsymbol{\theta}$ par :

   $$\widehat{\boldsymbol{\theta}}_1 = \frac{\sum_{i=1}^n Y_i \boldsymbol{X}_i}{\sum_{i=1}^n Y_i}; \quad \widehat{\boldsymbol{\theta}}_0 = \frac{\sum_{i=1}^n (1-Y_i) \boldsymbol{X}_i}{\sum_{i=1}^n (1-Y_i)}.$$

5. On considère le jeu de données de la Table 1. Prédire en appliquant la règle (3) la valeur de Y pour le dernier vecteur.

$\boldsymbol{X}$		Y
0	1	1
1	0	1
1	0	0
0	1	0
0	1	1
1	1	?

TABLE 1 – Jeu de données