CORRECTIONS DE L'EXAMEN DE CLASSIFICATION DU LUNDI 26 NOVEMBRE 2018.

Exercice 1 : Analyse en composantes principales (6 points)

1. (1 pt) Quelle est la normalisation à effectuer avant d'effectuer l'ACP des individus ? Expliquer ce que devient la note 6 obtenue par l'individu $\boldsymbol{x}_1$ dans la matière "Maths" après normalisation (sans effectuer les calculs).
   Réponse: il faut centrer et réduire les données du tableau, c'est-à-dire enlever la moyenne de la variable considérée et diviser par son écart-type. Par exemple, il faut soustraire de la note de maths de l'individu $\boldsymbol{x}_1$ la moyenne des notes de math et diviser le tout par l'écart-type des notes de maths.

2. (1 pt) Les valeurs propres de la matrice de covariance sont $\lambda_5 = 0.0004, \lambda_4 = 0.0039, \lambda_3 = 0.9831, \lambda_2 = 1.1507$ et $\lambda_1 = 2.8618$. Expliquer comment choisir le nombre de composantes principales p à retenir pour l'ACP des individus.
   Réponse: En général on choisit un nombre de valeurs propres p tel que
   $$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{5} \lambda_i}$$
   est proche de 1. On obtient ici
   $$\frac{\sum_{i=1}^{2} \lambda_i}{\sum_{i=1}^{5} \lambda_i} = 0.8025 \quad \text{et} \quad \frac{\sum_{i=1}^{3} \lambda_i}{\sum_{i=1}^{5} \lambda_i} = 0.9991.$$
   On doit donc choisir $p = 3$ pour conserver au moins 95% de l'information.

3. (1 pt) Les résultats de l'ACP des individus dans le plan des deux premières composantes principales sont représentés sur la figure 1a. Expliquer la signification des deux premières composantes principales pour cet exemple.
   Réponse: la première composante sépare les individus les mieux notés situés à droite du premier plan principal de ceux mal notés situés à gauche de ce plan. La seconde composante principale a tendance à opposer les matières scientifiques (Maths, Sciences) aux matières non scientifiques (French, Latin, Music).

4. (1 pt) La projection du premier individu dans le premier plan principal est $\widetilde{\boldsymbol{x}}_1 = (-2.7857, -0.6765)^T$. Déterminer la qualité de la représentation de cet individu dans le plan formé par les deux premières composantes principales.
   Réponse: D'après le cours, la qualité de représentation d'un individu $\boldsymbol{x}_i$ sur les deux premiers axes principaux sont $\cos^2(\theta_{i,1}) = \frac{||\boldsymbol{u}^t \boldsymbol{x}_i||^2}{||\boldsymbol{x}_i||^2}$ et $\cos^2(\theta_{i,2}) = \frac{||\boldsymbol{v}^t \boldsymbol{x}_i||^2}{||\boldsymbol{x}_i||^2}$, où $\boldsymbol{u}$ et $\boldsymbol{v}$ sont les deux vecteurs propres associés aux valeurs propres les plus grandes de la matrice de covariance des données. Pour l'individu $\boldsymbol{x}_1$, on obtient $\cos^2(\theta_{i,1}) = (-2.7857)^2/8.7641 = 0.8854$ et $\cos^2(\theta_{i,2}) = (-0.6765)^2/8.7641 = 0.0522$, d'où $\cos^2(\theta_{i,1}) + \cos^2(\theta_{i,2}) = 0.9376$. Cet individu est donc bien représenté dans le premier plan principal.

5. (1 pt) Analyser les corrélations entre les différentes variables à l'aide de la figure 1b.
   Réponse: les variables "French" et "Latin" sont très corrélées, de même que les variables "Maths" et "Sciences". Les paires "(French, Latin)" et "(Maths, Sciences)" sont quasi indépendantes car l'angle entre ces paires est proche de 90 degrés.

6. (1 pt) Expliquer à l'aide de quelques exemples comment on peut utiliser la figure 1c.
   Réponse: La représentation simultanée permet d'analyser comment un individu se projette sur les axes portés par les différentes variables. Par exemple, l'individu $\boldsymbol{x}_4$ a les meilleurs résultats tandis que les individus $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ ont de mauvais résultats dans toutes les matières.

Exercice 2 : Classification Bayésienne (4 points)

On considère un problème de classification à trois classes ω_1, ω_2 and ω_3 de densités normales

$$f(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{x^2}{2\sigma_i^2}\right), x \in \mathbb{R} \quad i = 1,2,3 \quad \text{avec} \quad \sigma_1 > \sigma_2 > \sigma_3. \tag{1}$$

1. (2 pts) Déterminer la règle de classification associée à ce problème avec la fonction de coût $0-1$ et lorsque les trois classes sont équiprobables. Pour les applications numériques, on prendra $\sigma_1^2 = 5$, $\sigma_2^2 = 2$ et $\sigma_3^2 = 1$.
   Réponse: le classifieur Bayésien affecte x à la classe ω_1 (ce que l'on notera $d^*(x) = \omega_1$) si

   $$f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2) \quad \text{et} \quad f(x|\omega_1)P(\omega_1) \geq f(x|\omega_3)P(\omega_3)$$

   c'est-à-dire, en utilisant l'équiprobabilité des deux classes et le fait que $\sigma_1^2 > \sigma_2^2$ et $\sigma_1^2 > \sigma_3^2$

   $$d^*(x) = \omega_1 \Leftrightarrow x^2 \geq a_{12}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \text{ et } x^2 \geq a_{13}^2 = \frac{\sigma_1^2\sigma_3^2}{\sigma_1^2 - \sigma_3^2} \ln\left(\frac{\sigma_3^2}{\sigma_1^2}\right)$$

   c'est-à-dire

   $$d^*(x) = \omega_1 \Leftrightarrow x^2 \geq \max\{a_{12}^2, a_{13}^2\} = a_{12}^2.$$

   De même

   $$d^*(x) = \omega_2 \Leftrightarrow x^2 \leq a_{12}^2 \text{ et } x^2 \geq a_{23}^2 = \frac{\sigma_2^2\sigma_3^2}{\sigma_2^2 - \sigma_3^2} \ln\left(\frac{\sigma_3^2}{\sigma_2^2}\right)$$

   et

   $$d^*(x) = \omega_3 \Leftrightarrow x^2 \leq a_{23}^2.$$

2. (2 pts) On rappelle que si $X \sim \mathcal{N}(0,1)$ alors $Y = X^2$ suit une loi du chi deux à 1 degré de liberté dont la fonction de répartition est notée ϕ. Expliquer comment calculer la probabilité d'erreur associée à ce classifieur en fonction de $\sigma_1, \sigma_2, \sigma_3$ et de la fonction ϕ (on pourra se limiter au calcul d'un des termes de cette probabilité d'erreur).
   Réponse: la probabilité d'erreur du classifieur est définie par

   $$P_e = \sum_{i=1}^{3} \sum_{j \neq i} P[d^*(x) = \omega_i | x \in \omega_j] P(x \in \omega_j).$$

   Par exemple, le premier terme se calcule comme suit

   $$P_1 = P[d^*(x) = \omega_1 | x \in \omega_2] P(x \in \omega_2) = P[x^2 > a_{12}^2 | x \sim \mathcal{N}(0, \sigma_2^2] P(x \in \omega_2).$$

   Si $x \sim \mathcal{N}(0, \sigma_2^2$ alors $\frac{x^2}{\sigma_2^2}$ suit une loi du chi deux à 1 degré de liberté. Donc

   $$P_1 = P\left[\frac{x^2}{\sigma_2^2} > \frac{a_{12}^2}{\sigma_2^2} | x \sim \mathcal{N}(0, \sigma_2^2\right] P(x \in \omega_2).$$

   soit

   $$P_1 = P(x \in \omega_2)\left[1 - \phi\left(\frac{a_{12}^2}{\sigma_2^2}\right)\right].$$

Questions sur l'article (10 **points)**

1. (1 pt) Expliquer le terme "Novelty Detection" utilisé dans le titre.
   Réponse : l'algorithme proposé dans cet article permet de détecter des signaux ayant des comportements anormaux, c'est-à-dire des anomalies. Ces anomalies n'ayant pas été observées auparavant, elles sont qualifiées de "nouveautés".

2. (1 pt) A quoi correspond la courbe d'équation $(\boldsymbol{w}.\Phi(\boldsymbol{x})) - \rho = 0$?
   Réponse : cette équation est l'équation d'une courbe englobant la majorité des vecteurs de la base d'apprentissage contenant des signaux normaux. Les vecteurs $\boldsymbol{x}$ situés à l'intérieur de cette courbe vérifieront $(\boldsymbol{w}.\Phi(\boldsymbol{x})) - \rho > 0$ tandis que ceux situés à l'extérieur de cette courbe seront déclarés comme anomalies et vérifieronf $(\boldsymbol{w}.\Phi(\boldsymbol{x})) - \rho < 0$.

3. (1 pt) Expliquer le role de la variable ξ_i dans (4) et pourquoi le terme $\sum_i \xi_i$ apparait dans la fonction à minimiser.
   Réponse : la variable ξ_i est associée au vecteur $\boldsymbol{x}_i$ de la base d'apprentissage et permet à ce vecteur de violer la contrainte $(\boldsymbol{w}.\Phi(\boldsymbol{x_i})) - \rho > 0$. Lorsque $\xi_i = 0$, cette contrainte est vérifiée tandis que si $\xi_i > 0$ la contrainte n'est pas satisfaite. L'idée est de trouver la courbe contenant le maximum de vecteurs $\boldsymbol{x}_i$ de la base d'apprentissage, c'est-à-dire qu'on cherche à avoir un maximum de variables ξ_i nulles, ce qu'on va obtenir en ajoutant le terme $\sum_i \xi_i$ dans la fonction de coût à minimiser.

4. (1 pt) Quel est l'intérêt du terme de régularisation $\frac{1}{2}||\boldsymbol{w}||^2$ intervenant dans (3) ?
   Réponse : on montre que la marge du classifieur est inversement proportionnelle à $||\boldsymbol{w}||$. Donc, minimiser $\frac{1}{2}||\boldsymbol{w}||^2$ va permettre de chercher un hyperplan séparateur dans le domaine transformé le plus loin de l'origine.

5. (1 pt) Quelle est l'expression du Lagrangien associé au problème d'optimisation défini par (3) et (4) ?
   Réponse : Le Lagrangien s'écrit

$$L(\boldsymbol{w}, \xi, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{\nu l}\sum_i \xi_i - \rho - \sum_i \alpha_i((\boldsymbol{w}.\Phi(\boldsymbol{x}_i)) - \rho + \xi_i) - \sum_i \beta_i \xi_i.$$

6. (1 pt) Expliquer comment on obtient le problème dual défini par (6).
   Réponse : on annule les dérivées du Lagrangien par rapport à $\boldsymbol{w}$, ξ_i et ρ et on remplace les expressions de $\boldsymbol{w}$ et la contrainte $\sum_i \alpha_i = 1$ dans le Lagrangien.

7. (1 pt) Expliquer le rôle du paramètre ν ?
   Réponse : ce paramètre est une borne maximale du pourcentage d'éléments de la base d'apprentissage situés à l'extérieur de la frontière d'équation $(\boldsymbol{w}.\Phi(\boldsymbol{x})) - \rho = 0$ (appelés "outliers"). C'est également une borne inférieure du pourcentage de vecteurs supports. Fixer ν à une faible valeur permet de s'assurer que le nombre d'outliers est faible.

8. (1 pt) Expliquer les résultats de la figure 1.
   Réponse : cette figure montre que lorsqu'on diminue ν, le nombre d'outliers diminue (ce qui illustre la réponse à la question précédente). Cette figure montre également l'effet du paramètre c sur les résultats de l'algorithme. Plus c est petit, plus la frontière est autorisée à être irrégulière.

9. (1 pt) Expliquer comment les auteurs proposent de résoudre un problème de classification (par exemple le problème de reconnaissance de caractères qui donne les résultats de la figure 2).
   Réponse : le problème considéré est un problème de classification à 10 classes. Les auteurs proposent d'ajouter dix dimensions au vecteur $\boldsymbol{x}_i$, ce qui permet de coder le numéro de la classe du vecteur $\boldsymbol{x}_i$.

10. (1pt) A quoi correspondent les caractères représentés sur la figure 2 ?
    Réponse : les caractères représentés sur la figure 2 sont les éléments anormaux détectés par l'algorithme.