CORRECTIONS DE L'EXAMEN DE CLASSIFICATION DU LUNDI 27 DÉCEMBRE 2017.

Exercice 1 : Classification Bayésienne pour des lois de Rayleigh (4 points)

On considère un problème de classification à deux classes ω_1 and ω_2 de densités

$$f(x|\omega_i) = \frac{x}{\sigma_i^2} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) I_{\mathbb{R}^+}(x) \quad i = 1, 2 \tag{1}$$

où $I_{\mathbb{R}^+}(x)$ est la fonction indicatrice sur $\mathbb{R}^+$ ($I_{\mathbb{R}^+}(x) = 1$ si $x > 0$ et $I_{\mathbb{R}^+}(x) = 0$ sinon) et $\sigma_1^2 > \sigma_2^2$.

1. (1.5 pt) Déterminer la règle de classification associée à ce problème avec la fonction de coût $0-1$ et lorsque les deux classes sont équiprobables.
   Réponse: le classifieur Bayésien affecte x à la classe ω_1 (ce que l'on notera $d^*(x) = \omega_1$) si

   $$f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2)$$

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

   $$d^*(x) = \omega_1 \Leftrightarrow x^2 \geq a^2 = \frac{2(\sigma_1^2\sigma_2^2)}{\sigma_2^2 - \sigma_1^2} \ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$$

   c'est-à-dire, en remarquant que $x > 0$

   $$d^*(x) = \omega_1 \Leftrightarrow x > a$$

   avec

   $$a = \sqrt{\frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2 - \sigma_1^2} \ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)}$$

2. (1.5 pts) Déterminer la probabilité d'erreur associée.
   Réponse: la probabilité d'erreur d'un classifieur est définie par

   $$P_e = P[d^*(x) = \omega_1 | x \in \omega_2]P(x \in \omega_2) + P[d^*(x) = \omega_2 | x \in \omega_1]P(x \in \omega_1)$$

   ce qui donne dans notre cas

   $$P_e = \frac{1}{2}P[x > a | x \in \omega_2] + \frac{1}{2}P[x < a | x \in \omega_1]$$

   soit

   $$P_e = \frac{1}{2}\int_a^\infty \frac{x}{\sigma_2^2} \exp\left(-\frac{x^2}{2\sigma_2^2}\right) dx + \frac{1}{2}\int_0^a \frac{x}{\sigma_1^2} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) dx.$$

   Des calculs élémentaires conduisent à

   $$P_e = \frac{1}{2} \exp\left(-\frac{a^2}{2\sigma_2^2}\right) + \frac{1}{2}\left[1 - \exp\left(-\frac{a^2}{2\sigma_1^2}\right)\right]$$

   avec la valeur de a déterminée précédemment.

3. (1 pt) Si le paramètre σ_1^2 est inconnu, expliquer comment l'estimer à partir de données d'apprentissage de la classe ω_1 en utilisant la méthode du maximum de vraisemblance.
   Réponse: soient $x_1, ..., x_n$ les données de la base d'apprentissage appartenant à la classe ω_1. La vraisemblance de ces n données s'écrit

   $$L(x_1, ..., x_n; \sigma_1^2) = \prod_{i=1}^n \left[\frac{x_i}{\sigma_1^2} \exp\left(-\frac{x_i^2}{2\sigma_1^2}\right)\right] \propto \frac{1}{\sigma_1^{2n}} \exp\left(-\frac{1}{2\sigma_1^2}\sum_{i=1}^n x_i^2\right).$$

L'estimateur du maximum de vraisemblance du paramètre σ_1^2 s'obtient en maximisant la vraisemblance $L(x_1, ..., x_n; \sigma_1^2)$ par rapport) σ_1^2. Des calculs élémentaires permettent d'obtenir

$$\widehat{\sigma}_1^2 = \frac{1}{2n} \sum_{i=1}^{n} x_i^2.$$

Exercice 2 : Arbres de décision (4 points)

On cherche à construire un arbre de décision permettant de décider si un individu doit jouer au tennis ou non. Une base d'apprentissage a été construite comme suit

	Ciel	Température	Vent	Jouer
$\boldsymbol{x}_1$	soleil	chaud	faible	Oui
$\boldsymbol{x}_2$	soleil	chaud	fort	Oui
$\boldsymbol{x}_3$	couvert	chaud	faible	Non
$\boldsymbol{x}_4$	pluie	froid	faible	Non
$\boldsymbol{x}_5$	pluie	froid	faible	Non
$\boldsymbol{x}_6$	pluie	froid	fort	Oui

1. (1 pt) Déterminer l'indice de Gini associé à cette base d'apprentissage vis-à-vis des deux classes "Jouer au Tennis" et "Ne pas jouer au Tennis".
   Réponse: L'indice de Gini de la base d'apprentissage s'écrit

$$\sum_{i=1}^{2} \frac{n_i}{n} \left(1 - \frac{n_i}{n}\right) = \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2}.$$

2. (2pts) Déterminer la variation de l'indice de Gini lorsqu'on coupe les données à l'aide des variables "Ciel", "Température" et "Vent". En déduire la variable qui sera utilisée au premier niveau de l'arbre de décision.
   Réponse: La variable "Ciel" coupe la base d'apprentissage en trois sous ensembles associés aux valeurs "Soleil", "Couvert" et "Pluie" qui correspondent aux valeurs de "Jouer" égales à $\{0, 0\}$, $\{\mathrm{N}\}$ et $\{\mathrm{N}, \mathrm{N}, \mathrm{O}\}$. L'indice de Gini associé à ces trois sous ensembles est

$$\left(\frac{2}{6} \times 0\right) + \left(\frac{1}{6} \times 0\right) + \left(\frac{3}{6} \times 2 \times \frac{2}{3} \times \frac{1}{3}\right) = \frac{2}{9} \approx 0.22.$$

   Pour la variable "Température", on obtient deux sous ensembles $\{\mathrm{O}, \mathrm{O}, \mathrm{N}\}$ et $\{\mathrm{N}, \mathrm{N}, \mathrm{O}\}$, d'où l'indice de Gini

$$\left(\frac{1}{2} \times 2 \times \frac{2}{3} \times \frac{1}{3}\right) \times 2 = \frac{4}{9} \approx 0.44.$$

   Enfin pour la variable "Vent", on obtient $\{\mathrm{O}, \mathrm{N}, \mathrm{N}, \mathrm{N}\}$ et $\{\mathrm{O}, \mathrm{O}\}$ avec un indice de Gini égal à

$$\frac{2}{3} \times 2 \times \frac{1}{4} \times \frac{3}{4} = \frac{1}{4} = 0.25.$$

   On en conclut que la variable "Ciel" permet d'obtenir la réduction la plus importante de l'indice de Gini. Ce sera donc cette variable qui sera au premier niveau de l'arbre de décision.

3. (1pt) Expliquer comment on pourrait procéder si la variable "Température" était une valeur en degrés celcius.
   Réponse: Dans ce cas, on sépare les valeurs de températures vérifiant $x_i > S$ et $x_i < S$ pour toutes les valeurs possibles des seuils S et on garde à chaque fois la valeur du seuil qui permet d'obtenir la diminution la plus importante de l'indice de Gini.

Questions sur l'article (12 **points)**

1. (1 pt) Expliquer pourquoi la matrice $\boldsymbol{X}$ représentée sur la Fig. 1 est idéalement diagonale par blocs.
   Réponse : les premières colonnes de la matrice $\boldsymbol{Y}$ représentées en vert dans la Fig. 1 correspondent aux vecteurs de la base d'apprentissage de la classe $\#1$. Idéalement, ces élément se décomposent comme une combinaison linéaire des éléments du dictionnaire associés à cette classe, ce qui correspond au premier bloc diagonal de $\boldsymbol{X}$ représenté en vert. De même les colonnes rouge de la matrice $\boldsymbol{Y}$ de la Fig. 1 correspondent aux vecteurs de la base d'apprentissage de la classe $\#2$ qui se décomposent idéalement comme une combinaison linéaire des éléments du dictionnaire associés à la seconde classe, et ainsi de suite. La matrice $\boldsymbol{X}$ ainsi obtenue est donc bien idéalement diagonale par blocs.

2. (1 pt) Expliquer la phrase "It has been shown that learning a dictionary from the training samples instead of using all of them as a dictionary can further enhance the performance of SRC".
   Réponse : Ceci signifie que les performances de classification sont meilleures lorsqu'on estime (apprend) les éléments du dictionnaire à partir des éléments de la base d'apprentissage et des vecteurs d'observation, plutôt que de construire le dictionnaire $\boldsymbol{D}$ en mettant dans chaque colonne un vecteur de cette base d'apprentissage (ce qui éviterait d'estimer le dictionnaire).

3. (1 pt) Quel est l'intérêt d'ajouter le dictionnaire partagé $\boldsymbol{D}_0$ aux dictionnaires des différentes classes $\boldsymbol{D}_1, ...\boldsymbol{D}_C$?
   Response : l'idée de cet article est de supposer que chaque vecteur se décompose comme une combinaison linéaire d'atomes discriminants qui sont propres à une classe et d'autres atomes qui sont communs à toutes les classes rangés dans $\boldsymbol{D}_0$. Les éléments de $\boldsymbol{D}_0$ peuvent correspondre aux vecteurs situés à l'intersection de plusieurs classes comme cela est illustré sur la figure 3.

4. (1 pt) Les auteurs de l'article précisent (bas de la page 5162) qu'avec le dernier terme $\|\boldsymbol{X}\|_F^2$ la fonction de coût devient convexe par rapport à $\boldsymbol{X}$. Quel est l'intérêt d'avoir une fonction de coût convexe ?
   Réponse : le fait d'avoir une fonction de coût convexe assure l'existence d'un minimum global de cette fonction de coût et la convergence de l'algorithme d'optimisation alternée vers ce minimum global.

5. (1 pt) Expliquer comment les auteurs justifient le fait que le dictionnaire $\boldsymbol{D}_0$ doit être de rang faible (voir section II C).
   Response : la contrainte de rang faible sur $\boldsymbol{D}_0$ permet d'éviter que $\boldsymbol{D}_0$ contienne des atomes discriminants d'une des différentes classes. Si on n'imposait aucune contrainte sur le dictionnaire $\boldsymbol{D}_0$, on pourrait au pire se retrouver dans une situation où $\boldsymbol{D}_0$ contienne tous les atomes discriminants et alors les autres dictionnaires $\boldsymbol{D}_1, ..., \boldsymbol{D}_C$ seraient vides !

6. (1 pt) Comment est définie la norme nucléaire d'une matrice ("nuclear norm") et que peut-on dire d'une matrice $\boldsymbol{A}$ de norme nucléaire faible ?
   Réponse : la norme nucléaire d'une matrice $\boldsymbol{A}$ est définie par

$$\|\boldsymbol{A}\|_* = \text{Trace}\left(\sqrt{\boldsymbol{A}^*\boldsymbol{A}}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{A})$$

   où $\boldsymbol{A}^*$ est la matrice transposée conjuguée de $\boldsymbol{A}$ et où $\sigma_i(\boldsymbol{A})$ désigne la $i$ème valeur singulière de la matrice $\boldsymbol{A}$. Une matrice de de norme nucléaire faible est une matrice de faible rang.

7. (1 pt) Expliquer la présence du terme de régularisation $\|\boldsymbol{X}^0 - \boldsymbol{M}^0\|_F^2$ dans (4).
   Réponse : la matrice $\boldsymbol{M}^0$ est une matrice dont toutes les colonnes sont égales à la moyenne des éléments de $\boldsymbol{X}^0$. On cherche à minimiser le terme de régularisation $\|\boldsymbol{X}^0 - \boldsymbol{M}^0\|_F^2$ de manière à ce

que toutes les colonnes de $\boldsymbol{X}^0$ soient proches comme illustré sur la figure 3. Ceci permet d'obtenir des contributions de $\boldsymbol{X}^0$ proches dans chaque classe, ce qui évite que $\boldsymbol{X}^0$ contienne des éléments discriminants associés à l'une des C classes.

8. (1 pt) Expliquer la règle de classification donnée dans (7).
   Réponse : le vecteur $\bar{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{D}^0\boldsymbol{x}^0$ permet d'enlever la partie de $\boldsymbol{y}$ contenue dans le dictionnaire partagé. On va ensuite chercher la classe la plus proche du vecteur $\bar{\boldsymbol{y}}$ en cherchant la classe c qui minimise
   $$\|\bar{\boldsymbol{y}} - \boldsymbol{D}_c\boldsymbol{x}^c\|_2^2.$$
   Le second terme $\|\boldsymbol{x}_c - \boldsymbol{m}_c\|_2^2$ permet de s'assurer que le vecteur $\boldsymbol{x}_c$ est proche de $\boldsymbol{m}_c$.

9. (1 pt) Expliquer le principe utilisé pour optimiser la fonction $f_{\boldsymbol{Y}}(\boldsymbol{D}, \boldsymbol{X})$ de (3).
   Réponse : Pour optimiser la fonction de coût (3), les auteurs proposent d'utiliser une méthode d'optimisation alternée. On commence par estimer $\boldsymbol{D}$ pour $\boldsymbol{X}$ fixé (à l'aide de l'algorithme ODL) puis on estime $\boldsymbol{X}$ pour $\boldsymbol{D}$ fixé (à l'aide l'algorithme FISTA).

10. (1pt) Qu'est ce qu'on entend par "five-fold cross validation"? (voir Section IV A)
    Réponse : Comme expliqué dans [46], la méthode de validation croisée "k-fold cross-validation" consiste à découper l'ensemble d'apprentissage en k sous ensemble $E_1, ..., E_k$. Pour chaque sous-ensemble E_i, on construit un classifier à l'aide de toutes les données sauf celles de E_i et on teste le classifier à l'aide des données de E_i ce qui donne une erreur de classification e_i. L'erreur finale est obtenue en moyennant les erreurs e_i, soit
    $$e = \frac{1}{k}\sum_{i=1}^{k} e_i.$$
    Ici les auteurs proposent d'utiliser ce principe avec $k = 5$.

11. (1pt) Expliquer le term "Random projection matrix" utilisé pour la base de données "Extended YaleB" (voir Section IV A).
    Réponse : Comme expliqué dans [23], on projette les données en les multipliant par une matrice appelée "Randomface" qui est définie dans l'article de J. Wright [5]. Un visage aléatoire (randomface) est défini par (voir définition 2 page 217 de [5]) une matrice dont les éléments sont générés suivant une loi normale de moyenne nulle et telle que chaque ligne est normalisée de manière à avoir une norme ℓ_2 égale à 1. Les auteurs choisissent d'utiliser une telle matrice avec 504 lignes, ce qui fournira un vecteur avec 504 variables (features).

12. (1pt) Rappeler le principe de l'analyse en composantes principales utilisée pour les données "AR gender".
    Réponse : le principe de l'analyse en composantes principales est de projeter les données sur les vecteurs propres les plus discriminants de la matrice de covariance de ces données. Les auteurs proposent dans cet article de se limiter aux 300 vecteurs les plus discriminants (associés aux 300 valeurs propres les plus grandes de cette matrice de covariance), ce qui donne un vecteur à 300 variables.

13. Questions BONUS

- (1pt) Expliquer le principe de l'algorithme "Online Dictionary Learning (ODL)".
  Réponse : L'algorithme ODL proposé dans [35] consiste à mettre à jour les colonnes de $\boldsymbol{D}$ de manière itérative. Plus précisément, pour $t = 1, ..., T$, on choisit un élément $\boldsymbol{x}_t$ de la base d'apprentissage, on calcule son code $\boldsymbol{\alpha}$ en résolvant le problème suivant

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

  où $\boldsymbol{D}_{t-1}$ est le dictionnaire obtenu à l'itération précédente. Ensuite on cherche la matrice $\boldsymbol{D}$ solution du problème

$$\underset{\boldsymbol{D}}{\operatorname{argmin}} \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1.\right)$$

  La solution de ce problème s'obtient en minimisant la fonction ci-dessus par rapport à chaque colonne de $\boldsymbol{D}$, ce qui mène à une solution analytique car le critère est quadratique par rapport à chaque colonne de $\boldsymbol{D}$.

- (1pt) Pour minimiser la fonction de coût

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{x}\|_1$$

  l'algorithme "Iterative Shrinkage-Thresholding Algorithm (ISTA)" utilise la récursion

$$\boldsymbol{x}_{k+1} = \mathcal{T}_{\lambda t}\left(\boldsymbol{x}_k - 2t\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b})\right)$$

  où t est un pas judicieusement choisi. Pouvez vous rappeler le rôle de la fonction $\mathcal{T}_{\boldsymbol{a}}(\boldsymbol{x})$ définie de $\mathbb{R}^n$ dans $\mathbb{R}^n$?
  Réponse : la $i$ème composante de la fonction $\mathcal{T}_{\boldsymbol{a}}$ est définie pour $a > 0$ par

$$\mathcal{T}_{\boldsymbol{a}}(\boldsymbol{x})_i = (|x_i| - a)_+ \text{signe}(x_i)$$

  avec pour tout $u \in \mathbb{R}$

$$(u)_+ = \begin{cases} u \text{ si } u > 0 \\ 0 \text{ sinon} \end{cases}$$

  Elle effectue un seuillage doux ("shrinkage" en Anglais) de $|x_i| - a$. En effet, $\mathcal{T}_{\boldsymbol{a}}(\boldsymbol{x})_i$ vaut 0 lorsque $|x_i| \leq a$, $x_i - a$ si $x_i > a$ et $-(x_i - a)$ si $x_i < a$.