Partiel Analyse de Données

Documents autorisés :
1 feuille A4 Recto/Verso
Durée :
1h30 (+30 min tiers temps)

Questions de cours

1. (1pt) Expliquer ce qu'on entend par "classes linéairement séparables".
   On dit que les classes sont linéairement séparables lorsqu'il existe des hyperplans (droites dimension 2, plans en dimension 3, hyperplans en dimension supérieure) qui permettent de séparer les données des différentes classes.
2. Pourquoi impose-t-on la contrainte $y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 1$ dans le classifieur SVM ? Lorsque les classes sont linéairement séparables, cette contrainte permet d'assurer que tous les points d'une même classe sont situés du même coté de l'hyperplan séparateur d'équation $\boldsymbol{w}^T\boldsymbol{x}_i - b = 0$. De plus, les points situés sur l'hyperplan séparateur vérifient dans le cas de deux classes $\boldsymbol{w}^T\boldsymbol{x}_i - b = \pm 1$.
3. (1pt) Qu'appelle-t-on méthode du "leave-one out" ?
   Cette méthode consiste à partir d'un ensemble d'apprentissage $\Omega = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ à construire n classifieurs. Le $i$ème classifieur d_i est construit à partir de $\Omega\{\boldsymbol{x}_i\}$ et utilise le point $\boldsymbol{x}_i$ seul pour le test. la probabilité d'erreur du classifieur est obtenue en moyennant les erreurs obtenues pour les n classifieurs $d_1, ..., d_n$.
4. (1pt) Dans le classifieur "soft-margin" SVM chaque vecteur $\boldsymbol{x}_i$ de classe y_i doit vérifier la contrainte $y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 1 - \xi_i$. Quelle est l'utilité de la variable ξ_i ?
   Pour un problème à deux classes linéairement séparables, Le classifieur SVM cherche un hyperplan séparateur tel que les points d'une même classe sont situés du même coté de l'hyperplan. Lorsque les classes ne sont pas linéairement séparables, on peut autoriser certains points à être du mauvais coté de l'hyperplan. Pour autoriser ou pas un point $\boldsymbol{x}_i$ à être du mauvais coté de l'hyperplan, il suffit d'introduire une variable latente $\xi_i \geq 0$ pour chaque point $\boldsymbol{x}_i$ telle que $y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 1 - \xi_i$. Lorsque $\xi_i = 0$, le point $\boldsymbol{x}_i$ est du bon coté de l"hyperplan tandis que lorsque $\xi_i > 0$, le point $\boldsymbol{x}_i$ est du mauvais coté de l"hyperplan.
5. (1pt) Donner au moins deux inconvénients de l'algorithme K-means.
   Comme expliqué en cours, l'algorithme K- means peut poser problème dans les cas suivants :
   - Les données des différentes classes ont des densités différentes
   - Les données d'apprentissage sont mal représentées (certains classes ont peu d'exemples et certaines classes ont beaucoup d'exemples)

— Les classes formées par l'algorithme K-means dépendent de la manière dont l'algorithme est initialisé.

— ...

6. (1pt) Expliquer ce que sont les points "core", "border" et "noise" dans l'algorithme DBSCAN.
L'algorithme DBSCAN est un algorithme de classification non supervisée qui identifie trois types de points : les points du coeur du nuage de points ("core points") sont les points qui contiennent plus de MinPts voisins (paramètre que l'utilisateur doit choisir), les points du bord ("Border points") sont les points qui contiennent moins de MinPts voisins et qui sont connectés à des points du coeur du nuage. Les points qui contiennent moins de MinPts voisins et qui ne sont pas connectés à des points du coeur du nuage sont des points de bruit, i.e., des "outliers". Notons que le voisinage d'un point $\boldsymbol{x}$ est défini comme l'ensemble des points $\boldsymbol{y}$ qui vérifient $d(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon$) où ϵ est un paramètre fixé par l'utilisateur.

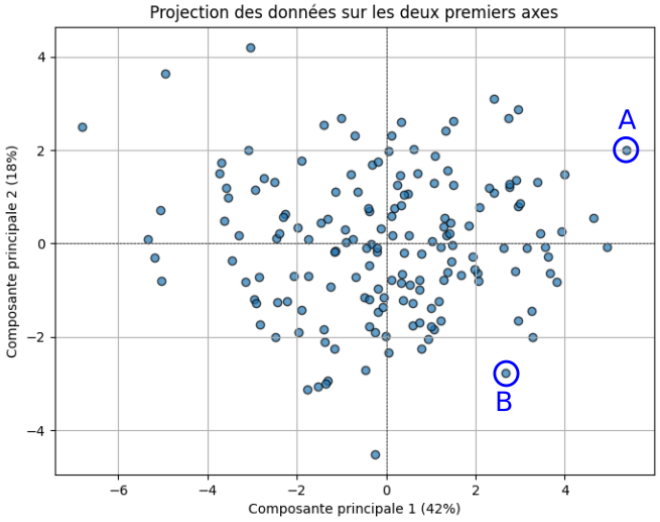
Exercice 1 : ACP

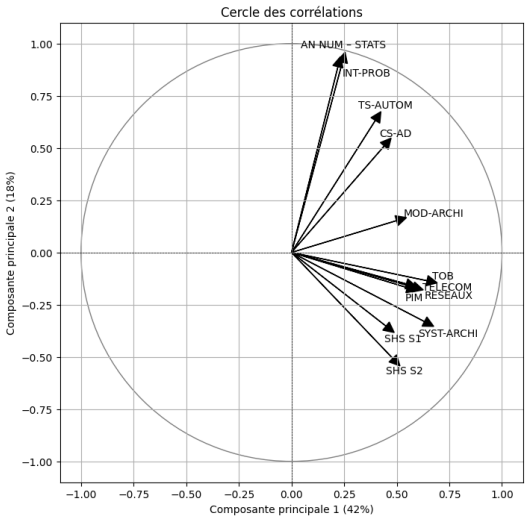
Pour cet exercice, on dispose d'un bordereau des notes obtenues par les 1SN pendant l'année 2022-2023. Chaque étudiant.e a obtenu 12 notes, une pour chacune des UEs suivantes : Sciences Humaines et Sociales Semestres 1 et 2, Architecture des ordinateurs et Système (`SYST-ARCHI`), Programmation Impérative (`PIM`), Technologies Objet (`TOB`), Telecommunications, Réseaux, Modélisation et Architecture (`MOD-ARCHI`), Calcul Scientifique et Analyse de Données (`CS-AD`), Traitement du Signal et Automatique (`TS-AUTOM`), Intégration et Probabilités (`INT-PROB`), et Analyse Numérique et Statistiques (`AN NUM-STATS`).

On se propose dans cet exercice de faire une Analyse en Composantes Principales de ce tableau de données.

1. (1pt) D'après vous, est-il nécessaire de centrer et/ou réduire ces données pour appliquer l'ACP ? Justifiez votre réponse.
   Il faut **toujours** centrer les données avant de faire une ACP. La réduction est utile lorsque les variables ont des ordres de grandeur et/ou des unités différentes, ce qui n'est pas le cas ici car les notes de chaque UE sont entre 0 et 20. Il n'est donc pas forcément nécessaire de réduire les données dans ce cas.
2. (1pt) Combien la matrice de variance-covariance de ces données compte-t-elle de valeurs propres ? Expliquez comment l'on peut construire la base d'un espace à deux dimensions dans lequel projeter les données.
   Le problème compte 12 variables, la matrice de variance-covariance est donc de taille 12×12, et il y a donc 12 valeurs propres (certaines peuvent être nulles). La base de l'ACP s'obtient en sélectionnant les vecteurs propres associés aux 2 plus grandes valeurs propres.

Voici une représentation des données projetées sur les 2 premiers axes ainsi que le cercle des corrélations :





3. (1pt) Entourez sur la figure le point correspondant, selon vous, à la projection de l'étudiant.e qui majore la promotion. Notez ce point `A`.

L'axe 1 est positivement corrélé à toutes les variables, autrement dit un.e étudiant.e qui aurait eu de bonnes notes à tous les examens aurait une grande valeur sur l'axe 1. On choisit donc le point le plus à droite de la figure.

4. (1pt) Entourez sur la figure un point correspondant à un.e étudiant.e qui serait excellent.e dans les matières informatiques (type `PIM` ou `TOB`), mais qui obtiendrait de mauvaises notes en mathématiques (`INT-PROB`, `AN NUM-STATS`). Notez ce point `B`.
L'axe 2 est positivement corrélé aux variables liées aux mathématiques, et négativement corrélé aux variables liées à l'informatique. Autrement dit un.e étudiant.e qui aurait eu de bonnes notes en informatique et de mauvaises notes en mathématiques devrait avoir une valeur fortement négative sur l'axe 2. L'axe 1 est également plus fortement corrélé aux variables informatique qu'aux variables mathématiques. Ainsi un.e étudiant.e qui réussit bien en informatique et moins en mathématiques devrait avoir une valeur légèrement positive sur l'axe 1 et fortement négative sur l'axe 2.

5. (1pt) D'après vous, que peut-on dire de la note d'`INT-PROB` pour un.e étudiant.e qui a eu une bonne note en `AN NUM-STATS` ?
L'angle entre les variables `INT-PROB` et `AN NUM-STATS` est très faible, ce qui signifie que les variables semblent fortement corrélées. En outre, les flêches représentant ces 2 variables sur le cercle des corrélations sont longues, ce qui signifient que les variables sont **bien représentées** sur cette projection. On peut donc en conclure qu'un.e étudiant.e qui a eu une bonne note en `AN NUM-STATS` a de très grandes chances d'avoir également eu une bonne note en `INT-PROB`.

6. (1pt) D'après vous, que peut-on dire de la note de `RESEAUX` pour un.e étudiant.e qui a eu une mauvaise note en `TELECOM` ?
L'angle entre les variables `RESEAUX` et `TELECOM` est très faible, ce qui signifie que les variables semblent fortement corrélées. Cependant, les flêches représentant ces 2 variables sur le cercle des corrélations sont assez courtes, ce qui signifient que les variables sont **mal représentées** sur cette projection. On ne peut donc pas conclure avec certitude que les 2 variables sont corrélées. Ainsi, on ne peut rien dire sur la note de `RESEAUX` d'un.e étudiant.e qui a eu une mauvaise note en `TELECOM`.

7. (1pt) D'après vous, que peut-on dire de la note de `SYST - ARCHI` pour un.e étudiant.e qui a eu une mauvaise note en `AN NUM-STATS` ?
L'angle entre les variables `SYST - ARCHI` et `AN NUM-STATS` est d'environ 90°, ce qui signifie que les variables semblent très faiblement corrélées. Ainsi, on ne peut rien dire sur la note de `SYST - ARCHI` d'un.e étudiant.e qui a eu une mauvaise note en `AN NUM-STATS`.

Exercice 2 : Moindres carrés

On cherche à ajuster une ellipse à des données expérimentales en utilisant la méthode des moindres carrés. Une ellipse peut être décrite par l'équation générale suivante :

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

En réalité, les coefficients sont définis à une constante près, car pour tout $\alpha \in \mathbb{R}$, on a

$$\alpha ax^2 + \alpha bxy + \alpha cy^2 + \alpha dx + \alpha ey + \alpha f = 0$$

On choisit donc de poser $f = -1$, et on s'intéresse maintenant à l'équation de l'ellipse :

$$ax^2 + bxy + cy^2 + dx + ey = 1$$

On cherche à déterminer les paramètres de cette ellipse. On dispose pour cela de n points P_i de coordonnées (x_i, y_i) dont on sait qu'ils sont au voisinage de l'ellipse.

1. (2 pts) Formulez le problème d'ajustement des paramètres de l'ellipse au sens des moindres carrés. Donnez la formulation matricielle en explicitant le vecteur de paramètres $\boldsymbol{\beta}$ et les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ associées.
   Le vecteur des paramètres $\boldsymbol{\beta}$ contient les paramètres de l'ellipse à estimer, i.e. $\boldsymbol{\beta} = [a, b, c, d, e]^T$
   Le problème d'estimation de $\boldsymbol{\beta}$ au sens des moindres carrés s'écrit donc :

   $$\min_{\boldsymbol{\beta} \in \mathbb{R}^6} \frac{1}{n} \sum_{i=1}^{n} \left(ax_i^2 + bx_iy_i + cy_i^2 + dx_i + ey_i - 1 \right)^2$$

   ce que l'on peut réécrire

   $$\min_{\beta \in \mathbb{R}^2} \frac{1}{2} ||\mathbf{A}\boldsymbol{\beta} - \mathbf{B}||^2$$

   en posant $\boldsymbol{\beta} = [a', b]^T$ et

   $$\boldsymbol{A} = \begin{bmatrix} x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 \\ x_2^2 & x_2y_2 & y_2^2 & x_2 & y_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_n^2 & x_ny_n & y_n^2 & x_n & y_n \end{bmatrix} \text{ et } \boldsymbol{B} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$$

2. (1 pt) Expliquez comment estimer les paramètres de l'ellipse à partir des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$.
   Si $\boldsymbol{A}^T\boldsymbol{A}$ est inversible, i.e. si n est suffisamment élevé pour que la matrice $\boldsymbol{A}$ soit de rang 5, alors

   $$\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}.$$

   où $(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ est la matrice pseudo-inverse de $\boldsymbol{A}$.

Exercice 3 : Classification Bayésienne (5 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. Montrer que la règle de classification associée à ce problème lorsque les deux classes sont équiprobables consiste à classifier x dans la classe ω_1 si $x > a$, où a est une constante dépendant de σ_1^2 et de σ_2^2.
   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 $\sigma_1^2 > \sigma_2^2$
   $$d^*(x) = \omega_1 \Leftrightarrow \frac{1}{\sigma_1^2} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \geq \frac{1}{\sigma_2^2} \exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$
   ou
   $$d^*(x) = \omega_1 \Leftrightarrow x^2 \geq a^2 = \frac{2(\sigma_1^2\sigma_2^2)}{\sigma_2^2 - \sigma_1^2} \ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$$
   En remarquant que $x > 0$, on obtient
   $$d^*(x) = \omega_1 \Leftrightarrow x > a$$
   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. Déterminer la probabilité d'erreur associée.
   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. Que devient la règle de classification de la première question lorsque $P(\omega_1) > P(\omega_2)$? Commenter le résultat obtenu.
   Lorsque $P(\omega_1) = 2P(\omega_2)$, le classifieur Bayésien affecte x à la classe ω_1 (ce que l'on notera $d^*(x) = \omega_1$) si
   $$f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2),$$
   soit
   $$d^*(x) = \omega_1 \Leftrightarrow \frac{P(\omega_1)}{\sigma_1^2} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \geq \frac{P(\omega_2}{\sigma_2^2} \exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$
   ou
   $$x^2 \geq b^2 = \frac{2(\sigma_1^2\sigma_2^2)}{\sigma_2^2 - \sigma_1^2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right)\right].$$
   Par rapport à la question 1, on voit que $b^2 < a^2$ car $P(\omega_1) > P(\omega_2)$ (et donc $\ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > 0$ et $\sigma_2^2 - \sigma_1^2 < 0$. Donc on accepte plus souvent la classe ω_1, ce qui est normal car elle est plus probable.
4. 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.
   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 à σ_1^2. Des calculs élémentaires permettent d'obtenir
   $$\widehat{\sigma}_1^2 = \frac{1}{2n}\sum_{i=1}^{n}x_i^2.$$