Unsupervised Learning - Monday 2 October 2023 - Duration : 60 min

No document, no phone, no computing machine.

Name : First name : Signature :

Part 1 :	Part 2 :	Total /20 :

Exercise 1 (Questions on Part 1, ≃ 8 pts)

1. When we estimate a covariance matrix Σ, why is it relevant to use a skrinkage method? Describe briefly how the shrinkage is applied to an ordinary estimate $\hat{\Sigma}$ of Σ.

   [Lorsqu'on estime une matrice de covariance Σ, pourquoi est-il pertinent d'utiliser une méthode de skrinkage? Décrire brièvement comment le shrinkage est appliqué à une estimation ordinaire $\hat{\Sigma}$ de Σ.]

sometimes $\hat{\Sigma}$ is not invertible for numerical reasons this happens also when there are more features than samples and this make it ill-conditioned so the eigen values estimation is bad.
so we use shrinkage estimator $\Sigma_s = (1-\alpha)\Sigma + \alpha \frac{\text{tr}(\Sigma)}{p} I_p$
that have a little bias but give us a well-conditioned matrix

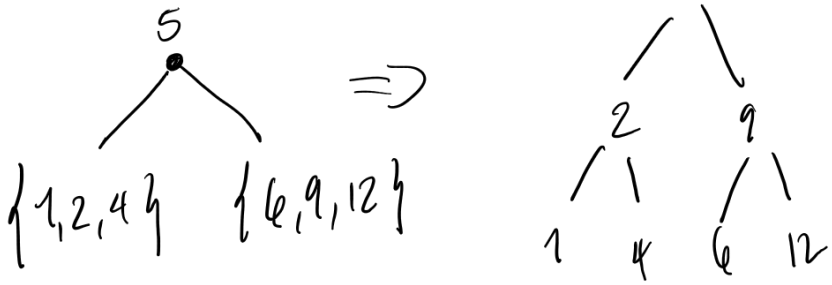
2. Let us consider the 6 numbers {1, 2, 4, 6, 9, 12}. Compute and draw the 1d-tree of this set of numbers (the 1d-tree must look like a tree). You will use the median for the splitting step.

   Reminder : The median is the middle value of a set of data. If there is an odd amount of numbers, the median value is the number that is in the middle, with the same amount of numbers below and above. If there is an even amount of numbers in the list, the median is the average of the two middle values. As an illustration, the median of $\{1,1,3,5,100\}$ is 3 and the median of $\{1,2,2,6,100,200\}$ is 4.

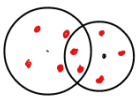
   [Considérons les 6 nombres {1, 2, 4, 6, 9, 12}. Calculez et dessinez le 1d-tree de cet ensemble de nombres (l'arbre 1d-tree doit vraiment ressembler à un arbre). Vous utiliserez la médiane pour l'étape de fractionnement.

   Rappel : La médiane est la valeur médiane d'un ensemble de données. S'il y a un nombre impair de nombres, la valeur médiane est le nombre qui se trouve au milieu, avec la même quantité de nombres en dessous et au-dessus. S'il y a un nombre pair de nombres dans la liste, la médiane est la moyenne des deux valeurs les plus au milieu. À titre d'illustration, la médiane de $\{1,1,3,5,100\}$ est de 3 et la médiane de $\{1,2,2,6,100,200\}$ est de 4.]

{1, 2, 4, 6, 9, 12} median = 5



3. Assume we run DBSCAN with $p_{\text{min}} = 6$ and $\epsilon = 0.1$ for a dataset and we obtain 4 clusters and 5% of the objects in the dataset are classified as outliers. Then we run DBSCAN with $p_{\text{min}} = 8$ and $\epsilon = 0.1$. How do expect the clustering results to change?
[Supposons que nous exécutons DBSCAN avec $p_{\text{min}} = 6$ et $\epsilon = 0.1$ pour un ensemble de données et que nous obtenions 4 clusters et que 5% des objets de l'ensemble de données sont classés comme valeurs aberrantes. Ensuite, nous exécutons DBSCAN avec $p_{\text{min}} = 8$ et $\epsilon = 0.1$. Comment pensez-vous que les résultats du clustering changeront?]



When we increase the requirement p_{min} to 8 what we'll produce is less core points and less core points and less border points so will be more outliers and also will be less clusters or depending on the structure equal or less clusters

Exercise 2 (Questions on Part 2, ≃ 12 pts)

1. Figure 1 shows data points in a coordinate system that shall be divided into clusters with the help of agglomerative clustering. Each square side measures one unit. The distance between data points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ can be computed with the Euclidean Distance. We will use single link similarity to cluster the data points

$$\text{sim}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \| x - y \|_2^2$$

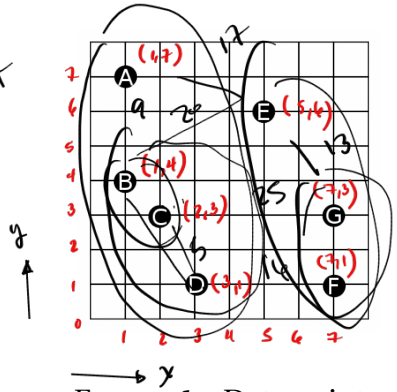
where C_i and C_j are two arbitraty clusters. Draw the clustering dendogram with well annotated axes, including numerical values.

[La Figure 1 montre des points de données dans un système de coordonnées qui doivent être divisés en clusters à l'aide d'un clustering agglomératif. La coté de chaque carré mesure 1 unité. La distance entre les points $x = (x_1, x_2)$ et $y = (y_1, y_2)$ peut être calculée avec la distance euclidienne. Nous utiliserons la similarité de lien unique pour regrouper les points

$$\text{sim}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \| x - y \|_2^2$$

où C_i et C_j sont deux clusters arbitraires. Dessinez le dendogramme du clustering avec des axes clairement annotés qui incluent des valeurs numériques.]

$D(A,E) = 4^2 + 1^2 = 16+1 = 17$



$D(B,E) = 4^2 + 2^2$
$16 + 4 = 20$

$D(G,F) = 3^2 + 2^2$
$9 + 4 = 13$

FIGURE 1 – Data points.

$D(B,C) = (1-2)^2 + (4-3)^2 = 2$

$D(A,B) = (3)^2 = 9$

$C_{BC} = \{B, C\}$

$\text{sim}\{C_{B,C}, D\} = (1)^2 + (2)^2 = 5$

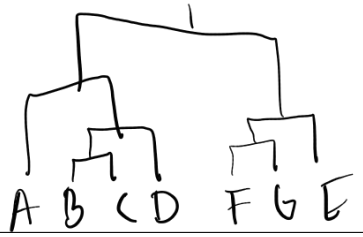
$D(G,F) = (0)^2 + (2)^2 = 4$

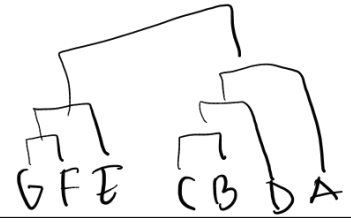
$C_{B,C,D} = \{\{B,C\}, D\}$

$C_{GF} = \{G, F\}$

$C_{B,C,D,A} = \{C_{B,C,D}, A\}$ $D(G,F) = (2)^2 + (3)^2$

$= 4 + 9 = 13$





2. We use the polynomial kernel defined by

$$K(x, y) = (1 + x^T y)$$

where $x = (x_1, x_2)$, $y = (y_1, y_2)$ and x^T is the transpose of x. Give the feature transformation $\phi(x)$ that generates the polynomial kernel. A detailed proof is required.
[Nous utilisons le noyau polynomial défini par

$$K(x, y) = (1 + x^T y)$$

où $x = (x_1, x_2)$, $y = (y_1, y_2)$ et x^T est la transposée de x. Donnez la transformation de fonctionnalités $\phi(x)$ qui génère le noyau polynomial. Une preuve détaillée est demandée.]

$$k(x,y) = 1 + \sum_{1}^{2} x_i y_i$$

$$k(x,y) = 1 + x_1 y_1 + x_2 y_2 = \langle \phi(x), \phi(y) \rangle$$

x_1, x_2

$$\phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} x_1, x_2, 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ 1 \end{bmatrix}$$

$$\phi(y) = \begin{bmatrix} \sqrt{c}\, y_1 \\ \sqrt{c}\, y_2 \\ \sqrt{a} \end{bmatrix}$$

$$= c x_1 y_1 + x_2 y_2 + 1$$

$$\phi(x)^T \phi(y)$$

3. We have computed a clustering for a given dataset. Give and describe briefly the three kinds of performance metrics that can be used to evaluate the clustering quality. You must precise the assumptions required by each kind of metrics.
[Nous avons calculé un clustering pour un ensemble de données. Donnez et décrivez brièvement les trois types de mesures de performances qui peuvent être utilisées pour évaluer la qualité du clustering. Vous devez préciser les hypothèses requises pour chaque type de métrique.]

4. Let the one-dimensional data-set given by $x_1 = 1$, $x_2 = 6$, $x_3 = 8$. The goal is to fit a Gaussian Mixture Model (GMM) with two components (C_1, C_2) onto the data. The initial model parameters are given by the means $\mu_1 = 1$, $\mu_2 = 4$, variances $\sigma_1^2 = \sigma_2^2 = 1$ and priors $\pi_1 = \pi_2 = 0.5$. The variances are not modified by the learning algorithm.

   (a) Draw the points on the horizontal axis and sketch the mean and probability density function of the initial Gaussians (vertical axis is the probability density) in Figure 2. Reminder : $1/\sqrt{2\pi} \approx 0.4$.

   (b) Compute the EM-Algorithm for Gaussian Mixture Models for the first iteration and report new priors and means in the answer box.

   (c) Sketch the Gaussian components after the first iteration into Figure 3.

   [Soit l'ensemble de données unidimensionnelles défini par $x_1 = 1$, $x_2 = 6$, $x_3 = 8$. Le but est d'ajuster un modèle de mélange de Gaussiennes (GMM) à deux composantes (C_1, C_2) sur les données. Les paramètres initiaux du modèle sont donnés par les moyennes $\mu_1 = 1$, $\mu_2 = 4$,

les variances $\sigma_1^2 = \sigma_2^2 = 1$ et les probabilités a priori $\pi_1 = \pi_2 = 0.5$. Les variances ne sont pas modifiées par l'algorithme d'apprentissage.

(a) Dessinez les points sur l'axe horizontal et esquissez les moyennes et les densités de probabilité des Gaussiennes initiales (l'axe vertical est la densité de probabilité) dans la Figure 2. Rappel : $1/\sqrt{2\pi} \approx 0.4$.

(b) Calculez l'algorithme EM pour les modèles de mélange gaussien pour la première itération et écrivez les nouvelles probabilités a priori et moyennes obtenues dans la boîte de réponse.

(c) Esquissez les composantes Gaussiennes après la première itération dans la Figure 3.]

$$r_{1,1} = \frac{0.5\, e^{(0)}}{0.5e^{(0)} + 0.5e^{-4.5}} = \frac{0.5}{0.5 + 0.5e^{4.5}} \approx 1$$

$$r_{2,1} = \frac{0.5e^{-12.5} \simeq 0}{0.5e^{-12.5} + 0.5e^{-2}} \approx 0$$

$$r_{3,1} = \frac{0.5e^{-24.5} \approx 0}{0.5e^{-24.5} + 0.5e^{-8}} \approx 0$$

$$T_1 = \frac{1}{r_{3,1} + r_{2,1} + r_{1,1}} = \frac{1}{1}$$

$$x = \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix}$$

$$\|x - M_1\|^2 = \begin{bmatrix} 0 \\ 5 \\ 7 \end{bmatrix}$$

$$0 + 25 + 49 = 74$$

$$\|x - M_2\|^2 = \begin{bmatrix} -3 \\ 2 \\ 4 \end{bmatrix}$$

$$9 + 4 + 16 = 29$$

$$M_1 = \frac{1}{T}\left[(r_{1,1} \cdot 1) + (r_{2,1} \cdot 6) + (r_{3,1} \cdot 8)\right]$$

$$\approx \frac{1}{1}\left(1 + (6(0) + 8(0)\right) \approx 1$$

$$r_{1,2} = \frac{0.5\, e^{-4.5}}{0.5e^{(0)} + 0.5e^{-4.5}} \approx 0$$

$$r_{2,2} \approx 1$$

$$r_{3,2} \approx 1$$

$$M_2 = \frac{1}{2}\left[6(1) + 6(1) + 8(1)\right]$$

$$\approx 0.5(14) \approx 7$$

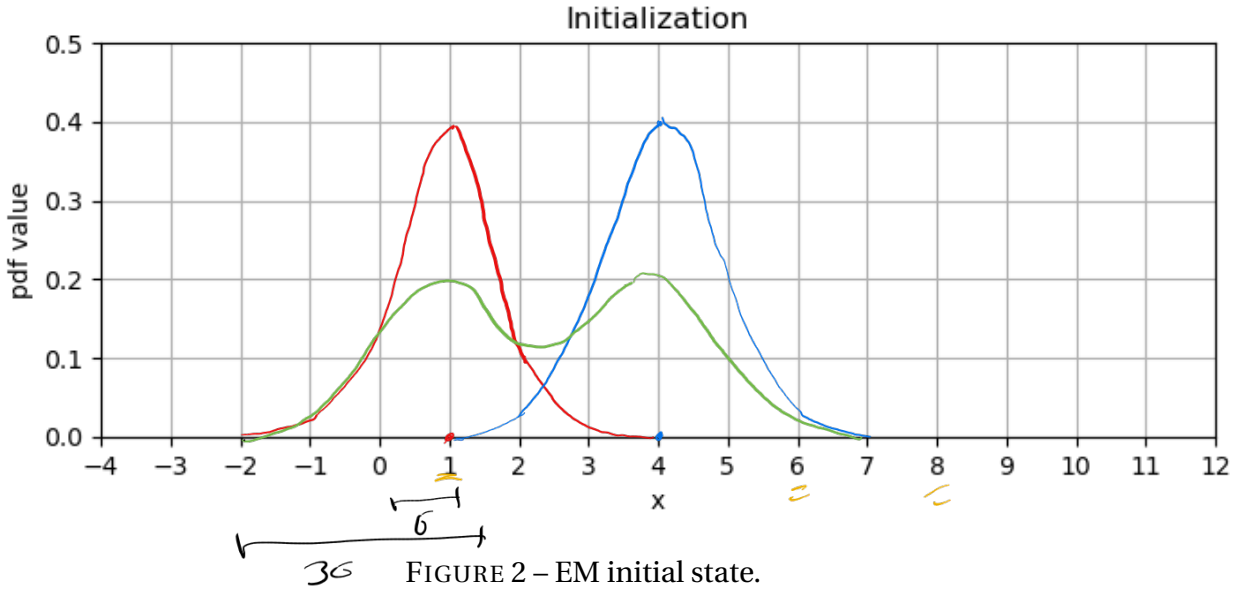


FIGURE 2 – EM initial state.

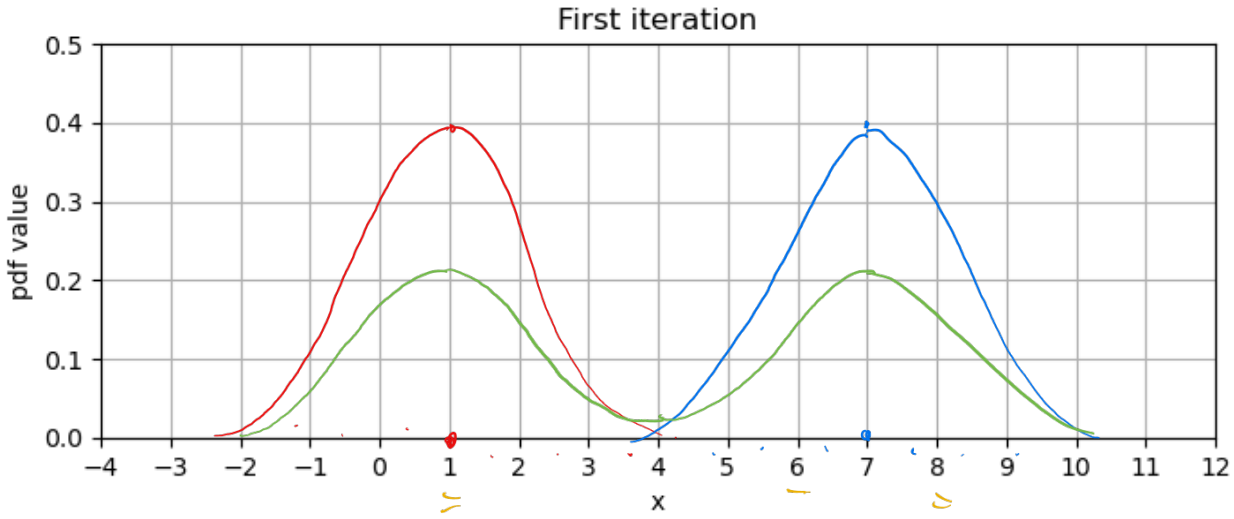


FIGURE 3 – EM first iteration.