# Examen CSAA EMINF1F1 - Session 1

## Problème : capteurs de montres connectées

Pour étudier l'efficacité de capteurs de température sur des montres connectées, on considère les différences entre les mesures de deux capteurs de température par rapport à la température réelle.
On obtient les données suivantes sur 4 individus.

| Individu | Capteur 1 | Capteur 2 |
|---|---|---|
| Ind. 1 | 0 | 2 |
| Ind. 2 | -2 | -1 |
| Ind. 3 | 1 | 0 |
| Ind. 4 | 1 | -1 |

### Partie 1 : Prétraitement - réduction de dimensions

1. Existe-t-il une dépendance entre ces deux capteurs ? Expliquer votre réponse.

Entre les deux capteurs, il existe une dépendance linéaire. En effet, on peut représenter les données par une droite. Pour vérifier cette dépendance, on peut calculer la matrice de corrélation comme suit :

corr(X) = 1/n * X^T * X

Avec X la matrice des données. On obtient en detaillant les calculs :

$$corr(X) = 1/4 * \begin{pmatrix} 0 & 2 \\ -2 & -1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix}^T * \begin{pmatrix} 0 & -2 & 1 & 1 \\ 2 & -1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 1 & 1 \\ 2 & -1 & 0 & -1 \end{pmatrix} * \begin{pmatrix} 0 & -2 & 1 & 1 \\ 2 & -1 & 0 & -1 \end{pmatrix}^T = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$$

On obtient donc une matrice de corrélation diagonale, ce qui confirme la dépendance linéaire entre les deux capteurs.

2. Calculer le premier axe principale de ces points.

Pour calculer le premier axe principale de ces points nous allons utiliser la méthode des valeurs propres. Pour cela, nous allons utiliser la matrice de corrélation calculée précédemment.

$$\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} * \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

On obtient donc $\lambda_1 = \lambda_2 = 5$.

3. Représenter les données et la premier axe principal.

4. Calculer les coefficients de projection 1D de l'ensemble des données sur le premier axe principal.

Pour calculer les coefficients de projection 1D de l'ensemble des données sur le premier axe principal, nous allons utiliser la formule suivante :

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T * \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 * x_1 \\ \lambda_2 * x_2 \end{pmatrix}$$

On obtient donc :

$$\begin{pmatrix} 0 \\ 2 \end{pmatrix}^T * \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 10 \end{pmatrix}$$

$$\begin{pmatrix} -2 \\ -1 \end{pmatrix}^T * \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} -10 \\ -5 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}^T * \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}^T * \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} 5 \\ -5 \end{pmatrix}$$

On obtient donc les coefficients de projection suivants :

| Individu | Capteur 1 | Capteur 2 |
|---|---|---|
| Ind. 1 | 0 | 10 |
| Ind. 2 | -10 | -5 |
| Ind. 3 | 5 | 0 |
| Ind. 4 | 5 | -5 |

### Partie 2 : Classification hiérarchique

On décide à partir de ces mesures de faire une étude sur les individus. On considère les composantes principales 1D (CP) suivantes des 4 individus :

| Personne | CP |
|---|---|
| Ind. 1 | 2 |
| Ind. 2 | -3 |
| Ind. 3 | 1 |
| Ind. 4 | 0 |

1. A partir de ces composantes principales 1D, calculer la matrice des distances euclidiennes entre les individus.

Pour calculer la matrice des distances euclidiennes entre les individus, nous allons utiliser la formule suivante :

$$d_{ij} = \sqrt{(x_i - x_j)^2}$$

On obtient donc :

$d_{12} = \sqrt{(2 - (-3))^2} = \sqrt{25} = 5$

$d_{13} = \sqrt{(2 - 1)^2} = \sqrt{5} = 2.236$

$d_{14} = \sqrt{(2 - 0)^2} = \sqrt{4} = 2$

$d_{23} = \sqrt{(-3 - 1)^2} = \sqrt{16} = 4$

$d_{24} = \sqrt{(-3 - 0)^2} = \sqrt{9} = 3$

$d_{34} = \sqrt{(1 - 0)^2} = \sqrt{1} = 1$

On obtient donc la matrice des distances euclidiennes suivante :

| Individu | Ind. 1 | Ind. 2 | Ind. 3 | Ind. 4 |
|---|---|---|---|---|
| Ind. 1 | 0 | 5 | 2.236 | 2 |
| Ind. 2 | 5 | 0 | 4 | 3 |
| Ind. 3 | 2.236 | 4 | 0 | 1 |
| Ind. 4 | 2 | 3 | 1 | 0 |

2. Réaliser une classification hiérarchique de ces données avec la distance du lien simple et celle du lien complet.

Rappel : Soient G et H deux groupes de données et d la distance euclidienne classique : Distance du lien simple/single linkage
$d_{SL}(G,H) = min_{i \in G, j \in H} d\_ij$

Distance du lien complet/complete linkage
$d_{CL}(G,H) = max_{i \in G, j \in H} d\_ij$

On obtient donc les matrices de distances suivantes :

Distance du lien simple :

| Individu | Ind. 1 | Ind. 2 | Ind. 3 | Ind. 4 |
|---|---|---|---|---|
| Ind. 1 | 0 | 5 | 2.236 | 2 |
| Ind. 2 | 5 | 0 | 4 | 3 |
| Ind. 3 | 2.236 | 4 | 0 | 1 |
| Ind. 4 | 2 | 3 | 1 | 0 |

Distance du lien complet :

| Individu | Ind. 1 | Ind. 2 | Ind. 3 | Ind. 4 |
|---|---|---|---|---|
| Ind. 1 | 0 | 5 | 4 | 3 |
| Ind. 2 | 5 | 0 | 4 | 3 |
| Ind. 3 | 4 | 4 | 0 | 1 |
| Ind. 4 | 3 | 3 | 1 | 0 |

3. Représenter les dendrogrammes. Peut-on obtenir la classification telle qu'énoncée ci-dessous ?

On obtient donc les dendrogrammes suivants :

| Personne | CP |
|---|---|
| Ind. 1 | $C_1$ |
| Ind. 2 | $C_2$ |
| Ind. 3 | $C_1$ |
| Ind. 4 | $C_1$ |

Si oui, pour quelle(s) distance(s) et quelle(s) valeur(s) ?

Oui, pour la distance du lien simple et la valeur 2.236.

4. Que proposeriez-vous comme autre méthode de classification (supervisée ou non) pour obtenir les classes $C_1$ et $C_2$ ? Expliquer votre réponse (choix de la méthode, paramètres à régler...)

On pourrait utiliser la méthode des k-means. On pourrait choisir k = 2, car on a deux classes. On pourrait choisir la distance euclidienne comme distance de référence.

## Exercice 2 : lois Gaussiennes et frontières de décision

On s'intéresse à un problème à deux classes $C_1$ et $C_2$ dans $IR^2$. On suppose que les observations de chaque classe suivent une loi Gaussienne avec pour paramètres respectifs : $m_1$ et $m_2$ (vecteurs moyennes) et $\Sigma_1$ et $\Sigma_2$ (matrices de covariances) pour les deux classes. On rappelle l'expression de la fonction de densité de probabilité conditionnelle multivariée pour une classe $C_i$ (ici, i vaut 1 ou 2) :

$$p(x|C_i) = \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp\left(-\frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i)\right)$$

Soit $g_i$ la fonction correspondant au logarithme népérien de la fonction $p(x|C_i)*P(C_i)$, où $P(C_i)$ est la probabilité a priori de la classe $C_i$.
Elle est utilisée lors de la phase de décision de la classification d'observations entre les deux classes :

$$g_i(x) = \ln(p(x|C_i) * P(C_i))$$

1. Développer l'expression de $g_i$ en fonction de x, $m_i$, $\Sigma_i$, $P(C_i)$ en distribuant le log et sans faire d'hypothèses spéciales sur les gaussiennes en jeu.

On a donc :

$g_i(x) = \ln(p(x|C_i) * P(C_i))$
$= \ln(p(x|C_i)) + \ln(P(C_i))$
$= \ln(\frac{1}{\sqrt{2\pi|\Sigma_i|}}) + \ln(\exp\left(-\frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i)\right)) + \ln(P(C_i))$
$= -\frac{1}{2}\ln(2\pi|\Sigma_i|) - \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) + \ln(P(C_i))$

2. On rappelle que la frontière entre les 2 classes C1 et C2 est donnée par :

$$g_2(x) - g_1(x) = 0$$

Donner l'expression de cette égalité la plus simplifiée possible lorsque les déterminants des matrices de covariances sont égaux : $|\Sigma_1|$ = $|\Sigma_2|$ (égalité des déterminants des deux matrices de covariances, mais attention on reste dans le cas général où ces matrices peuvent être différentes).

L'expression de cette égalité est donc :

$g_2(x) - g_1(x) = 0$
$-\frac{1}{2}\ln(2\pi|\Sigma_i|) - \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) + \ln(P(C_i)) - (-\frac{1}{2}\ln(2\pi|\Sigma_i|) - \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) + \ln(P(C_i))) = 0$

### Application numérique

Voici les moyennes, matrices de covariance et probabilités a priori des deux classes :

$$m_1 = \begin{pmatrix} -4 \\ -4 \end{pmatrix},$$
$$\Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix},$$
$$P(C_1) = 0.25,$$
$$m_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$
$$\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix},$$
$$P(C_2) = 0.75$$

3. On rappelle que les courbes de niveau ou courbes d'équidensité sont les courbes qui relient les points x pour lesquels $(x-m_i)' \Sigma_i^{-1} (x-m_i) = k$, où k est une constante positive. Simplifier le plus possible cette équation dans le cas de la classe $C_1$ puis de $C_2$ (deux équations à donner). Ces courbes sont-elles des cercles ou des ellipses ?

Les courbes sont donc des ellipses car les matrices de covariance ne sont pas des matrices diagonales comme nous pouvons le voir ci-dessous :

$$\Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}$$
$$\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$

4. On rappelle que différentes frontières de décision peuvent être obtenues en fonction des caractéristiques des observations : une droite, deux lignes qui se croisent, deux lignes parallèles, une parabole, une hyperbole, un cercle, une ellipse... A l'aide de votre réponse à la question 2, calculer l'équation de la frontière de décision (simplifier l'expression le plus possible). Quel type de frontière obtient-on ?

La frontière de décision est donc une ellipse. On a donc :

$g_2(x) - g_1(x) = 0$
$-\frac{1}{2}\ln(2\pi|\Sigma_i|) - \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) + \ln(P(C_i)) - (-\frac{1}{2}\ln(2\pi|\Sigma_i|) - \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) + \ln(P(C_i))) = 0$
$-\frac{1}{2}\ln(2\pi|\Sigma_i|) - \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) + \ln(P(C_i)) + \frac{1}{2}\ln(2\pi|\Sigma_i|) + \frac{1}{2}(x - m_i)'\Sigma_i^{-1}(x - m_i) - \ln(P(C_i)) = 0$
$-\frac{1}{2}\ln(2\pi|\Sigma_i|) + \ln(P(C_i)) + \frac{1}{2}\ln(2\pi|\Sigma_i|) - \ln(P(C_i)) = 0$
$-\frac{1}{2}\ln(2\pi|\Sigma_i|) + \frac{1}{2}\ln(2\pi|\Sigma_i|) = 0$
$\ln(2\pi|\Sigma_i|) = 0$
$\ln(2\pi|\Sigma_i|) = \ln(2\pi|\Sigma_i|)$
$2\pi|\Sigma_i| = 2\pi|\Sigma_i|$
$|\Sigma_i| = 1$