

CS2 - TD AFD Classification**Exercice 1**

Considérons le cas où la distribution de deux populations suit dans chaque groupe une loi normale $\mathcal{N}(\mu_i, \Sigma)$ avec $i = 1, 2$. On considère les probabilités a priori égales.

- Montrer que la procédure optimale classe x en 1 si

$$a^t x \geq a^t \times \frac{\mu_1 + \mu_2}{2}$$

avec $a = \Sigma^{-1}(\mu_1 - \mu_2) = \Sigma^{-1}\delta$ et $\delta = \mu_1 - \mu_2$,

en 2 sinon.

- Tracer la frontière pour $\mu_1 = -\mu_2 = (2, 1)^t$ et $\Sigma = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$ puis $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$.
- Montrer que l'hyperplan séparateur des deux zones de classification est orthogonal à $\delta = \mu_1 - \mu_2$ ssi δ est vecteur propre de Σ .
- Dans le cas $\mu_1 = (r, s)^t$ et $\mu_2 = (0, 0)^t$ et $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$, étudier la direction de a dans les cas (a) $s = 0$, (b) $s \neq 0$ et $\sigma_2^2 >> |s|$, (c) $s \neq 0$ et $\sigma_2^2 << |s|$.

Exercice 2

Considérons le cas où la distribution de deux populations suit dans chaque groupe une loi normale $\mathcal{N}(\mu_i, \Sigma)$ avec $i = 1, 2$. On considère les probabilités a priori égales. On adopte comme règle de classification la projection sur a en classant x en 1 si $a^t x \geq c$, c réel.

- On note $S(x) = a^t x$. Quelles sont les lois conditionnelles de $S(x)$ sachant 1 ou 2 ?
- On note p_{ij} la probabilité de classer en j un élément de i . Calculer p_{12} et p_{21} .
- Trouver c tel que les taux d'erreurs soient égaux.
- En déduire alors $p_{12} = p_{21}$ en fonction de Δ_p^2 .

Exercice 3 On considère toujours deux classes 1 et 2. On introduit maintenant le coût d'erreur, c_{ij} , associé au classement en j d'un élément de i . On suppose $c_{ii} = 0$. On imagine très bien un tel coût en médecine par exemple. Un patient considéré comme susceptible d'avoir une maladie nécessite des analyses complémentaires, serait-ce plus économique d'attendre des complications avant d'agir ? Suivant le patient ou la sécu, le raisonnement diffère...

- On note n_{ij} les effectifs de la matrice de confusion. Calculer le coût moyen de mauvais classement (misclassification).
- On note p_{ij} la probabilité de classer en j un élément de i . Montrer que le coût moyen de mauvais classement est : $EMC = c_{12}p_{12}p_1 + c_{21}p_{21}p_2$.
- Soit un élément x , la règle de Bayes avec coût consiste à choisir le groupe j qui minimise :

$$\sum_{k=1}^q c_{kj} P(k|x).$$

Trouver le coût moyen a posteriori d'un classement en 1 puis en 2.

- On note f_1 et f_2 les densités conditionnelles de x . En déduire que x est classé en 1 si :

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c_{21}p_2}{c_{12}p_1}.$$

Exercice 1 =

1. On affecte l'individu x au grp 1 si

$$\rho_1 f_2(x; \mu_1, \Sigma) > \rho_2 f_1(x; \mu_2, \Sigma) \Leftrightarrow f_1(x; \mu_1, \Sigma) > f_2(x; \mu_2, \Sigma)$$

Cours page 8 :

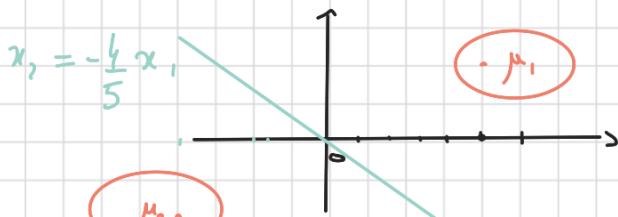
$$x^\top \Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2} (\mu_1 + \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) + \log\left(\frac{\rho_2}{\rho_1}\right) \text{ ou } \rho_1 = \rho_2$$

$$x^\top a \geq \frac{(\mu_1 + \mu_2)^\top - a}{2}$$

2. $a^\top x \geq a^\top \times \frac{\mu_1 + \mu_2}{2}$ pour trouver plan séparateur

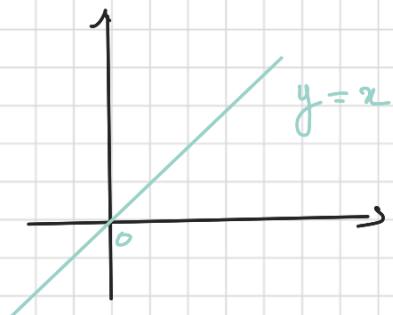
1^{er} cas :

$$\underline{a^\top x = 0} \quad \text{avec } a^\top = \begin{pmatrix} 4/5 \\ 1 \end{pmatrix}$$



2^e cas :

$$a^\top = \begin{pmatrix} 2 \\ -3 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \times 2 - 3 \times 2 \\ -3 \times 4 + 2 \times 5 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$



$$3. \quad a^\top x = a^\top \frac{(\mu_1 + \mu_2)}{2}$$

$$a^\top \left(x - \frac{\mu_1 + \mu_2}{2} \right) = 0$$

$$\mu_1 \xrightarrow{\mu_1 + \mu_2} \mu_2 \quad \text{et } x \in \mathcal{P}_{\text{mediateur}}$$

$$x - \frac{\mu_1 + \mu_2}{2} \perp a.$$

donc a colinéaire à δ .

Donc il faut δ colinéaire à $a = \Sigma^{-1} \delta = \lambda \delta$.

δ vect prop. de Σ^{-1}

$$L_1 - \mu_1 = (r, s)^T \quad \mu_2 = (0, 0)^T$$

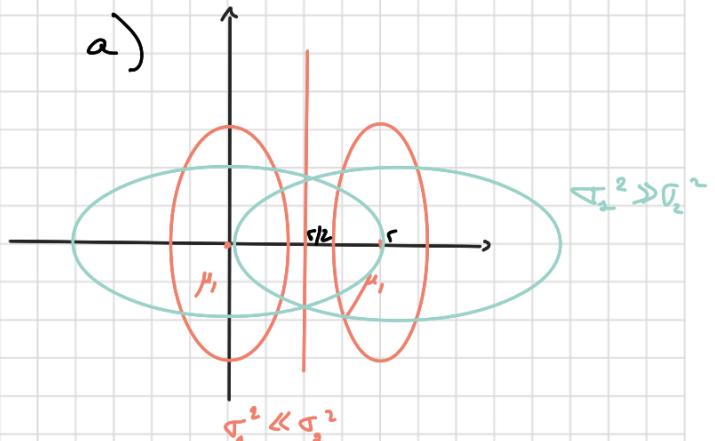
$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$a = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} r \\ s \end{bmatrix} = \begin{bmatrix} r/\sigma_1^2 \\ s/\sigma_2^2 \end{bmatrix}$$

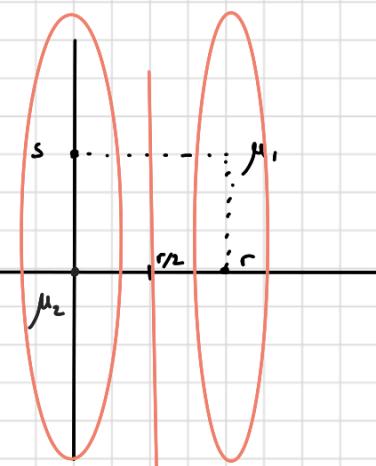
$$a^T x \geq a^T \begin{bmatrix} r/2 \\ 0 \end{bmatrix}$$

$$\frac{r}{\sigma_1^2} x_1 \geq \frac{r^2}{2\sigma_1^2}$$

$$x_1 \geq \frac{r}{2}$$



b)



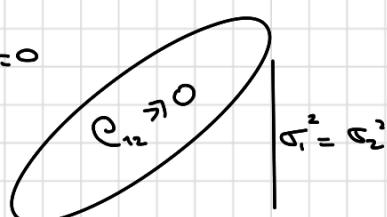
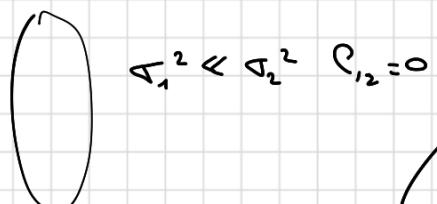
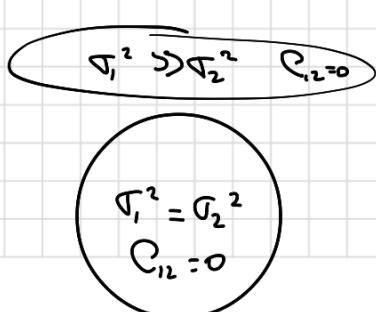
$$a = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} r \\ s \end{bmatrix} = \begin{bmatrix} r/\sigma_1^2 \\ s/\sigma_2^2 \end{bmatrix}$$

$$a^T x \geq a^T \begin{bmatrix} r/2 \\ s/2 \end{bmatrix}$$

$$\frac{r}{\sigma_1^2} x_1 + \underbrace{\frac{s}{\sigma_2^2} x_2}_{\approx 0} \geq \frac{r^2}{2\sigma_1^2} + \underbrace{\frac{s^2}{2\sigma_2^2}}_{\approx 0}$$

car $\sigma_2^2 \gg s$.

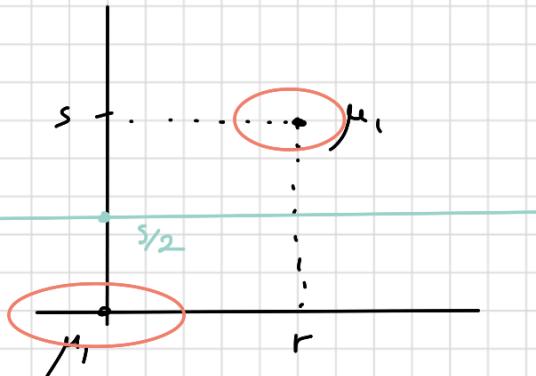
$$a) \quad \Sigma = \begin{bmatrix} \sigma_1^2 & C_{12} \\ C_{12} & \sigma_2^2 \end{bmatrix}$$



$$c) \frac{r}{\sigma_1^2} x_1 + \frac{s}{\sigma_2^2} x_2 \geq \frac{r^2}{2\sigma_1^2} + \frac{s^2}{2\sigma_2^2}$$

$\Rightarrow 0$ or $\sigma_2^2 \ll s$

$$\Leftrightarrow x_2 \geq \frac{s}{2}$$



Exercice 3 =

Exercice 3 On considère toujours deux classes 1 et 2. On introduit maintenant le coût d'erreur, c_{ij} , associé au classement en j d'un élément de i . On suppose $c_{ii} = 0$. On imagine très bien un tel coût en médecine par exemple. Un patient considéré comme susceptible d'avoir une maladie nécessite des analyses complémentaires, serait-ce plus économique d'attendre des complications avant d'agir ? Suivant le patient ou la sécu, le raisonnement diffère...

1. On note n_{ij} les effectifs de la matrice de confusion. Calculer le coût moyen de mauvais classement (misclassification).
2. On note p_{ij} la probabilité de classer en j un élément de i . Montrer que le coût moyen de mauvais classement est : $EMC = c_{12}p_{12}p_1 + c_{21}p_{21}p_2$.
3. Soit un élément x , la règle de Bayes avec coût consiste à choisir le groupe j qui minimise :

$$\sum_{k=1}^q c_{kj} P(k|x).$$

Trouver le coût moyen a posteriori d'un classement en 1 puis en 2.

4. On note f_1 et f_2 les densités conditionnelles de x . En déduire que x est classé en 1 si :

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c_{21}p_2}{c_{12}p_1}.$$

	y					
\hat{y}	n_{11}		n_{12}			

coût max = $\frac{c_{21}n_{12} + n_{21}c_{12}}{n}$

"
matrice de confusion

2) C : coût d'un patient

C	0	c_{12}	c_{21}
$P(C=c)$	$\frac{1}{P_1 P_{21}} - \frac{P_1 P_{12}}{P_2 P_{21}}$	$P_1 P_{12}$	$P_2 P_{21}$

$$\begin{aligned} E[C] &= 0 \times P(C=0) + c_{12} P(C=c_{12}) + c_{21} P(C=c_{21}) \\ &= c_{12} P_1 P_{12} + c_{21} P_2 P_{21} \end{aligned}$$

3- On classe x en 1 : le coût en 1 est

$$\text{en } 1 : \underbrace{c_{11} P(1|x)}_{=0} + c_{21} P(2|x)$$

$$\text{en } 2 : \underbrace{c_{22} P(2|x)}_{=0} + c_{12} P(1|x).$$

$$P(1|x) = \frac{P_1 f_1(x)}{P_1 f_1(x) + P_2 f_2(x)}$$

$$\text{en } 1 : \frac{c_{21} P_2 f_2(x)}{P_1 f_1(x) + P_2 f_2(x)}$$

$$\text{en } 2 : \frac{c_{12} P_1 f_1(x)}{P_1 f_1(x) + P_2 f_2(x)}$$

$$c_{21} P(2|x) \leq c_{12} P(1|x)$$

$$c_{21} P_2 f_2(x) \leq c_{12} P_1 f_1(x)$$

$$\frac{c_{21} P_2}{c_{12} P_1} \leq \frac{f_1(x)}{f_2(x)}$$

$$5- P_1 f_1(x) > P_2 f_2(x) \Leftrightarrow \log(P_1 f_1(x)) > \log(P_2 f_2(x))$$

$$\log(P_1 f_1(x)) = \log(P_1) - \frac{1}{2} \log(\Sigma_k) - \frac{(x - \mu_1)^T \Sigma^{-1}}{(x - \mu_1)}$$

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c_{21} p_2}{c_{12} p_1} \quad (=) \quad \frac{\frac{1}{(\Sigma_1)^{1/2} (\Sigma_2)^{1/2}} \exp\left(-\frac{(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2}\right)}{1}$$

$$\geq \frac{c_{21} p_2}{c_{12} p_1}$$

$$(\Rightarrow) \frac{(x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) - (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2} \geq \log\left(\frac{c_{21} p_2}{c_{12} p_1}\right)$$

$$(\Rightarrow) \frac{a^T x - a^T \frac{\mu_1 + \mu_2}{2}}{2} \geq \log\left(\frac{c_{21} p_2}{c_{12} p_1}\right)$$

$$a = \Sigma^{-1}(\mu_1 - \mu_2)$$

5. En déduire avec les définitions de l'exercice 1, la règle de classement dans le cas gaussien que x est classé en 1 si :

$$a^t x - a^t \times \frac{\mu_1 + \mu_2}{2} \geq \log \frac{c_{21}p_2}{c_{12}p_1}.$$

6. Que peut-on dire de la frontière ?
7. On pose $\mu_1 = (1, 1)^t$ et $\mu_2 = (1, -1)$ et $\Sigma = diag(\sigma_1^2, \sigma_2^2)$. Trouver l'équation de la frontière.
8. Que devient la frontière si $p_1 = p_2$ et $c_{12} = c_{21}$.
9. Pourquoi les composantes de x sont-elles indépendantes ? Comment est caractérisée géométriquement la distribution de x dans le cas d'indépendance des composantes ? Qu'en déduit on pour quant à la position de la frontière.
10. On choisit $p_1 = p_2$ et $c_{12} = 1$ et $c_{21} = 4$. Déterminer la nouvelle frontière.
11. On prend maintenant $\Sigma = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$, $p_1 = p_2$ et $c_{12} = c_{21}$, $\mu_1 = (1, 1)^t$ et $\mu_2 = (1, -1)^t$. Trouver une équation de la frontière. Trouver les axes principaux de la matrice Σ . Sur un graphique construire les centres de gravité, la frontière et l'ellipse de confiance des lois 1 et 2.

Exercice 4

L'objectif de cet exercice est de faire la classification de chiens selon leur race : race pure (Pure) ou race non pure (NPure) (métis). La classification se base sur deux variables Taille et Couleur. La variable Taille peut prendre deux valeurs grand ou petit (G, S respectivement) et la variable Couleur peut également prendre deux valeurs (U) pour uni et (M) pour multiples taches de couleur différente.

Individu	Taille	Couleur	Class
1	P	U	Pure
2	P	U	Pure
3	G	U	Pure
4	G	M	Pure
5	G	M	NPure
6	G	M	NPure
6	P	U	NPure
6	P	M	NPure

1. Calculez l'entropie de la variable Couleur.
2. Calculez l'entropie de la variable Taille.
3. La variable Taille constitue t-elle un bon prédicteur de la race ?

Justifiez votre réponse

Exercice 5 On veut apprendre un modèle permettant de déterminer si un client est intéressé à acheter un certain produit (Oui ou Non), en fonction de son sexe (Homme ou Femme), son âge (≤ 18 , $18 \leq 35$ ou ≥ 35), son état civil (Célibataire ou Marié), et son revenu (Faible, Moyen ou élevé). Soit l'échantillon suivant d'exemples d'entraînement :

ID	Sexe	Age	Etat civil	Revenu	Achat	
1	Homme	$18 \leq 35$	Marié	Moyen	Non	
2	Homme	≤ 18	Célibataire	Faible	Non	
3	Homme	≥ 35	Marié	élevé	Oui	
4	Femme	≤ 18	Célibataire	Moyen	Non	
5	Homme	$18 \leq 35$	Célibataire	Moyen	Non	
6	Femme	$18 \leq 35$	Célibataire	élevé	Oui	
7	Femme	$18 \leq 35$	Marié	Faible	Non	Construisez l'arbre de décision résultant de
8	Homme	$18 \leq 35$	Marié ?	Elevé	Oui	
9	Homme	≥ 35	Célibataire	Faible	Oui	
10	Femme	≤ 18	Célibataire	Moyen	Non	
11	Femme	≥ 35	Célibataire	Moyen	Oui	
12	Femme	≥ 35	Marié	élevé	Oui	
13	Homme	$18 \leq 35$	Célibataire	Faible	Non	
14	Femme	$18 \leq 35$	Marié	Moyen	Oui	

ces exemples en utilisant l'algorithme ID3. On suppose qu'on arrête la subdivision uniquement lorsque les nœuds sont purs (entropie de 0). Exprimez la classe des exemples positifs sous la forme d'un prédictat logique.

Exercice de mise en œuvre des arbres
de décision.