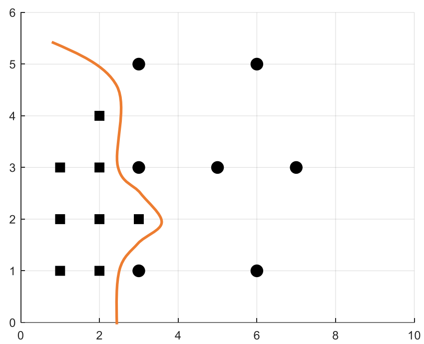
Nom	Prénom

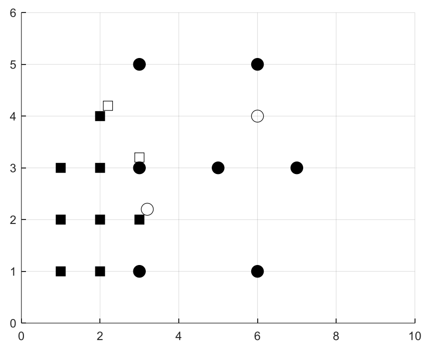
Exercice KPPV – toutes les questions sont indépendantes

Considérons un problème de classification à deux classes. On donne les moyennes et matrices de covariance de chacune des classes : $\mu_1 = \begin{bmatrix}1.75\\2.25\end{bmatrix}$, $\mu_2 = \begin{bmatrix}4.7\\3.0\end{bmatrix}$, $C_1 = \begin{bmatrix}0.7^2 & 0.07\\0.07 & 1^2\end{bmatrix}$, $C_2 = \begin{bmatrix}1.7^2 & 0\\0 & 1.6^2\end{bmatrix}$.

1. Les deux classes sont représentées ci-dessous. Tracer les frontières de décision obtenues avec la méthode du 1PPV. (1pt)



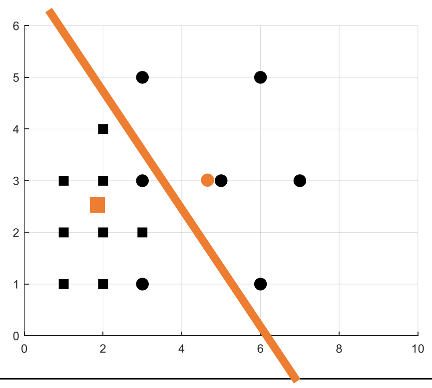
2. Donner la matrice de confusion obtenue en classant les exemples (dessinés en blanc) avec l'algorithme du 1ppv. Quel est le taux de reconnaissance ? (1pt)



	□	o
□	1	1
o	1	1

Taux de reconnaissance : 50%

3. On décide maintenant d'utiliser l'algorithme nearest mean et la distance euclidienne. Tracer la nouvelle frontière de décision. Quel sera le gain en temps de calcul pour classer un nouvel exemple par rapport au 1PPV? (1pt)

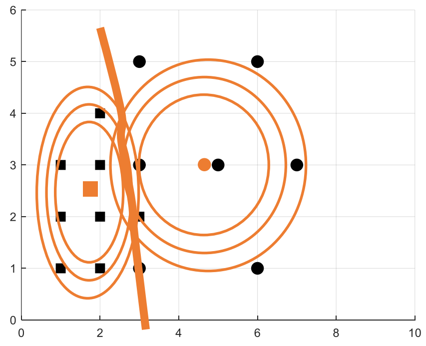


Avec le 1 ppv, il faut calculer la distance entre l'exemple inconnu et les 15 exemples de la base de référence

Avec le nearest mean, il faut calculer la distance entre l'exemple inconnu et les 2 centres

⇨ Gain de 7.5 fois plus rapide

4. Tracer grossièrement la frontière obtenue avec l'algorithme nearest-mean et la distance de Mahalanobis. Justifier votre raisonnement. (1pt)

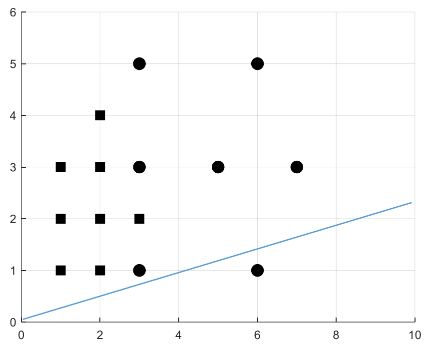


Le nuage des points représentés par des carrés a une forme d'ellipse tandis que celui représenté par des ronds à plutôt une forme ronde, plus étendue en x

5. Afin de gagner en temps de calcul, on décide de réaliser une ACP pour ne conserver qu'une dimension. Les vecteurs propres et valeurs propres associées sont données par :

$$v_1 = \begin{bmatrix} 0.27 \\ -0.9 \end{bmatrix} \quad v_2 = \begin{bmatrix} -0.9 \\ -0.2 \end{bmatrix} \; \lambda_1 = 1.6 \quad \lambda_2 = 4$$

Représenter l'axe de projection ci-dessous. Quelle sera la nouvelle coordonnée du point $\boldsymbol{x} = \begin{bmatrix} 1 \\ 6 \end{bmatrix}$? (1pt)



L'axe de projection correspond à la plus grand valeur propre et donc v_2

$$\begin{bmatrix} 1 & 6 \end{bmatrix} \begin{bmatrix} -0.9 \\ -0.2 \end{bmatrix} = -2.1$$

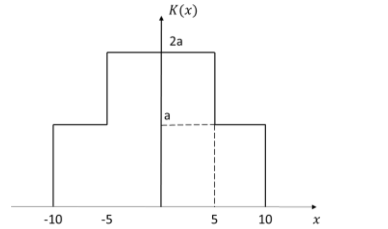
Exercice classification bayésienne

On souhaite reconnaitre deux espèces de champignons à partir d'une mesure m réalisée suite à des prélèvements. On dispose de 10 champignons de chaque espèce dont les prélèvements sont donnés ci-dessous :

C_1	244	238	277	280	219	249	245	265	271	275
C_2	278	318	316	266	262	300	346	284	309	272

Afin d'utiliser une classification bayésienne, on a besoin d'estimer la vraisemblance des observations $p(m/C_1)$ et $p(m/C_2)$.

1. Dans un premier temps, on estime ces vraisemblances avec des histogrammes dont l'origine est située en 0 et le pas de discrétisation vaut 10. Que valent $p(m/C_1)$ et $p(m/C_2)$ pour m = 264 ? Justifier votre réponse. (1pt)
   Ce sera e nombre de points entre 260 et 270 divisé par le nombre de points de la classe.
   $p(m/C1\,) = 1/10$ et $p(m/C2\,) = 2/10$
2. On utilise maintenant une estimation par noyau en utilisant le noyau $K(x)$. Que valent $p(m/C_1)$ et $p(m/C_2)$ pour m = 264 et a = 1/40 ? Justifier votre réponse. (1.5pt)



Les points situé à une distance inférieure à 5 voteront 2a et ceux situés à une distance entre 5 et 10 voteront a. Les autres voteront 0.
Pour C1, 265->2a et 271->a ➔ $p(m/C1\,) = 3a/10$
Pour C2, 266->2a, 262->2a 272 ->a ➔ $p(m/C2\,) = 5a/10$

3. Réaliser la classification du champignon représenté par m = 264 en utilisant l'estimation par noyau et en sachant que P(C_1)=0.25. (1pt)

On utilise la règle de Bayes :

P(C1/m)=$\frac{\mathrm{P(m/C1)*P(C1)}}{\mathrm{cte}} = \frac{3a*0.25}{10*cte} = 0.75\frac{a}{10*cte}$

P(C2/m)=$\frac{\mathrm{P(m/C2)*P(C2)}}{\mathrm{c}} = \frac{5a*0.75}{10*cte} = 3.75\frac{a}{10*cte}$ ➔Le champignon est classé C2

4. Dans un dernier temps, on réalise la classification par seuillage avec un seuil de 270.5 : tous les champignons avec un prélèvement inférieur à 270.5 sont classés C_1. Les autres sont classés C_2. Réalisés la classification des exemples de test ci-dessous. Donner la matrice de confusion et en déduire la sensibilité et la spécificité. (1.5pt)

C_1	275	226	251	270	289	296	255	214	215	226
C_2	334	275	331	274	343	285	270	275	312	297

3 exemples de la classe C1 seront mal classé et 1 seul de la classe C2.
La matrice de confusion est :

$$\begin{array}{ccc} & C1 & C2 \\ C1 & 7 & 3 \\ C2 & 1 & 9 \end{array}$$

Si on considère la classe C1 comme les exemples positifs, sensibilité = $\frac{7}{10}$ et spécificité = $\frac{9}{10}$, sinon, c'est l'inverse