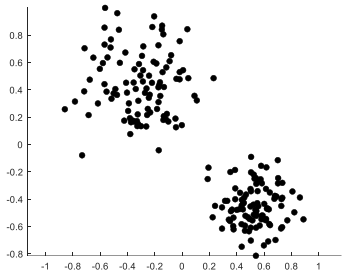
Nom:	
Prénom:	

Exercice 1 – 6 pts

On considère les données suivantes, **centrées**, dont on souhaite réduire la dimension.



1. **(1pts)** Quelle matrice de covariance correspond à ce nuage de points ? Justifier votre réponse.

$$C_1 = \begin{pmatrix} 0.2082 & -0.1864 \\ -0.1864 & 0.2370 \end{pmatrix} \quad C_2 = \begin{pmatrix} 0.2082 & 0 \\ 0 & 0.2370 \end{pmatrix} \quad C_3 = \begin{pmatrix} 0.2082 & 0.1864 \\ 0.1864 & 0.2370 \end{pmatrix}$$

C_1

2. **(1pts)** Quelle est la variance portée par le premier axe (variance des points projetés sur le premier axe) ?

0.2082 (Première case de la matrice de covariance)

3. **(1pts)** On donne les deux vecteurs propres associés à la matrice de covariance des points :

$$v_1 = \begin{pmatrix} -0.7338 \\ -0.6794 \end{pmatrix} \quad v_2 = \begin{pmatrix} -0.6794 \\ 0.7338 \end{pmatrix}$$

Lequel devra être utilisé pour projeter les points si on souhaite réduire la dimension des données par ACP ? Justifier votre réponse.

Le second, axe principal du nuage de points sur la figure

4. **(1pts)** Comment se transformera le point de coordonnées $\begin{pmatrix} -0.4238 \\ 0.1738 \end{pmatrix}$ dans le nouveau système de représentation ?

$$\begin{pmatrix} -0.4238 \\ 0.1738 \end{pmatrix} (-0.6794 \quad 0.7338) = 0.41.5$$

$$\begin{pmatrix} -0.4238 \\ 0.1738 \end{pmatrix} (-0.7338 \quad -0.6794) = 0.1929$$

5. **(2pts)** Quel sera la variance portée par cet axe dans le nouveau système de représentation ? Conclusion.

La variance des points projetés sur $\boldsymbol{v_2}$ est définie par :

$$\sigma = \frac{1}{N}\sum_{i=1}^{N} {x_i}'^2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}'^T {x_i}' = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v_2}'^T \boldsymbol{x_i} \boldsymbol{x_i}^T \boldsymbol{v_2}' = \boldsymbol{v_2}'^T \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i}\boldsymbol{x_i}^T\right) \boldsymbol{v_2'}$$
$$= \boldsymbol{v_2}'^T C_1 \boldsymbol{v_2}'$$

$$\sigma = \begin{pmatrix} -0.6794 & 0.7338 \end{pmatrix} \begin{pmatrix} 0.2082 & -0.1864 \\ -0.1864 & 0.2370 \end{pmatrix} \begin{pmatrix} -0.6794 \\ 0.7338 \end{pmatrix} = 0.4096$$

Conclusion : elle est bien plus grande que la variance des points sur le premier axe dans le repres initial

Exercice 2 – 3pts

On considère deux classifieurs dont on estime les performances sur une petite base de données. Pour cela, on utilise la K-fold validation avec K=4.
Le taux de reconnaissance de chaque classifieur sur chacun des « folds » est donné par :

	Fold1	Fold 2	Fold 3	Fold 4
Classifieur 1	74	75	75	74
Classifieur 2	95	45	70	92

1. **(1.5pts)** Donner le taux de reconnaissance obtenu par chaque classifieur

Classifieur 1 : 74.5
Classifieur 2 : 75.5

2. **(1.5pts)** Qu'est ce qui différencie chaque classifieur ? Lequel recommanderiez-vous à un industriel ?

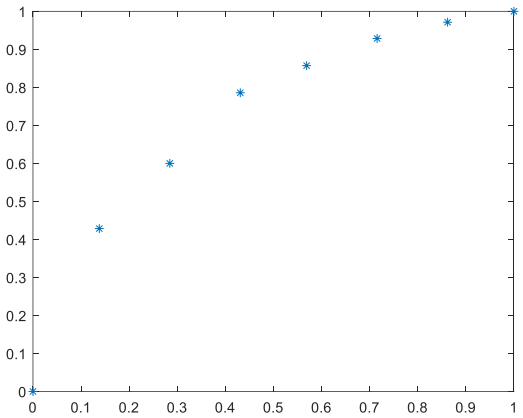
Le second classifieur a une très grande variance. Même si son taux de reconnaissance est un peu plus grand que le premier, on lui préférera le second qui est beaucoup plus stable.

Exercice 3 – 5pts

Des médecins mettent à notre disposition un test pour détecter le virus du covid. Un réglage permet de régler la sensibilité du test. On a ainsi, en fonction du réglage, pour 70 personnes réellement malades et 580 saines :

	Réglage1	Réglage2	Réglage3	Réglage4	Réglage5	Réglage6	Réglage7	Réglage8
TP	0	30	42	55	60	65	68	70
FP	0	80	165	250	330	415	500	580
Sp	1	0.86	0.72	0.57	0.43	0.28	0.14	0
Se	0	0.43	0.6	0.79	0.86	0.93	0.97	1
TR	89	81	70	59	48	35	23	11
TR1	50	64	66	68	64	61	55	50

1. **(2pts)** Déterminer, pour chaque réglage, la sensibilité (Se) et la spécificité (Sp) du test et tracer la courbe ROC correspondant



2. **(1.5pts)** Déterminer l'expression analytique du taux de bonne reconnaissance TR en fonction de Se, Sp, du nombre de positifs et de négatifs. Estimer ce taux TR pour chacun des réglages (compléter le tableau). Conclusion sur le meilleur réglage en fonction de TR. Conclusion.

Se=TP/70
Sp=TN/580
TR = (TP+TN) / (70+580)*100 = (Se*70+Sp*580)/(70+580)*100
Le premier réglage amène au meilleur taux de reconnaissance car les classes sont très déséquilibrées. On aura tendance à dire que tout le monde est sain.

3. **(1.5pts)** Déterminer l'expression analytique du taux de reconnaissance si on tient compte du déséquilibre des classes (TR1) et ré-estimer ce taux pour chacun des réglages. Conclusion sur le réglage le plus adapté.

TR= (TP/70+TN/580) / 2*100 = (Se+Sp)/(2)*100
Le meilleur réglage est maintenant le 4, ce qui correspond bien à la courbe ROC

Exercice 4 – 5pts

On dispose d'un ensemble de 10 points dont on souhaite estimer la densité de probabilité.

$$x = \{10, 12,12,13,15,15,16,18,18,20\}$$

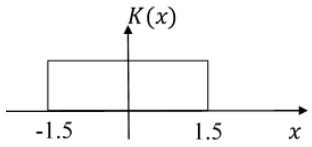
1. **(1pt)** Dans un premier temps, on utilise un histogramme, avec pour origine -0.5 et des cases de largeur 1. Que vaut $\hat{p}(15)$?

$\hat{p}(15)$ =2/10

2. **(1pt)** Même question avec des cases de largeur 2.

$\hat{p}(15)$ =2/10

3. **(1.5pts)** On utilise maintenant une fenêtre de Parzen. Que vaut $\hat{p}(15)$?



La hauteur de K(x) vaut 1/3.
$\hat{p}(15)$ =1/10*(3*1/3)=1/10

4. **(1.5pts)** Que vaut $\hat{p}(15)$ si on modélise les données par une gaussienne ?

On aura
m=14.9
sigma=3.01
`p=[1/(sqrt(2*pi)*s)*exp(-(15-m)^2/(2*s*s))]=0.13`