Nom, prénom : ______________________ **Signature :** ______________

Répondre sur ce document uniquement. La qualité de la présentation sera prise en compte dans la notation. Aucune copie supplémentaire ne sera acceptée. Pour les questions à choix multiples, le nombre de réponses correctes peut varier de 0 au nombre maximal de réponses proposées. Le barème indiqué est seulement indicatif.

1. 4 points Dans quels cas peut-on s'attendre à ce qu'une méthode d'apprentissage flexible (comportant beaucoup de paramètres) ait de meilleures performances qu'une méthode simple avec peu de paramètres :
   - ☐ Le nombre N d'exemples d'apprentissage est très grand, et le nombre p de prédicteurs est petit.
   - ☐ Le nombre p de prédicteurs est très grand, et le nombre N d'exemples d'apprentissage est petit.
   - ☐ La relation entre la variables à expliquer Y et les prédicteurs X_j est fortement non linéaire.
   - ☐ La variance des erreurs $\sigma^2 = \mathrm{Var}(\epsilon)$ est très grande.
2. 4 points Représenter l'allure typique du biais, de la variance et de l'erreur de test, en fonction du nombre de paramètres d'une méthode d'apprentissage. (Faire trois courbes sur le même graphique).

3. 2 points On considère un problème de régression linéaire avec une variable à expliquer Y et deux variables explicatives X_1 et X_2. On suppose que X_1 est une variable quantitative, et que X_2 est une variable qualitative à 3 modalités A, B et C. Ecrivez un modèle de régression linéaire pour ce problème.

   ....

   ....

   ....

   ....

   ....

   ....

4. On a obtenu les résultats suivants en appliquant la régression linéaire à un ensemble de données :

```
Call:
lm(formula = y ~ x1 + x2 + x3)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7597102  0.1807646   4.203 5.91e-05 ***
x1           0.7051162  0.1915360   3.681 0.000383 ***
x2           0.0102298  0.0002026  50.504  < 2e-16 ***
x3          -0.0131784  0.1738365  -0.076 0.939729
---
Residual standard error: 0.5429 on 96 degrees of freedom
Multiple R-squared:  0.9646,Adjusted R-squared:  0.9635
F-statistic: 873.1 on 3 and 96 DF,  p-value: < 2.2e-16
```

(a) $\boxed{\text{1 point}}$ Donnez un intervalle de confiance approché à 95% sur le coefficient de la variable X_2.

....

(b) $\boxed{\text{4 points}}$ Cocher la ou les bonnes réponses :

- ☐ La variable X_1 a plus d'influence sur Y que la variable X_2.
- ☐ Les coefficients des variables X_1 et X_2 sont significativement non nuls.
- ☐ Une variation de X_3 a très peu d'influence sur la variable Y.
- ☐ La suppression de X_3 conduirait à un meilleur modèle, donc à une augmentation du R^2.

5. $\boxed{\text{6 points}}$ On considère un problème de classification à 2 classes, avec 5 prédicteurs. Combien y a-t-il de paramètres à estimer dans les méthodes suivantes :

(a) Régression logistique.

(a) ________________

(b) Analyse discriminante linéaire.

(b) ________________

(c) Analyse discriminante quadratique.

(c) ________________

(d) Classifieur bayésien naïf (*naive Bayes*)

(d) ________________

(e) Règle des k plus proches voisins

(e) ________________

(f) Réseau de neurones avec 3 neurones cachés.

(f) ________________

6. $\boxed{\text{6 points}}$ La régression logistique

- ☐ est basée sur la méthode du maximum de vraisemblance ;
- ☐ nécessite un algorithme de programmation quadratique ;
- ☐ suppose que les classes sont linéairement séparables ;
- ☐ suppose que les classes sont gaussiennes ;
- ☐ estime les probabilités a priori des classes ;
- ☐ produit des résultats facilement interprétables.

7. $\boxed{\text{7 points}}$ Cochez la ou les bonnes réponses :

- ☐ Les méthodes de sélection ascendante et descendante (*forward/backward selection*) permettent de sélectionner un ensemble optimal de prédicteurs.
- ☐ En régression linéaire, la méthode de sélection ascendante n'est pas applicable lorsque le nombre p de prédicteurs est supérieur au nombre n d'exemples.
- ☐ La méthode de sélection descendante nécessite d'évaluer $1 + p(p + 1)/2$ modèles.
- ☐ Le R^2 ajusté croît de manière monotone avec le nombre de prédicteurs.
- ☐ La régression ridge permet de sélectionner les variables dans un modèle de régression.
- ☐ La pénalisation lasso permet de sélectionner les variables dans un modèle de régression.
- ☐ La pénalisation lasso permet d'améliorer l'erreur de prédiction lorsque le nombre d'exemples d'apprentissage est petit relativement au nombre de prédicteurs.

8. $\boxed{\text{3 points}}$ L'analyse factorielle discriminante

- ☐ Maximize le rapport de la variance inter-classe à la variance totale.
- ☐ Maximize le rapport de la variance inter-classe à la variance intra-classe.
- ☐ Permet d'extraire au plus $\max(N, K - 1)$ nouvelles variables, K étant le nombre de classes et N le nombre de vecteurs d'apprentissage.

9. $\boxed{\text{5 points}}$ Dans un problème de classification à deux classes, on considère deux classifieurs C_1 et C_2. On suppose que C_1 a de meilleures performances que C_2. Tracer l'allure des courbes COR des deux classifieurs. Expliquez la signification des axes.

10. $\boxed{\text{4 points}}$ La validation croisée
    - ☐ est utilisée pour choisir un modèle ayant la plus petite erreur de prédiction ;
    - ☐ donne une estimation sans biais de l'erreur de prédiction ;
    - ☐ généralise la méthode *leave-one-out* ;
    - ☐ a une plus grande variance, mais un biais plus faible, lorsque le nombre de blocs est plus grand.
11. $\boxed{\text{5 points}}$ Cochez la ou les bonnes réponses :
    - ☐ Une spline cubique avec un seul noeud a 5 degrés de libertés.
    - ☐ Une spline cubique est une fonction partout dérivable jusqu'à l'ordre 3.
    - ☐ Les splines de lissage ont deux hyperparamètres : le nombre de noeuds et le paramètre λ de régularisation.
    - ☐ Les modèles additifs généralisés peuvent représenter les interactions entre prédicteurs.
    - ☐ Les modèles additifs généralisés peuvent représenter des fonctions non linéaires.
12. $\boxed{\text{5 points}}$ L'algorithme EM :
    - ☐ Converge vers un minimum local de la fonction de vraisemblance.
    - ☐ Converge vers un maximum global de la fonction de vraisemblance.
    - ☐ Augmente à chaque étape la vraisemblance.
    - ☐ Est basé sur la minimisation à chaque étape d'une fonction majorante.
    - ☐ Est basé sur la maximisation à chaque étape d'une fonction minorante.
13. $\boxed{\text{6 points}}$ Cochez la ou les bonnes réponses :
    - ☐ L'apprentissage des réseaux de neurones nécessite de résoudre un problème d'optimisation non linéaire sous contraintes linéaires.
    - ☐ L'algorithme de rétropropagation du gradient consiste à propager l'erreur de la couche de sortie vers la couche d'entrée.
    - ☐ La méthode de descente de gradient converge vers un minimum global de l'erreur.
    - ☐ La régression logistique correspond à un réseau de neurones sans couche cachée.

- ☐ Dans l'apprentissage en ligne, les poids sont mis à jour à chaque présentation d'un nouvel exemple.
- ☐ L'avantage des réseaux de neurones est qu'ils sont facilement interprétables, car ils modélisent le fonctionnement des neurones biologiques.

14. $\boxed{\text{5 points}}$ Expliquez le principe de la méthode "Mixture of Regressions". Ecrivez le modèle en explicitant toutes les notations.

15. $\boxed{\text{5 points}}$ Représentez un réseau de neurones avec trois entrées, deux neurones cachés et une sortie linéaire. Ecrivez les équations de propagation dans le réseau. (Explicitez les notations sur le graphique).

16. $\boxed{\text{4 points}}$ Cochez la ou les bonnes réponses :
    - ☐ L'apprentissage des SVM consiste à minimiser la marge.
    - ☐ L'apprentissage des SVM nécessite de résoudre un problème d'optimisation linéaire.
    - ☐ Avec les SVM, la fonction de décision s'exprime en fonction des produits scalaires entre l'entrée X et les vecteurs de support.
    - ☐ Les SVM avec une fonction noyau non linéaire consistent à rechercher une frontière de décision linéaire dans un nouvel espace de dimension plus faible que l'espace initial.

17. On considère la régression à vecteurs de supports (*Support Vector Regression*).

   (a) $\boxed{\text{1 point}}$ La fonction de coût optimisée par cette méthode est :

   - ☐ $\min(0, |f(x) - y| - \epsilon)$.
   - ☐ $\max(\epsilon, |f(x) - y|)$.
   - ☐ $\max(\epsilon, |f(x) - y| + \epsilon)$.
   - ☐ $\max(\epsilon, |f(x) - y| - \epsilon)$.

   (b) $\boxed{\text{2 points}}$ Représentez graphiquement le coût en fonction de la différence $f(x) - y$.

18. $\boxed{\text{5 points}}$ L'ACP à noyaux (*Kernel PCA*)

   - ☐ nécessite de diagonaliser une matrice de taille $N \times N$, où N est le nombre d'exemples.
   - ☐ nécessite de diagonaliser une matrice de taille $p \times p$, où p est le nombre de variables.
   - ☐ permet de construire q nouvelles variables, $q \leq p$.
   - ☐ revient à faire une ACP dans un nouvel espace de représentation défini par une fonction de noyau.
   - ☐ permet de construire de nouvelles variables, fonctions linéaires de variables initiales.

19. $\boxed{\text{5 points}}$ On considère les données suivantes :

obs.	1	2	3	4	5	6	7
X_1	3	2	4	1	2	4	4
X_2	4	2	4	4	1	3	1
classe	$+1$	$+1$	$+1$	$+1$	-1	-1	-1

Représentez ces données. Tracez l'hyperplan séparateur optimal, et indiquez les vecteurs de support. Exprimez la fonction de décision sous la forme $f(X) = \text{sign}(\beta^T X + \beta_0)$, où $X = (X_1, X_2)^T$ et (β, β_0) le vecteur de paramètres que l'on précisera. Que vaut la marge ?

20. On considère un modèle de mélange gaussien à $K = 2$ composantes en dimension $p = 2$.
    (a) $\boxed{\text{3 points}}$ Représentez graphiquement la forme des classes dans les trois cas suivants : (a) classes de même forme mais de volumes et orientations différents (b) classes de mêmes forme et orientation mais de volumes différents ; (c) classes de mêmes forme et volume mais d'orientations différentes.

    (b) $\boxed{\text{3 points}}$ Pour chacun des modèles précédents, donnez le nombre de paramètres à estimer (proportions comprises) :
    (a) Même forme, volumes et orientations différents.

    (a) ______________

    (b) Mêmes forme et orientation, volumes différents.

    (b) ______________

    (c) Mêmes forme et volume, orientations différentes.

    (c) ______________

21. $\boxed{\text{5 points}}$ Vous disposez d'un ensemble de $N = 1000$ exemples et souhaitez entraîner un SVM avec un noyau gaussien. Quels sont les hyperparamètres ? Comment les déterminez-vous ? Comment obtenez-vous une estimation sans biais de la probabilité du meilleur classifieur obtenu ?