# Introduction to Statistical Supervised Machine Learning

## Examen 2023 - 2024

## Master 1 MIASHS

**Guillaume Metzler**

**Institut de Communication (ICOM)**
**Université de Lyon, Université Lumière Lyon 2**
**Laboratoire ERIC UR 3083, Lyon, France**

**guillaume.metzler@univ-lyon2.fr**

**Durée : 2h00**

**Les documents personnels, notes de cours, téléphones et ordinateurs ne sont pas autorisés pour cet examen. Vous aurez cependant le droit à deux feuilles A4 recto-verso avec vos notes mauscrites.**

**Abstract**

Les exercices sont tous indépendants et peuvent être traités dans l'ordre qui vous conviendra. On prendra cependant soin de bien indiquer les numéros des questions traitées ainsi que les exercices correspondants. Pour certaines questions, vous pouvez illustrer vos propos à l'aide de graphiques, dessins ou autres tableaux si cela vous semble pertinent.

## Exercice 1 : SVM à noyaux

Les paramètres d'un modèle de séparateurs à vastes marges sont obtenus en résolvant le problème d'optimisation :

$$\begin{array}{ll} \min\limits_{\xi\in\mathbb{R}^m,\,(\mathbf{w},b)\in\mathbb{R}^{d+1}} & \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{m}\sum_{i=1}^m \xi_i \\ s.t. & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1-\xi_i, \quad \text{pour tout } i = 1, ..., m, \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad \text{pour tout } i = 1, ..., m. \end{array} \tag{1}$$

### Vers la formulation duale

L'objectif est d'obtenir la version duale de ce problème. Pour cela, on considère la fonction suivante, que l'on appelle le **lagrangien** du problème d'optimisation, qui va permettre d'étudier notre problème dans un autre espace. Ce lagrangien est donné par

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{m}\sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i\left(1-\xi_i - y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right)\right) - \sum_{i=1}^m \beta_i\xi_i.$$

où $\alpha_i$ et $\beta_i$, pour $i \in [\![1, n]\!]$ sont appelées les variables lagrangiennes associées aux deux contraintes du problème (1).

On va commencer par exprimer notre problème en fonction des variables lagrangiennes uniquement afin de déterminer un problème d'optimisation équivalent.

1. En considérant les trois équations suivantes

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0, \quad \frac{\partial\mathcal{L}}{\partial b}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0, \quad \text{and} \quad \frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0,$$

montrer que l'on

$$\mathbf{w} = \sum_{i=1}^m y_i\alpha_i\mathbf{x}_i, \quad \sum_{i=1}^m \alpha_i y_i = 0, \quad \text{and} \quad \frac{C}{m} - \alpha_i - \beta_i = 0, \iff \alpha_i, \beta_i \geq 0.$$

On pourra admettre ce résultat pour la suite des questions si besoin.

2. En utilisant les expressions obtenues dans la question précédente et les injectant dans le lagrangien, montrer que ce dernier peut s'écrire :

$$-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{m} \alpha_i. \tag{2}$$

3. Montrer que le problème (2) peut s'écrire sous la forme

$$-\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + \sum_{i=1}^{m} \alpha_i,$$

où $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ et $\mathbf{K}$ une matrice dont précisera la dimension et la définition. On étudiera également la convexité de ce problème.

4. Dans cette formulation, $\mathbf{K}$ est une fonction à **noyaux**. Quelles sont les hypothèses que doit vérifier cette fonction à noyaux ?

5. Le noyau présenté ci-dessus est appelé **noyau linéaire**. Donnez deux autres exemples de noyaux vus en cours.

### Classification

On considère le jeu de données étiquetées suivant :

| $y$ | -1 | -1 | -1 | -1 | +1 | +1 | +1 |
|---|---|---|---|---|---|---|---|
| $x_1$ | 2 | 4 | -1 | 0 | 1 | 6 | 5 |
| $x_2$ | 1 | 3 | 4 | 7 | -6 | -3 | -5 |
| $\alpha$ | 0.1 | 0.3 | 0 | 0.5 | 0.7 | 0 | 0.2 |

On souhaite déterminer l'étiquette de la donnée $\mathbf{x}'$ définie par $\mathbf{x}' = (1, 5)$ et on considère un noyau **linéaire**.

1. Rappeler la règle de classification pour un SVM à noyaux.

2. Déterminer l'étiquette prédite par le modèle pour la donnée $\mathbf{x}'$ précédemment définie.

### Etude des points supports (Difficile)

Dans cette dernière partie, on cherchera à déterminer quels sont les points supports d'un SVM.

1. Rappeler ce qu'est, graphiquement, un point support pour un séparateur à vaste marge. On fera également le lien avec la valeur $\xi_i$ associée à un tel point.

2. Dans la théorie de l'optimisation, les relations dites de KKT conduisent à toute une série d'équations dont certaines sont utilisées pour obtenir la formulation duale d'un problème d'optimisation. C'est ce que vous avons fait dans la première partie de cet exercice.

   Mais d'autres équations dites de *complementary slackness conditions*, données par :

$$\begin{aligned} \alpha_i(1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) &= 0, \quad (3) \\ \beta_i \xi_i &= 0. \quad (4) \end{aligned}$$

   En utilisant les équations (3) and (4) ainsi que le problème (1). Montrer que si $\alpha_i = 0$ alors la données $\mathbf{x}_i$ n'est pas un point support et si $0 < \alpha_i \leq \frac{C}{m}$, alors point est un point support.

## Exercise 2 : Forêts aléatoires

### Quelques questions de cours

1. En utilisant vos propres mots, expliquez la signification du terme présenté ci-dessous ainsi que le rôle de ce dernier dans un contexte de *bagging*.

$$\frac{1}{T} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{X}, y \sim \mathcal{Y}} \left[ \sum_{t=1}^{T} (h_t(\mathbf{x}) - y)^2 \right] \geq \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{X}, y \sim \mathcal{Y}} \left[ (y - H_T(\mathbf{x}))^2 \right],$$

   où $H_T = \sum_{t=1}^{T} h_t$, $y \in \mathbb{R}$ représente la valeur prédite par la combinaison de modèles, $\mathbf{x} \in \mathbb{R}^d$ désigne un individu et $h_t$ sont les apprenants.

2. Expliquer le principe du double échantillonnage qui est couramment utilisé dans les modèles de forêts aléatoires. On précisera notamment l'intérêt de ce dernier sur les plans théorique et pratique.

3. Expliquer comment est effectué la validation des modèles de forêts aléatoires. On prendra soin de préciser quel est, en moyenne, le pourcentage des exemples utilisés pour la validation de ces derniers.

**Construction d'une forêt aléatoire**

On considère le jeu de données de classification suivant :

| $y$ | -1 | -1 | -1 | -1 | -1 | -1 | +1 | +1 | +1 | +1 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | -3 | -4 | 2 | 4 | -1 | 0 | 1 | 6 | 5 | 8 |
| $x_2$ | -4 | -2 | 1 | 3 | 4 | 7 | -6 | -3 | -5 | 9 |

Notre objectif est de construire une forêt constituée de deux arbres ayant chacun une profondeur de 1, *i.e.*, une seule séparation (ou split) sera effectué.

**A. Construction d'un premier arbre** Pour ce premier arbre, seule la **première** variable est utilisée ($x_1$). De plus, on considérera uniquement les exemples $\mathbf{x}_i$ pour $i \in [\![2, 9]\!]$ pour entraîner notre arbre et les autres exemples serviront à la validation.

1. Evaluer l'indice de Gini à la racine de cet arbre.
2. Déterminer la split optimal et évaluer le gain de Gini (on pourra s'aider d'un dessin pour trouver le split optimal)
3. Déterminer les performances, en *accuracy*, de ce modèle $h_1$, sur l'ensemble d'apprentissage et sur l'ensemble de validation.

**B. Construction d'un deuxième arbre** Pour ce premier arbre, seule la **deuxième** variable est utilisée ($x_2$). De plus, on considérera uniquement les exemples $\mathbf{x}_i$ pour $i \in [\![1, 8]\!]$ pour entraîner notre arbre et les autres exemples serviront à la validation.

1. Evaluer l'indice de Gini à la racine de cet arbre.
2. Déterminer la split optimal et évaluer le gain de Gini (on pourra s'aider d'un dessin pour trouver le split optimal)
3. Déterminer les performances, en *accuracy*, de ce modèle $h_2$, sur l'ensemble d'apprentissage et sur l'ensemble de validation.

**C. Performance de la forêt** On considère la forêt aléatoire défini par $H_T = \frac{1}{2}(h_1 + h_2)$ et le jeu de données suivant :

| $y$ | -1 | -1 | +1 | +1 |
|---|---|---|---|---|
| $x_1$ | -1 | 2 | 3 | 4 |
| $x_2$ | -5 | 6 | -1 | -2 |

Evaluer les performances, en terme d'accuracy, de la forêt aléatoire sur ce jeu de données.

## Exercise 3 : Une méthode ensembliste pour la régression

**Dans cet exercice, les valeurs numériques seront données à une précision de l'ordre de $10^{-2}$, soit deux chiffres après la virgule. On peut aussi faire le choix de les donner sous forme de fraction**

### Régression linéaire

Dans cette première partie, on considère le modèle de régression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

où $\mathbf{y} \in \mathbb{R}^m, \mathbf{X} \in \mathscr{M}_{m,d+1}$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

On considère le jeu de données suivant :

| $y$ | 1 | 4 | 2 | 5 | 2 | 4 |
|---|---|---|---|---|---|---|
| $x_1$ | 4 | 0 | -2 | -1 | 1 | -2 |
| $x_2$ | 1 | 0 | 0 | 0 | -2 | 1 |

On considère le problème de régression suivant :

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2, \tag{5}$$

1. Donner l'expression du vecteur $\boldsymbol{\theta}$ en fonctinon $\mathbf{X}$ et $\mathbf{y}$.
2. Déterminer la valeur de $\boldsymbol{\theta}$ à l'aide des données de l'énoncé.

### Régression quadratique

On considère à présent le modèle suivant

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \varepsilon,$$

En utilisant le fait que

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 18 \\ -11 \\ 1 \end{pmatrix}$$

and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.378 & -0.090 & -0.049 \\ -0.090 & 0.077 & -0.021 \\ -0.049 & -0.021 & 0.011 \end{pmatrix}.$$

En déduire la valeur du paramètre $\boldsymbol{\theta}$.

**Un modèle ensembliste**

On considère un modèle ensembliste $H_T$ qui combine les résultats de la régression linéaire simple et de la régression quadratique en moyennant les prédictions. Déterminer les valeurs prédites par ce modèle ensembliste sur les vecteurs $\mathbf{x}_1, \mathbf{x}_3$ et $\mathbf{x}_5$.