



Introduction to Statistical Supervised Machine Learning

Examen 2022 - 2023

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 : 1h30

Les documents personnels, notes de cours, téléphones et ordinateurs ne sont pas autorisés pour cet examen.

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.

Autour de l'apprentissage

Soit $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$ un échantillon d'apprentissage, ℓ une loss et \mathbf{w} les paramètres d'un modèle. On considère un algorithme de classification quelconque qui cherche à résoudre le problème d'optimisation suivant :

$$\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^m \ell(h_{\mathbf{w}}(\mathbf{x}_i), y_i) + \lambda \|\mathbf{w}\|, \quad (1)$$

où $h_{\mathbf{w}}$ représente une hypothèse (un classifieur ou un régresseur) qui dépend du paramètre \mathbf{w} .

1. Dans le problème d'optimisation (1), identifier le terme que l'on appelle *risque empirique* et le *terme de régularisation*.
2. Quel est le nom de λ et quel est son rôle ?
3. On se concentre maintenant sur le terme $\|\mathbf{w}\|$
 - (a) Quel est l'impact, sur le modèle appris, en choisissant $\|\mathbf{w}\| = \|\mathbf{w}\|_1$?
 - (b) Quel est l'impact, sur le modèle appris, en choisissant $\|\mathbf{w}\| = \|\mathbf{w}\|_2$?
4. Rappeler quel est l'objectif général en Machine Learning (notamment vis à vis du risque empirique et du risque en généralisation).
5. Décrire précisément la procédure que vous devriez mettre en place pour résoudre le problème (1) permettant de répondre à la question précédente.

Algorithmes

Cet exercice se concentre sur l'étude de certains algorithmes, on va considérer que l'on dispose du jeu d'entraînement S suivant

Individu	x_1	x_2	y
1	-2	1	1
2	1	-1	1
3	3	4	1
4	0	-3	-1
5	1	-2	-1
6	-1	-4	-1

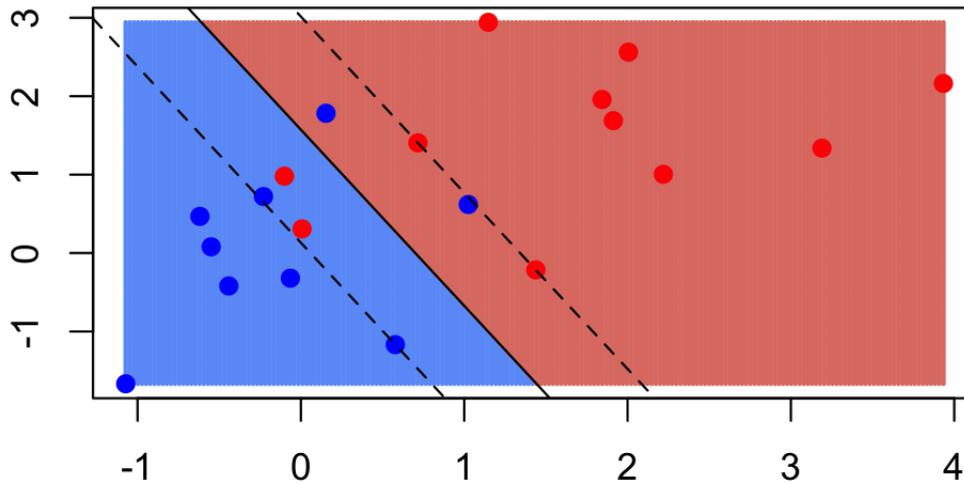


Figure 1: Classifieur SVM linéaire. La zone bleue représente la zone de prédiction négative, *i.e.* $y = -1$ et la zone rouge représente la zone de prédiction positive, *i.e.* $y = +1$.

Algorithme du plus proche voisin

1. Représenter les individus $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_6$ sur un dessin.
2. Expliquer le fonctionnement de l'algorithme du plus proche voisin (*i.e.* k -NN lorsque $k = 1$)
3. On considère les points $\mathbf{z}_1 = (1, 1), \mathbf{z}_2 = (0, 3)$ et $\mathbf{z}_3 = (-1, 1)$. Déterminer leur étiquette à l'aide de l'algorithme du plus proche voisin.
4. Sur le plan de l'apprentissage, quel est le spécificité de cet algorithme ?
5. Quel est le principal inconvénient de cet algorithme, notamment lorsque la taille de l'échantillon d'apprentissage est grande ?

Algorithme du SVM

On considère un jeu d'entraînement S' qui a conduit à l'obtention du SVM linéaire représenté en Figure 1 et dont les paramètres sont approximativement les suivants :

$$\mathbf{w} = (-1.5, -0.5) \quad \text{et} \quad b = -1.$$

1. Rappeler la règle de classification d'un SVM linéaire.
2. Donner la définition de la *hinge loss* et énoncer le problème d'optimisation à résoudre pour un SVM linéaire.

3. Sur la Figure 1, identifier :
 - (a) l'hyperplan séparateur
 - (b) les marges du SVM
 - (c) les vecteurs (ou points) supports
4. Prédire l'étiquette des individus \mathbf{x}_1 , \mathbf{x}_2 et \mathbf{x}_6 à l'aide du SVM (on demande de le faire par un calcul et non graphiquement).
5. Quelle est la valeur de la loss pour le point \mathbf{x}' de coordonnées $(1, -1)$ qui est un point dont le label est négatif, *i.e.* $y = -1$.

Régression logistique

1. Rappeler ce qu'est la régression logistique. On expliquera le principe et on rappellera la règle de classification.
2. Quelle est la loss que l'on cherche à optimiser dans la phase d'apprentissage. On rappellera la définition de cette loss.
3. On rappelle que le modèle logistique repose sur l'utilisation d'une fonction logistique

$$s(x) = \frac{1}{1 + \exp(-x)}.$$

Etudier la convexité de cette fonction.

Méthodes ensemblistes

1. Quelles sont les deux grandes méthodes ensemblistes vues en cours ? On décrira brièvement les principes de ces deux approches. On pourra également citer un algorithme emblématique des différentes approches ensemblistes.
2. Décrire les différentes étapes de l'algorithme Adaboost ci-dessous :

Input: Echantillon d'apprentissage S de taille m ,
un nombre T de modèles
Output: Un modèle $H_T = \sum_{t=0}^T \alpha_t h_t$

begin

- Distribution uniforme $w_i^{(0)} = \frac{1}{m}$
- for** $t = 1, \dots, T$ **do**
 - Apprendre un classifieur h_t à partir d'un algorithme \mathcal{A}
 - Calculer l'erreur ε_t de l'algorithme.
 - if** $\varepsilon_t > 1/2$ **then**
 - | Stop
 - else**
 - Calculer $\alpha_{(t)} = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right)$
 - $w_i^{(t)} = w_i^{(t-1)} \frac{\exp(-\alpha_t y_i h_t(\mathbf{x}_i))}{Z_t}$
- Poser $H_T = \sum_{t=0}^T \alpha_t h_t$
- return** H_T

3. Quelle est loss que l'on cherche à minimiser avec l'algorithme adaboost ?
On considère le modèle suivant :

$$H(\mathbf{x}) = \text{sign}(\alpha_1 h_1(\mathbf{x}) + \alpha_2 h_2(\mathbf{x}) + \alpha_3 h_3(\mathbf{x})),$$

où $(\alpha_1, \alpha_2, \alpha_3) = (1, 2, 3)$ et $h_1(\mathbf{x}) = \text{sign}(2)$, $h_2(\mathbf{x}) = \text{sign}(2x_1)$ et $h_3(\mathbf{x}) = \text{sign}(-x_1 + x_2 - 1)$.

- (a) En repartant de la description de l'algorithme Adaboost, quel est le *weak learner* qui a l'erreur la plus faible ?
- (b) Prédire l'étiquette des données $\mathbf{x}_1, \mathbf{x}_2$ et \mathbf{x}_3 de l'ensemble S défini à l'exercice précédent.