Machine Learning

Examen 2022 - 2023
Master 2 Informatique - BI&A

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 sont autorisées pour cet examen. En revanche, l'usage du téléphone portable est interdit.

Prénom :

Nom :

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.

1. Quels sont les deux grandes familles d'algorithmes ? Donnez deux exemples d'algorithmes pour chacune de ces familles.

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 :

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

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. Décriver 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

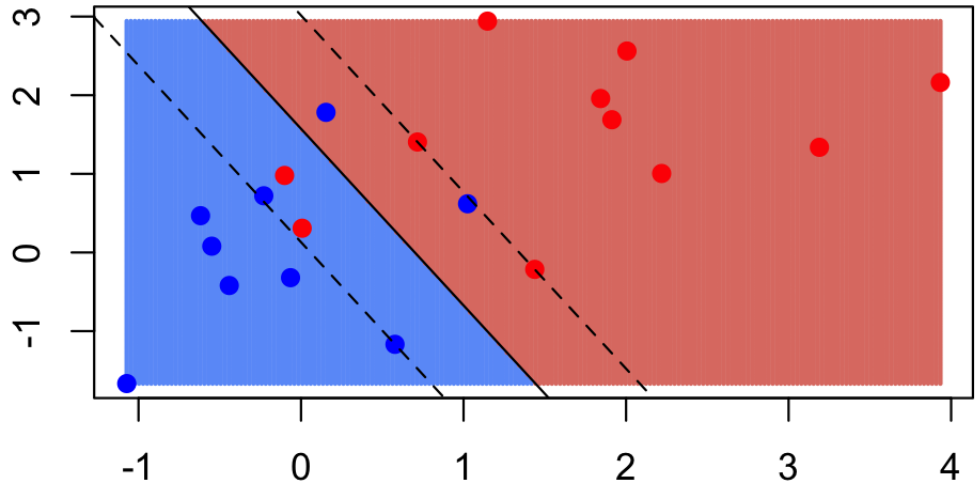

Figure 1: Classifieur SVM linéaire. La zone bleue représente la zone de prédiction négative, *i.e.* $y = -1$ et la zone rouge représente la zone de prédiction positive, *i.e.* $y = +1$.

Algorithme du plus proche voisin

1. Représenter les individus $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_6$ 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. Comment est déterminé l'hyperplan séparateur, sur le plan conceptuel ?
5. Prédire l'étiquette des indivus $\mathbf{x}_1, \mathbf{x}_2$ et $\mathbf{x}_6$ à l'aide du SVM (on demande de le faire par un calcul et non graphiquement).
6. 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.* $y = -1$.
7. Qu'est-ce qu'une méthode à noyau? Donnez un exemple de jeu de données pour lequel il est préférable d'utiliser une méthode à noyau plutôt qu'une version linéaire des SVM. Quelles sont les limites des méthodes à noyaux ?

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. Est-il possible de combiner les deux approches précédentes ?
3. 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 $w_i^{(0)} = \dfrac{1}{m}$
  for $t = 1, ..., T$ **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 $\alpha_{(t)} = \dfrac{1}{2} \ln \left(\dfrac{1 - \varepsilon_t}{\varepsilon_t} \right)$
      $w_i^{(t)} = w_i^{(t-1)} \dfrac{\exp(-\alpha_t y_i h_t(\mathbf{x}_i))}{Z_t}$
  Poser $H_T = \sum_{t=0}^{T} \alpha_t h_t$
  return H_T

---

4. Quelle est loss que l'on cherche à minimiser avec l'algorithme adaboost ?

   On considère le modèle suivant :

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

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

   (a) En reppartant de la description de l'algothime 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.

Imbalanced Learning

1. Quelles sont les difficultés de l'apprentissage dans un contexte déséquilibré ? Citez trois méthodes différentes qui peuvent aider à traiter cette problématique. Quand vous les connaissez, donnez des avantages et inconvénients pour les différentes méthodes.

2. Expliquer pourquoi il n'est pas toujours bon de chercher à rééquilibrer votre jeu de données dans ces contextes là.

3. Expliquer pourquoi l'*Accuracy* n'est pas une mesure de performance adaptée. On pourra donner un exemple et citer une mesure de performance alternative.

4. Tous les jeux de données déséquilibrés représentent-ils nécessairement des tâches complexes à résoudre ? Quid des jeux de données équilibrées ?

Autres

1. Rappeler quels sont les éléments qui caractérisent un réseau de neurones et quels sont les paramètres sur lesquels nous pouvons jouer pour éviter le sur-apprentissage. Citez au moins trois types de réseaux de neurones différents. Sur quel type de données les réseaux de neurones sont les plus adaptés ?

2. Nous avons vu une famille d'algorithme appelées les SVDD, dont le problème consistait à apprendre la plus petite sphère englobant vos données. Quel autre algorithme est proche de ce problème là et quel est sont usage ? Vous pourrez également indiquer s'il s'agit d'un algorithme d'apprentissage supervisé ou non.

3. Quel(s) autre(s) algorithmes d'apprentissage non supervisé connaissez-vous ? En donner une brève description.