Machine Learning

Examen 2023 - 2024
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 : 2h15

Les notes de cours ne pas sont autorisées pour cet examen tout comme l'usage du téléphone portable. Seules vos notes personnelles manuscrites sont autorisées.

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. Quelles sont les trois grandes catégories d'algorithme que l'on peut rencontrer en Machine Learning ? Pour deux de ces catégories, donner un exemple algorithme ainsi qu'une application pratique pour cet algorithme.

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. Rappeler quel est l'objectif général en Machine Learning (notamment vis à vis du risque empirique et du risque en généralisation) et les éventuelles problèmes que l'on peut rencontrer.
4. Rappeler le principe de k-fold cross-validation.

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

Régression Logistique

1. Est-ce un algorithme de régression ou de classification ? Décrire son fonctionnement.
2. Rappeler la définition de la loss logistique
3. Montrer que la
4. On considère maintenant que notre algorithme nous donne les résultat suivants sur notre jeu de données.

Individu	$\mathbb{P}(Y = 1 \mid X = \mathbf{x})$	y
1	0.8	1
2	0.7	1
3	0.65	-1
4	0.4	1
5	0.3	-1
6	0.2	1

   (a) On considère un seuil de 0.5 pour notre algorithme. En déduire l'Acccuracy, la Précision, le Rappel et la F-mesure de votre modèle.

   (b) Evaluer l'AUC-ROC de ce modèle.

5. Comment pourriez vous étendre cet algorithme pour faire traiter des problèmes multi-classes.

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 résodure pour un SVM linéaire.
3. Sur la Figure **??**, identifier :

   (a) l'hyperplan séparateur

   (b) les marges du SVM

   (c) les vecteurs (ou points) supports

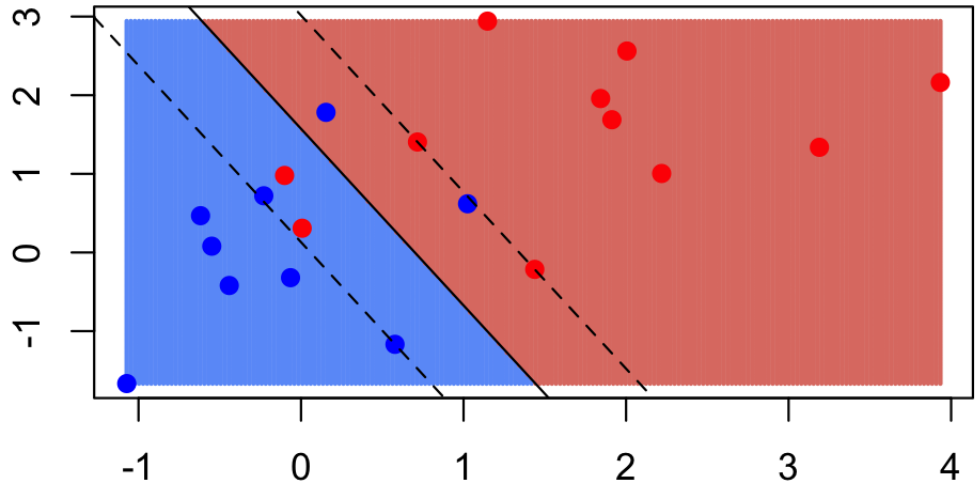

Figure 1: Classifieur SVM linéaire. La zone bleue représente la zone de prédiction négative, *i.e.* $y = -1$ et la zone rouge représente la zone de prédiction positive, *i.e.* $y = +1$.

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. Rappeler le fonctionnement de l'algorithme Adaboost (on peut écrire un pseudo-code). Les relations mathématiques sont attendues.
   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 $(\alpha_1, \alpha_2, \alpha_3) = (2, 4, 1)$ et $h_1(\mathbf{x}) = (-3x_1 + 5x_2 - 1), h_2(\mathbf{x}) = (2x_1 - 2x_2)$ et $h_3(\mathbf{x}) = (6x_1 - 7x_2 - 3)$.

(a) En partant de la description de l'algorithme Adaboost, quel est le *weak learner* qui a l'erreur la plus faible parmi les trois précédents ?

(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.

3. Quel autre algorithme de boosting pouvez-vous citer. En donner une brève description.

4. Décrire l'algorithme des forêts aléatoires. On pourra s'aider d'un schéma pour la présentation de ce dernier.

5. Sur quelle composante de l'erreur l'algorithme des forêts aléatoires va t-il agir, comparé à un arbre de décision.

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 ?