Fouilles de Données Massives

Examen 2022 - 2023
Master 2 Informatique - SISE

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 personnels, notes de cours sont autorisées pour cet examen, sous format électronique ou papier. En revanche, l'usage du téléphone portable est interdit ainsi que toute ressources en ligne autre que le cours.

Prénom :

Nom :

Abstract

L'examen se décompose en deux parties de durées équivalentes. La première partie est essentiellement composée de questions dont des éléments de réponses figurent dans le cours. La deuxième partie propose une mise en situation où le but sera d'exposer une démarche de résolution au problème proposé.

La qualité de la rédaction, notamment la clarté des explications fournies, sera un élément important dans l'évaluation de la copie.

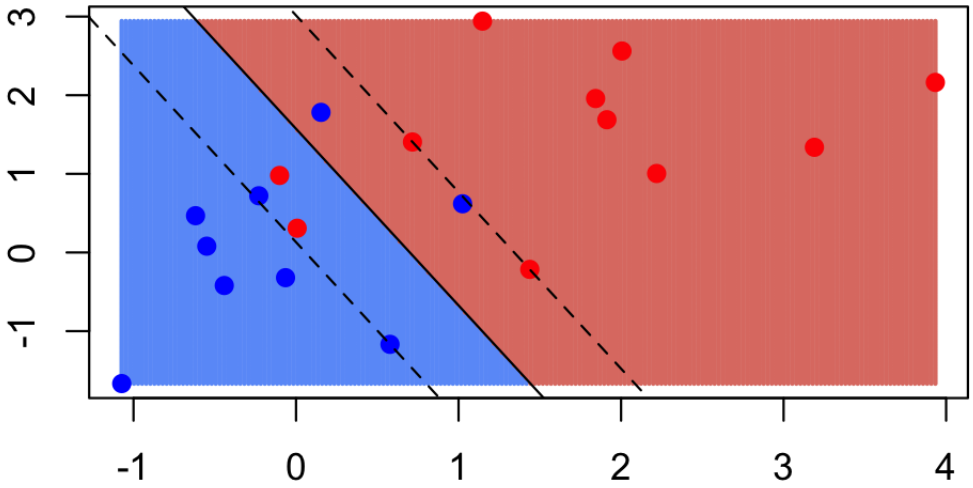

Figure 1: Classifieur SVM linéaire. La zone bleue représente la zone de prédiction négative, *i.e.* $y = -1$ et la zone rouge représente la zone de prédiction positive, *i.e.* $y = +1$.

Première Partie

Généralités

1. Quelles sont les principales caractéristiques du Big Data ? Qu'est-ce qui explique un tel engouement pour les données ?

2. Expliquer, avec vos propres mots, les différences et ressemblances entre le boosting classique (comme l'algorithme Adaboost) et un algorithme de Gradient Boosting.

Algorithme du SVM

On considère un jeu d'entraînement S qui a conduit à l'obtention du SVM linéaire représenté en Figure 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 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. 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 ?
6. En repartant d'une méthode à noyaux, expliquer comment il est possible de retrouver l'équation de l'hyperplan séparateur à partir des données et de la résolution du problème dual.
7. Comment peut-on approcher les méthodes à noyaux ?

Deuxième Partie

Données génomiques

Avec le développement des nouvelles technologies, il est actuellement plus simple de séquencer le génome des individus. A l'aide de ces avancées, et dans le cadre d'une étude clinique, on cherche à déterminer si certaines caractéristiques physiques, physiologiques ou encore génomiques pré-sélectionnées sont déterminantes dans la détection d'une maladie.

Pour cela les données étudiées se présentent comme suit :

- une suite de nucléotides qui donne la représentation d'un gène (et de ses différentes versions). Cet encodage se fait sur un vecteur de taille 100 000 environ.
- certaines caractéristiques spécifiques aux patients comme le sexe, l'age, le poids, la taille, le nombre globules rouges, de globules blancs

Etant donnée la difficulté liée à l'étude, cette dernière est effectuée sur 1000 individus environ et certaines informations, non liées au séquencage de l'ADN, sont manquantes.

Proposer une procédure complète : pré-traitement des données, procédure et choix de l'algorithme, permettant de discriminer au mieux les individus malades et non malades à l'aide des descripteurs disponibles.

Fraude fiscale

Tout comme les particuliers, les entreprises sont tenues de déclarer leur bilan comptable chaque année afin d'être assujetti aux impôts. Il arrive cependant que certaines entreprises, environ 5%, minimisent leur rentabilité pour réduire le montant de ses impôts,

i.e. certaines entreprises fraudes.

Dans le cadre de la lutte contre la fraude fiscale, la DGFiP (Direction Générale des Finance Publiques) souhaite mettre en place un système permettant de les détecter sur la base la déclaration de ces dernières. Les informations dont la DGFiP dispose sont des informations relatives à l'entreprise, comme le nombre de salariés, ses clients, son chiffre d'affaire, celui des clients, etc ... au total, plus d'une centaine de variables qui peuvent présenter certaines corrélations du au regroupement de différentes bases de données.

L'objectif de la DGFiP est de minimiser la perte due à la fraude fiscale. Proposer une procédure complète : pré-traitement des données, procédure et choix de l'algorithme, permettant de détecter mais aussi de minimiser les pertes engendrées par la fraude fiscale.