Fouilles de Données Massives

Examen 2023 - 2024
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 : 3h00

Les notes personnelles manuscrites sont autorisées pour cet examen : 3 feuilles A4 recto-verso En revanche, l'usage du téléphone portable, de l'ordinateur ou de tout autre matériel électronique est interdit. Les téléphones sont en mode silencieux (ou éteints) à l'intérieur de votre sac.

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.

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.

3. Expliquer l'intérêt du double échantillonnage que l'on retrouve dans l'algorithme des forêts aléatoires.

4. Dans l'algorithme de Boosting, plus précisément dans Adaboost, un résultat théorique énonce que l'on a l'inégalité suivante

$$\mathcal{R}_S \leq \exp(-2\gamma^2 T),$$

où $\mathcal{R}_S$ est le risque empirique sur l'échantillon S et $\gamma > 0$ et T est le nombre de modèles appris.
Que se passe-t-il lorsque $T \to \infty$? Est-ce réalisable en pratique ?

5. Rappelez quelles sont les différentes approches vues en cours qui permettent de traiter un problème de classification binaire dans un contexte déséquilibré. On illustrera ces propos en décrivant quelques méthodes.

Autour des SVM

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 α_i et β_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
α	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.

Approximation des méthodes à noyaux

1. Expliquez quelles sont les limites des méthodes à noyaux lorsque le nombre de données est considérable et pourquoi ? Les méthodes à noyaux sont-elles toujours problématiques dans un tel contexte ? Pourquoi ?
2. Expliquez quoi en l'usage de landmarks, peut-être intéressant dans le cadre de l'approximation de méthodes à noyaux.
3. Donnez une autre approche étudiée en cours qui permet d'approcher des méthodes à noyaux et expliquez brièvement l'idée et son fonctionnement.

Deuxième Partie

Données sur la transplantation

La santé est un domaine lequel l'utilisation des intelligences artificielles et des modèles Mathématiques est de plus en plus répandue. Le domaine de la transplantation d'organe en fait partie. La greffe permet de sauver un grand nombre de vies chaque années. Son taux de réussite dépend grandement de la compatibilité entre donneurs et receveurs d'organes. Pour créer ce *matching*, les médecins se basent sur des caractéristiques biologiques et physiologiques de l'organe concerné afin d'estimer la **probabilité de survie à un temps** t d'un greffon après transplantation.

Une base de données extrêmement riche est mise à disposition d'un data scientist afin qu'il puisse travailler sur l'élaboration d'un modèle permettant d'améliorer ce matching donneur receveur à l'aide des observations effectuées sur les précédentes greffes.

Il travaille dans le contexte suivant :

- la base de données étudiées comporte plus d'un million de greffe (une ligne = une greffe)
- et plus de 1000 caractéristique sur la greffe, greffon, donneur et receveur.

L'estimation de durée de vie n'étant pas un point étudié en cours, on va se contenter de voir cela comme un problème de **régression.**

Proposer une procédure complète : pré-traitement des données, procédure et choix de l'algorithme, mesure de performance, permettant d'estimer au mieux ce temps de survie et qui tienne compte des contraintes. Si vous effectuez des stratégies particulières, vous prendrez le soin de bien détailler tout cela.

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.