Examen : Big Data Mining

M2 SISE

Durée : 2h00

Résumé

Les questions du sujet sont globalement indépendantes, vous pouvez dons les traiter dans l'ordre que vous souhaitez.

Certaines questions n'appellent pas de réponses précises mais sont là pour avoir votre point de vue sur un problème ou encore la stratégie que vous allez mettre en place pour traiter un problème selon le contexte et le type de données.

La qualité de la rédaction sera prise en compte dans l'évaluation. Vous pouvez répondre aux différentes questions à l'aide de simples phrases ou encore de schémas si nécessaires.

Questions

1. Qu'est-ce qui caractérise le big data ? Quelles sont les trois grandes branches en Machine Learning. On illustrera chaque branche par un exemple d'application ou algorithme.

2. Etant donnés un ensemble d'apprentissage $\mathrm{S} = \{\mathbf{x}_i\}_{i=1}^m$, une fonction $f : \mathbb{R} \to \mathbb{R}^+$ et une fonction $\ell : \mathbb{R} \to \mathbb{R}^+$, on considère le problème d'optimisation :

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}, \mathbf{x}_i) + \lambda f(\mathbf{w}).$$

   (a) Donnez la signification de chaque chaque terme $\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{w}, \mathbf{x}_i)$, λ et $f(\mathbf{w})$.

   (b) Pour chacun des choix de fonction f suivant, rappelez l'impact sur le modèle appris :

   - $f : w \mapsto \|w\|_1$
   - $f : w \mapsto \|w\|_2$

   (c) Décrire un protocole expérimentale **général** (je ne parle pas d'algorithme là) permettant d'apprendre $\mathbf{w}$ et de cross-valider λ.

3. Nous avons vu que nous sommes capables d'établir des bornes en généralisation sur les performances de notre algorithme à partir de la notion de **stabilité uniforme**. Ces bornes prennent la forme suivante :

$$|\mathcal{R}^{\ell}(\theta_S) - \mathcal{R}^{\ell}_S(\theta_S)| \leq \mathcal{O}\left(m^{-1/2}, \lambda^{-1}\right),$$

où $\mathcal{R}^{\ell}(\theta_S)$ représente le risque du modèle sur notre distribution inconnue, $\mathcal{R}^{\ell}_S(\theta_S)$ et le dernier terme représente la valeur de notre borne.

   (a) Donnez la signification de cette borne et éventuellement une interprétation en fonction de la valeur de λ.

   (b) Rappelez l'hypothèse importante sur notre problème d'optimisation permettant d'établir ce type de bornes.

   (c) Quelles sont deux autres "outils" vus en classe qui permettant de construire ce type de bornes ?

4. Citer 5 algorithmes permettant de faire de la classification. Parmi les 5 algorithmes cités, on en choisira un dont on expliquera le fonctionnement.

5. Rappeler en quoi consiste le clustering et son usage. Expliquez aussi comment cette méthode peut être combinée à d'autres algorithmes.

6. Expliquez ce qui distingue un un arbre d'une forêt aléatoire et comment les forêts aléatoires sont construites. On prendra également le soin de préciser différentes règles de votes pour les forêts.

7. Expliquer ce qu'est le boosting et donner deux exemples d'algorithmes de boosting.

8. Expliquer ce qu'est le *Metric Learning* et comment cela fonctionne (on demande simplement d'expliquer l'intuition)
9. Rappelez quelles sont les différentes composantes d'un réseau de neurones, pour un réseau de neurones dit "classique" (on pourra faire un schéma pour illustrer ses propos).

10. Citez au moins 5 "paramètres sur lesquels on peut jouer dans l'apprentissage de notre réseau (d'un point de vue structure ou optimisation) avec quelques explications brèves concernant ces paramètres (si nécessaires !).

11. De quel type de réseau avons nous besoin pour faire de la classification ou de segmentation sémantique d'une image ? Expliquez rapidement le fonctionnement de ce type de réseaux.

Mise en situation

Dans cette partie là, on vous propose deux mises en situations en vous décrivant un certain problème avec des données particulières. Vous devrez expliquer le protocole mis en place pour répondre à la tâche de demandée. Cela va de l'éventuelle préparation des données (création de variables éventuellement, normalisation des données) au choix de(s) algorithme(s) et de la loss, mesure de performance ou encore des combinaisons des méthodes (clustering - sampling - réduction de dimension - ...)
On prendra également le soin de préciser les avantages et inconvénients des méthodes proposées.

1. On dispose d'un jeu de données brutes qui représentent des transactions par carte bancaire. Après un travail de *création de features*, notre jeu de données comporte un ensemble de 30 features, certaines de ces variables sont numériques et seule la variable *type d'enseigne* (qui peut prendre 3 valeurs différentes) est catégorielle. On dispose également de la date des transactions (une information que l'on pourra ou non prendre en compte). Ce jeu de données comporte environ 6 mois de transactions et seules 1% de ces transactions sont frauduleuses.
   L'objectif est alors déterminer quels sont les transactions frauduleuses de celles qui ne le sont pas.

   (a) Proposez une approche non supervisée pour la résolution de cette tâche.

   (b) Proposez une approche supervisée pour la résolution de cette tâche.

2. Notre deuxième tâche est la suivante : je dispose d'un jeu de données de taille $m = 1000$ (nombre d'observations) et comportant 10000 attributs qui sont uniquement numériques et la variable réponse est un nombre réel. On ne donne pas plus d'informations sur les descripteurs si ce n'est que ces derniers se trouvent sur des échelles de grandeurs différentes. L'objectif est donc de prédire la valeur de cette variable réponse.