Ecole Polytechnique de Thies
Département Génie Informatique et Télécommunications
TD AD2-ML SVM 2018-2019

Exercice 1 *SVM*

1. *Quel est le principe fondamental des machines à vecteurs support ? Sous quel autre nom les connaît-on ?*
2. *Q'est ce qu'un vecteur support ?*
3. *Pourquoi est'il important de normaliser les données d'entrée lorsqu'on utilise les SVM ?*
4. *Un classifieur SVM peut-il fournir un indice de confiance lorsqu'il classe une observation ? Et peut-il fournir une probabilité ?*
5. *Si votre jeu d'entraînement possède des millions d'observations et de centaines de variables, dvez vous utiliser la forme optimale du problème SVM ou sa forme duale pour entraîner le modèle ?*
6. *Supposons que vous ayez entraîné un classifieur SVM avec un noyau à base radiale. Il semble sous-ajuster le jeu d'entraînement : devez-vous augmenter ou diminuer γ (gamma) ? Même question pour C.*
7. *Comment devez-vous définir les paramètres de progarammation quadratique (H,f, A et b) pour résoudre un problème de classification SVM linéaire à magre souple en utilisant un solveur QP proposé sur le marché ?*

Exercice 2 *Arbre de décision*

1. *Quelle est la profondeur approximative d'une arbre de décision entraîné (sans restrictions) sur un jeu d'entraînement comportant 1 million d'observations ?*
2. *L'impureté Gini d'un noeud est-elle en général inférieure ou supérieure à celle du noeud parent ? Est-elle généralement inférieure/supérieure ou toujours inférieure/supérieure ?*
3. *Si un arbre de décision surajuste le jeu d'entraînement, est-il judicieux d'essayer de diminuer max_depth ?*
4. *Si un arbre de décision sous-ajuste le jeu d'entraînement, est-il judicieux d'essayer de normaliser les caractéristiques d'entrée ?*
5. *S'il faut une heure pour entraîner un arbre de décision sur un jeu d'entraînement comportant 1 million d'observations, combien de temps faudra-il approximativement pour entraîner un autre arbre de décision sur un jeu d'entraînement contenant 10 millions d'observations ?*

6. *Si votre jeu d'entraînement comporte 100 000 observations, allez-vous accélérer l'apprentissage en spécifiant* ***presort=True*** *?*

Exercice 3 *Forêt aléatoire*

1. *Si vous avez entraîné cinq modèles fifférentes sur les mêmes données d'entraînement et s'ils ont tous atteint une précision de* 95%*, y a-t-il un moyen de combiner ces modèles pour obtenir de meilleurs résultats ? Si oui, comment ? Sinon, pourquoi ?*
2. *Quelle est la différence entre les classificateurs à vote rigide et à vote souple ?*
3. *Est-il possible d'accélérer l'entraînement d'un ensemble de bagging en le distribuant entre plusieurs serveurs ? Même question pour des ensembles de type pasting, boosting, forêt aléatoire ou stacking ?*
4. *Quel est l'avantage de l'évaluation hors sélection ?*
5. *Qu'est-ce qui rend les extra-arbres plus aléatoires que les forêts aléatoires normales ? En quoi cette part de hasard supplémentaire peut-elle aider ? Les extra-arbres sont-ils plus lents ou plus rapides que les forêts aléatoires normales ?*
6. *Si votre ensemble AdaBoost sous-ajuste les données d'entraînement, sur quels hyperparamètres pouvez-vous jouer, et comment ?*
7. *Si votre ensemble à boosting de gradient surajuste le jeu d'entraînement, devez-vous augmenter ou diminuer le taux d'apprentissage ?*