DEVOIR I

Fondements Théorétiques en Sciences des Données

STT 3795 – Hiver 2020

Date limite pour remiser le devoir (sur StudiUM): 14 février à 23h59

Pour implémenter les Exercices 1-3, cliquez **ici** et suivez les instruction sur Google Colaboratory. Remettez la partie théorique comme fichier PDF. Vous pouvez numériser votre solutions écrites à la main si elles sont lisibles ou alors utiliser LaTeX.

Exercice 1

Le but de cet exercice est l'analyse d'un ensemble de tweets. Vous trouvez les données (tweets.txt) sur la page web du cours. Suivez les instruction suivantes pour l'implémentation:

- Mettez les données disponible sur Colab et les gérez dans une liste (ou un tableau) appelée «tweets»
- Trouvez les 10 mots de longueur d'au moins 5 caractères les plus fréquents apparaissant dans les tweets et les mémorisez dans une liste (ou un tableau) «terms»
- Vérifiez que le mot le plus fréquent est *iphone* et le négligez pour les étapes suivantes
- Créez un tableau (188x9) contenant les nombres des mots les plus fréquents pour chaque tweet
- Créez la matrice de corrélation (9x9) entre ces mots
- Générez un tableau (9x2) ou une liste des listes avec les mots trouvés avant d'après l'ordre alphabétique dans la première colonne et le mot comportant la corrélation maximale (excluant le mot soi-même) dans la deuxième colonne
- Imprimez cet objet dans votre code et l'ajoutez à votre fichier PDF

Exercice 2

Soit $S^k \subset \mathbb{R}^{k+1}$ la sphère de dimension k défini d'après

$$S^k \triangleq \left\{ x = (x_1, x_2, \ldots, x_k, x_{k+1}) \in \mathbb{R}^{k+1} : \quad \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2 + x_{k+1}^2} = 1 \right\}.$$

Pour un ensemble des points $X \subset \mathbb{R}^d$, définissons la distance D par

$$D \triangleq \{\|x - y\|_2 : x, y \in X\}.$$

Pour $k = 1, 2, 3$, effectuez les étapes suivantes:

- Générez 1000 points X_k uniformément dans $S^k \subset \mathbb{R}^{k+1}$
- Calculez les distances par paires D_k pour X_k
- Générez un histogramme de 25 boites avec une latitude égale
- Sauvegardez l'histogramme comme "eqwidth_k.png"
- Générez un histogramme de 25 boites chaque contenant le même nombre de points de D_k et chaque ayant la même superficie (notez que la latitude de chaque boite de cet histogramme est la longueur de l'intervalle minimale contenant les valeurs de ce boite)

Ajoutez les six histogrammes à votre fichier PDF.

Exercice 3

Dans ce problème vous allez implémenter une machine à vecteurs de support avec l'astuce du noyau (kernel SVM).

1. Implémentez la fonction d'après
$$\text{Md} = \text{svm\_train}(\text{c}, \text{X}, \text{K})$$
avec
   - $c \in \{-1, 1\}^n$ le vecteur contenant les labels
   - $X \in \mathbb{R}^{n \times m}$ la matrice des attributs
   - $K : (x_1, x_2) \mapsto K(x_1, x_2)$ une fonction anonyme qui retourne le noyau pour $x_1, x_2 \in \mathbb{R}^m$
   - Md une structure des données de votre choix qui représente le modèle

2. Implémentez une fonction de classification d'après
$$[\text{c\_hat}, d] = \text{svm\_classify}(\text{Md}, Y)$$
avec
   - Md la structure des données obtenue à l'étape précédente
   - $Y \in \mathbb{R}^{\ell \times m}$ la matrice des attributs
   - c_hat $\in \{-1, 1\}^\ell$ le vecteur contenant les labels prédits
   - $d \in \mathbb{R}^\ell$ le vecteur contenant les niveaux de confiance
   - Notez que pour un nouveau vecteur d'attributs $y \in \mathbb{R}^m$ le niveau de confiance peut être calculé avec la fonction sigmoïde d'après
$$\text{sigmoid}\left(\sum_{i=1}^{n} c_i \lambda_i K(x_i, y) - b\right)$$

3. Testez votre code en utilisant un noyau linéaire
   - Entraînez le modèle avec les données "simple_iris.mat" pour $K(x, y) = \langle x, y \rangle$
   - Générez le graphique suivant et l'ajoutez au fichier PDF:
     (a) Tracez les points avec label -1 en rouge
     (b) Tracez les points avec label 1 en bleu
     (c) Tracez les vecteurs de support en noir
     (d) Tracez les trois hyperplans
$$\{x : w^T x - b = -1\}, \quad \{x : w^T x - b = 0\}, \quad \{x : w^T x - b = 1\}$$

4. Testez votre code en utilisant un noyau quadratique
   - Entraînez le modèle avec les données "simple_nonlinear.mat" pour $K(x, y) = (\langle x, y \rangle + 1)^2$
   - Créez une grille (meshgrid) Y (20x20) couvrant les données d'entraînement X
   - Classifiez cette grille en utilisant votre fonction de classification
   - Générez les deux graphiques suivants et les ajoutez au fichier PDF:
     – Tracez les points de la grille avec les couleurs comme suggérées dans 3.(a)-(c)
     – Tracez les points de la grille colorés d'après leur niveau de confiance

Instructions pour l'implémentation:

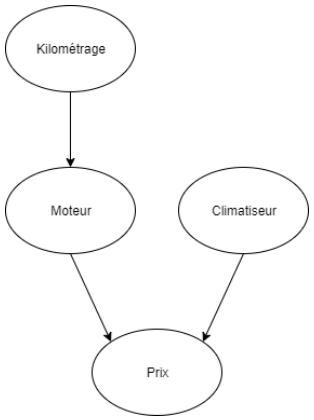
- Résolvez le problème en utilisant seulement les packages déjà téléchargés sur Google Colab
- Implémentez kernel SVM en utilisant une optimisation quadratique appropriée. Servez vous de la commande solvers. qp() pour trouver une solution
- Expliquez votre choix des paramètres pour l'optimisation quadratique dans le fichier PDF

Exercice 4

1. Considérez les données suivantes qui comportent des attributs binaires:

Kilométrage	Moteur	Climatiseur	fréquence (si prix élevé)	fréquence (si prix bas)
haut	bon	fonctionne	3	4
haut	bon	cassé	1	2
haut	mauvais	fonctionne	1	5
haut	mauvais	cassé	0	4
bas	bon	fonctionne	9	0
bas	bon	cassé	5	1
bas	mauvais	fonctionne	1	2
bas	mauvais	cassé	0	2

   - Complétez les tables contenant les probabilités conditionnelles pour le réseau bayésien suivant:

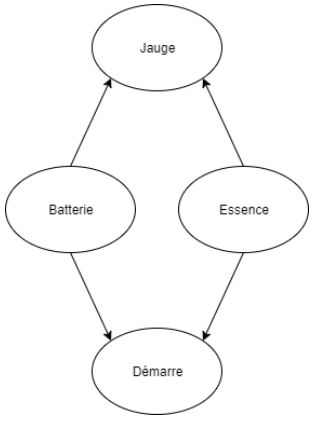


   - Utilisez les tables pour calculer $P(M = \text{mauvais}, C = \text{cassé})$. Expliquez le calcul.

2. Considérez le réseau bayésien suivant et calculez

   - $P(B = \text{bon}, E = \text{vide}, J = \text{vide}, D = \text{oui})$
   - $P(B = \text{mauvais}, E = \text{vide}, J = \text{non-vide}, D = \text{no})$
   - La probabilité que la voiture démarre soit la batterie est mauvais

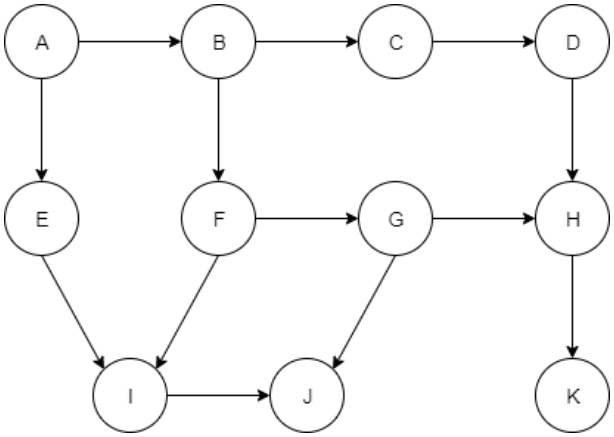
   Utilisez les tableaux des probabilités conditionnelles ci-dessous. Posons $P(B = \text{mauvais}) = 0.1$ et $P(E = \text{vide}) = 0.2$.



Batterie	Essence	Jauge vide
bon	non-vide	0.1
bon	vide	0.8
mauvais	vide	0.9
mauvais	non-vide	0.2

Batterie	Essence	ne démarre pas
bon	non-vide	0.1
bon	vide	0.8
mauvais	vide	1.0
mauvais	non-vide	0.9

Exercice 5



1. En utilisant le principe de d-séparation, répondez de manière brève aux questions ci-dessous. Une justification n'est pas nécessaire.
   (a) $B \perp E$?
   (b) $B \perp E \mid A$?
   (c) $B \perp E \mid A, K$?
   (d) $B \perp E \mid A, H, G$?
   (e) $A \perp I$?
   (f) $A \perp I \mid E$?
   (g) $B \perp G \mid F$?
   (h) $B \perp G \mid F, J$?
   (i) Existe-il un sommet S tel que $H \perp I \mid S$
   (j) Inversez exactement une arête telle que $E \perp G$
2. Déterminez la couverture de Markov du sommet G.
3. Montrez que sachant la couverture de Markov d'un sommet, ce dernier est indépendant des autres sommets du réseau.