Devoir 3
STT6415 - Régression
Échéance: mardi 15 mars.

1. **Les données *diabète.*** Obtenez l'arbre de régression optimal pour log(C-peptide) en fonction d'âge et du déficit de base. Utilisez la validation croissée. Montrez l'arbre original et l'arbre optimal. Faites le graphique de la validation croissée de la fonction constante par intervalles RSS(lambda). Commentez sur la relation entre les variabls prédictives et la réponse.

2. **Les données de CPU.** Obtenez l'arbre de régression optimal pour log(performance). Utilisez la validation croissée. Montrez l'arbre original et l'arbre optimal. Faites le graphique de la validation croissée de la fonction constante par intervalles RSS(lambda). Quelles variables ont été utilisées dans l'arbre de régression? Commentez sur la relation entre les variabls prédictives et la réponse.

3. **Considérez le problème suivant à deux classes.** Les deux classes, des données à deux dimensions (x_1, x_2) sont générées comme suit: si (r, θ) sont les coordonnées polaires, alors la distribution de la classe 1 est uniforme sur $3 \leq r \leq 4$, et $0 \leq \theta \leq \pi = 3.14159(...)$; et la distribution de la classe 2 est uniforme sur $4 \leq r \leq 5$, et $0 \leq \theta \leq \pi$.

   (a) Générez (simulez) 100 échantillons provenant de chaque distribution, et faites un graphique des données.

   (b) Utilisez l'analyse discriminante linéaire. Dessinez la frontière linéaire correspondante qui sépare les deux classes sur le graphique des données.

   (c) Utilisez l'analyse discriminante quadratique. Dessinez la frontière quadratique correspondante qui sépare les deux classes sur le graphique des données.

   (d) Construissez l'arbre de classification optimal. Dessinez les régions correspondantes dans le nuage des données.

   (e) Quelle est la frontière évidente de séparation optimale entre les classes?
   Remarque : Vous devez travailler avec la paire (x_1, x_2). Ne travaillez pas avec les coordonnées polaires.

4. **La diabète chez les amérindians Pima.** Une population de femmes de descendance amérindien Pima qui avaient au moins 21 ans, et vivant près de Phoenix, Arizona, a été testé pour le diabète selon les critères de l'Organisation Mondiale de la Santé. Les données ont été recueillies par l'Institut national américain du diabète et des maladies digestive et renal.

   Le fichier de données contient les lignes suivantes

   - npreg: nombre de grossesses
   - glu: la concentration plasmatique de glucose dans un test de tolérance au glucose oral

- pb: pression artérielle diastolique (mm Hg)
- skin: épaisseur de la peau de triceps (mm)
- bmi: indice de masse corporelle (poids en kg / (taille en m^2))
- ped: fonction du "pedigree" du diabète
- l'âge: années
- Type: Non / Oui

Effectuez une régression logistique additive afin d'analyser ces données.