Université du Québec à Montréal (UQAM)
Faculté des sciences

ACT6100– Examen Final
Analyse de données en actuariat

Enseignant : Noureddine Meraihi

2020/04/16

Cet examen contient 10 pages (incluant la page couverture) et 12 questions sur un total de 30 points.

Instructions

— L'examen commence à 09 :00 pour une durée de 180 minutes ;
— Vous avez le droit d'utiliser votre ordinateur **SEULEMENT** pour ;
  — Vous connecter à *Zoom Meetings*
  — consulter le questionnaire de l'examen
  — Écrire vos réponses sur le cahier de réponse BRUH123456.txt
— Il est strictement **interdit** d'utiliser un quelconque moyen de communication pendant l'examen ;
— Il est strictement **interdit** de faire des recherches sur le web ;
— Prévoyez un 5 minutes pour la remise de cahier de réponse ;
— Déposez votre fichier de réponse ici : http ://tiny.cc/act6100h20 ;
— L'examen compte pour $max(45{,}35)\%$ de la note finale ;
— Vous serez informés par courriel/Slack lorsque l'examen sera corrigé.

1. (13 points) Pour chacun des énoncés ci-dessous, indiquer si l'affirmation est vraie ou fausse.
   - i Le *clustering* cherche à trouver une représentation en basse dimension des observations qui expliquent une bonne fraction de la variance
   - ii Une procédure de forêts aléatoires (random forests) peut permettre de décorréler les modèles utilisés pour la prédiction finale.
   - iii Le modèle de régression linéaire est un des modèles les plus flexibles vus dans le cadre de ce cours.
   - iv Si on utilise toutes les variables explicatives à chacune des étapes d'une procédure de forêts aléatoires, on obtient des résultats similaires à ceux obtenus à partir d'une procédure de bagging.
   - v L'algorithme AdaBoost est construit de façon à minimiser la fonction de risque empirique

$$\frac{1}{n}\sum_{i=1}^{n} e^{-Y_i r(\mathbf{X}_i)}.$$

   - vi Un arbre de décisions qui contient davantage de feuilles aura généralement une variance plus faible.
   - vii En utilisant un paramètre de complexité de 0 pour élaguer un arbre de décisions, on obtient généralement un arbre qui contient uniquement un noeud (racine) avec deux feuilles.
   - viii Il n'est pas possible d'utiliser un modèle d'arbre de décisions si au moins une des variables explicatives est catégorielle.
   - ix Lorsqu'un méta-algorithme de Boosting est utilisé, l'erreur de généralisation va systématiquement exploser empiriquement lorsque le nombre d'étapes (T) devient grand.
   - x Lorsqu'on applique un modèle basé sur les arbres de décision, une fois que les régions R_1,...,R_J ont été créées ,nous prédisons la réponse pour une observation de test donnée en utilisant la moyenne des observations des données d'entraînement dans la région à laquelle cette observation de test appartient.
   - xi Nous décidons généralement du nombre de composantes principales principale nécessaires pour visualiser les données en examinant un graphique de validation croisée par rapport à la variance expliqué via un nuage de points (*scree plot*)
   - xii L'ACP cherche à trouver des sous-groupes homogènes parmi les observations.
   - xiii La proportion de la variance expliquée (PVE) par chaque composante principale est simplement la proportion de la variance des données qui n'est pas contenue dans les ces premières composantes principales.

2. (1 point) Pour modéliser la sévérité d'un sinistre, un actuaire considère, comme point de départ, un modèle de régression linéaire classique avec 7 variables explicatives donné par

$$Y = \beta_0 + \sum_{i=1}^{7} \beta_i X_i.$$

Il souhaite conserver, dans son modèle de régression linéaire final, $K \leq 7$ variables explicatives, afin de minimiser l'erreur quadratique moyenne. Combien de modèle(s) devra-t-il ajuster afin de sélectionner son modèle final ?

A. 7 modèles

B. 8 modèles

C. 128 modèles

D. 256 modèles

E. 5 040 modèles

3. (1 point) On a vu que pour la régression Ridge, on peut obtenir, de façon équivalente, les paramètres du modèle en résolvant

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

sous une certaine contrainte. Laquelle ?

A. $\displaystyle\sum_{j=1}^{p} \beta_j^2 \leq \psi$

B. $\lambda \leq \psi$

C. $\displaystyle\sum_{j=1}^{p} \beta_j^2 \geq \psi$

D. $\displaystyle\sum_{j=0}^{p} \beta_j^2 \leq \psi$

E. Aucune de ces réponses

4. (1 point) La Figure 1 illustre un résultat que l'on peut obtenir après une régression linéaire Lasso appliquée à une base de données contenant une variable réponse continue et 37 variables explicatives. Parmi les propositions ci-dessous, laquelle correspond à l'étiquette de l'axe des ordonnées ?

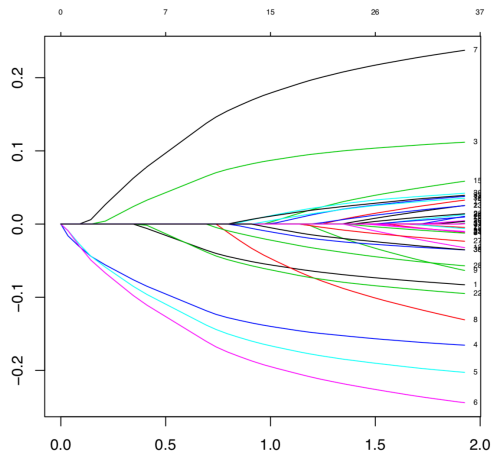

Figure 1: Résultats.

A. Erreur quadratique moyenne (MSE)
B. Déviance Poisson
C. Nombre de variables
D. Taux de mauvais classification
E. Aucune de ces réponses.

5. (1 point) Une base de données de taille $n = 1\,000$ est composée d'une variable réponse binaire (0 ou 1) et de deux variables explicatives : X_1 et X_2. Un modèle d'arbre de **régression** est ajusté et l'arbre final est présenté à la Figure 2. L'espace initial des variables explicatives est divisé en combien de régions par cet arbre ?

A. 2
B. 3
C. 7
D. 8
E. 10

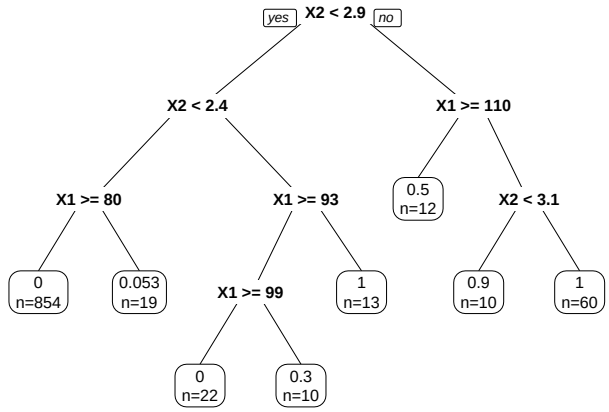

Figure 2: Arbre final.

6. (2 points) À partir de l'arbre final présenté à la Figure 2, parmi les observations pour lesquelles $X_2 \geq 3.1$, combien ont une variable réponse égale à 1 ?

   A. entre 45 et 57 (bornes incluses)
   B. exactement 60
   C. entre 60 et 66 (bornes incluses)
   D. exactement 66
   E. strictement plus de 66

7. (1 point) Pour construire un arbre de décisions, on utilise un algorithme **AAAAA** descendant fonctionnant par division binaire **BBBBB**. Quels mots ou expressions remplacent **AAAAA** et **BBBBB** respectivement ?

   A. de validation croisée, hiérarchique
   B. de validation croisée, descendante
   C. glouton (*greedy approach*), hiérarchique
   D. glouton (*greedy approach*), récursive
   E. de validation croisée, convexe

8. (1 point) Lorsque l'on cherche à élaguer (pruning) un arbre de décisions $\mathcal{T}_0$, on minimise la fonction

$$\sum_m \sum_{i:\mathbf{X}_i \in R_m} \left(Y_i - \widehat{Y}_{R_m}\right)^2 + \alpha \mathcal{T}.$$

Que représente $\mathcal{T}$?

A. le paramètre de complexité
B. la valeur absolue du paramètre de complexité
C. la taille de l'échantillon
D. le nombre de feuilles du sous-arbre $\mathcal{T}$
E. un paramètre lié à la taille de l'espace d'hypothèses

9. (4 points) Pour chacune des affirmations ci-dessous, indiquer si elle fait référence au *Bagging*, au *Boosting* ou au *Bagging* ET au *Boosting*.

(i) Construction de H modèles de façon séquentielle.
(ii) Peut permettre de réduire la variance de la prédiction.
(iii) N'aura pas un très bon pouvoir prédictif si les modèles utilisés sont fortement biaisés.
(iv) Une prédiction est obtenue en calculant une moyenne des prédictions faites par chacun des modèles utilisés.

10. (1 point) Le code présenté ci-dessous permet d'ajuster un modèle en utilisant un méta-algorithme de *Boosting*.

```
watchlist <- list(train = dtrain, test = dtest)
bst <- xgb.train(data = dtrain,
                 max.depth = 8,
                 eta = 0.3,
                 nthread = 2,
                 nround = 1000,
                 watchlist = watchlist,
                 objective = "reg:linear",
                 early_stopping_rounds = 50,
                 print_every_n = 500)

[1] train-rmse:7.02 test-rmse:6.25
Multiple eval metrics are present.
Will use test_rmse for early stopping.
Will train until test_rmse hasn't improved in 50 rounds.

Stopping. Best iteration:
[49]  train-rmse:0.05 test-rmse:4.3
```

Quelle est la valeur du taux d'apprentissage ?

A. 0.3
B. 0.5
C. 2
D. 4.3
E. 8

11. (1 point) En utilisant les informations présentées dans le code de la question précédente (10) pour construire un modèle final qui sera utilisé sur des données, quelle est la valeur de l'erreur quadratique moyenne d'entrainement ?

A. $[0.00, 0.04)$
B. $[0.04, 0.30)$
C. $[0.30, 4.00)$
D. $[4.00, 7.00)$
E. $[7.00, \infty)$

12. (3 points) Une étude a été réalisée à partir de cinq ménages afin de modéliser certaines habitudes de consommation. Pour chacun des ménages, le montant total du revenu net (mensuel) et le montant total consacré aux loisirs (mensuel) ont été mesurés. À partir de ces cinq points dans $\mathbb{R}^2$, la matrice de distances suivantes a été construite :

$$D_1 = \begin{bmatrix} 0 & 2 & 4 & 7 & 9 \\ & 0 & 8 & 9 & 8 \\ & & 0 & 3 & 7 \\ & & & 0 & 5 \\ & & & & 0 \end{bmatrix}.$$

En utilisant les distances simples (simple linkage) entre les groupes et un algorithme hiérarchique ascendant (agglomerative hierarchical algorithm), quels seront les deux **derniers** groupes (clusters) à être regroupés ?

A. $\{12\}$ et $\{345\}$
B. $\{1234\}$ et $\{5\}$
C. $\{124\}$ et $\{35\}$
D. $\{135\}$ et $\{24\}$
E. Aucune de ces réponses.

Fin de l'examen

Régularisation

— Régression linéaire sous forme matricielle :

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} \\ \widehat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{Y}. \end{aligned}$$

— Régression Ridge :

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{\boldsymbol{R}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}, \qquad \lambda > 0 \\ &= \left(\mathbf{X}^{**T}\mathbf{X}^{**} + \lambda\mathbf{I}_p\right)^{-1} \mathbf{X}^{**T}\mathbf{Y}, \end{aligned}$$

où $\mathbf{I}_p$ est une matrice identité $(p \times p)$ et

$$\mathbf{X}^{**} = \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \dots & \ddots & \dots \\ X_{n1} & \dots & X_{np} \end{bmatrix}.$$

— Régression Lasso :

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{L}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1} \beta_j, \qquad \lambda > 0.$$

Arbres de décisions

— Pour une région r, si le mode est utilisée comme prédiction, on peut utiliser le taux d'erreur de classification donné par

$$E_r = 1 - \max_{g \in \mathcal{G}} \widehat{p}_{rg},$$

où $\widehat{p}_{rg}$ est la proportion d'observations i pour lesquelles $\mathbf{X}_i \in r$ et dont la variable réponse est g.

— L'index Gini associé à la région r est donné par

$$G_r = \sum_{g \in \mathcal{G}} \widehat{p}_{rg}(1 - \widehat{p}_{rg}).$$

— L'entropie associée à la région r est donnée par

$$D_r = -\sum_{g \in \mathcal{G}} \widehat{p}_{rg} \ln(\widehat{p}_{rg}) \mathbb{I}_{\{\widehat{p}_{rg} \neq 0\}}.$$

Agrégation de modèles

— Erreur apparente :

$$\epsilon_t = p_t\left(r_t(\mathbf{X}_i) \neq Y_i\right) = \sum_{i:r_t(\mathbf{X}_i)\neq Y_i} p_t(i).$$

— Borne supérieure de l'erreur apparente :

$$\epsilon_T \leq \exp\left(-2\sum_t \gamma_t^2\right) \leq \exp\left(-2T\gamma^2\right),$$

où $\gamma = \min(\gamma_1, \ldots, \gamma_T)$.

— Borne supérieure de l'erreur de généralisation :

$$\mathrm{Err}_G \leq \epsilon_T + \sqrt{\frac{(T)(d_{\mathcal{H}})}{n}}.$$

Analyse en composantes principales

— Décomposition spectrale d'une matrice $\mathbf{R}$ $(k \times k)$:

$$\begin{aligned} \det(\mathbf{R} - \lambda \mathbf{I}_k) &= 0 \\ \mathbf{R}\mathbf{a}_i &= \lambda_i \mathbf{a}_i, \qquad i = 1, \ldots, k. \end{aligned}$$

— Pour des données centrées-réduites, la distance entre deux observations est donnée par

$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{k} m_j (x_j - y_j)^2}.$$

— La corrélation entre deux variables j_1 et j_2 est donnée par

$$\begin{aligned} r_{j_1 j_2} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij_1} - \bar{x}^{j_1}}{s_{j_1}}\right)\left(\frac{x_{ij_2} - \bar{x}^{j_2}}{s_{j_2}}\right) \\ &= \cos\theta\left(\mathbf{z}^{j_1}, \mathbf{z}^{j_2}\right), \end{aligned}$$

où $\theta(\mathbf{a}, \mathbf{b})$ indique l'angle entre les vecteurs $\mathbf{a}$ et $\mathbf{b}$.

— La projection ($\mathbf{M}$-orthogonale) d'un point $\mathbf{z}_i \in \mathbb{R}^k$ sur un axe D_α dont la direction est donnée par un vecteur $\mathbf{v}_\alpha$ de norme 1 a pour coordonnée

$$p_{i\alpha} = < \mathbf{z}_i^T, \mathbf{v}_\alpha >_{\mathbf{M}} = \mathbf{z}_i \mathbf{M} \mathbf{v}_\alpha.$$

— La projection ($\mathbf{N}$-orthogonale) d'une variable $\mathbf{z}^j \in \mathbb{R}^n$ sur un axe G_β décrit par un vecteur $\mathbf{u}_\beta$ a pour coordonnée

$$t_{j\beta} = < \mathbf{z}^j, \mathbf{u}_\beta > = (\mathbf{z}^j)^T \mathbf{N} \mathbf{u}_\beta.$$

— La qualité de la projection de l'observation i sur l'axe D_α est donnée par :

$$\cos^2(\theta_{i\alpha}) = \frac{\mathbf{p}_{i\alpha}^2}{||\mathbf{z}_i||^2}.$$

— La contribution de l'observation i sur l'axe D_α est mesurée par

$$C(i,\alpha) = \frac{\mathbf{p}_{i\alpha}^2}{\lambda_\alpha}.$$

— La qualité de la projection de la variable j sur l'axe G_β est donnée par

$$\cos^2(\theta_{j\beta}) = \frac{\mathbf{t}_{j\beta}^2}{||\mathbf{z}^j||^2} = \mathbf{t}_{j\beta}^2.$$

— La contribution de la variable j sur l'axe G_β est mesurée par

$$C(j,\beta) = \frac{\mathbf{t}_{j\beta}^2}{\lambda_\beta}.$$