Examen final

N. Saunier et François Bélisle

16 décembre 2021

Veuillez

- noter le barème (la note totale est sur 20) et le temps indicatif à consacrer à chaque exercice;
- indiquer clairement les numéros des questions que vous traitez et vos réponses correspondantes (et souligner ou encadrer les résultats numériques);
- apporter une attention particulière à la rédaction et à la définition des notations que vous employez;
- noter que certains exercices nécessitent des fichiers disponibles sur Moodle (Section "Examen final") (les fichiers texte sont fournis en version avec le point et la virgule pour les nombres décimaux, si nécessaire). Des tables statistiques sont disponibles sur Moodle si nécessaire.

Exercice 1 (fouille de données) 30 min (/4 pts)

Un arbre de décision a été appris à partir du jeu de données des survivants du Titanic (fichier `titanic.txt` disponible sur Moodle). Chaque personne est décrite par quatre attributs: son statut sur le bateau (équipage, 1ère, 2ème ou 3ème classe), son âge (adulte ou enfant), son genre (homme ou femme) et s'il a survécu ou pas. L'arbre de décision est présenté dans la figure 1 et sur Moodle pour la lisibilité.

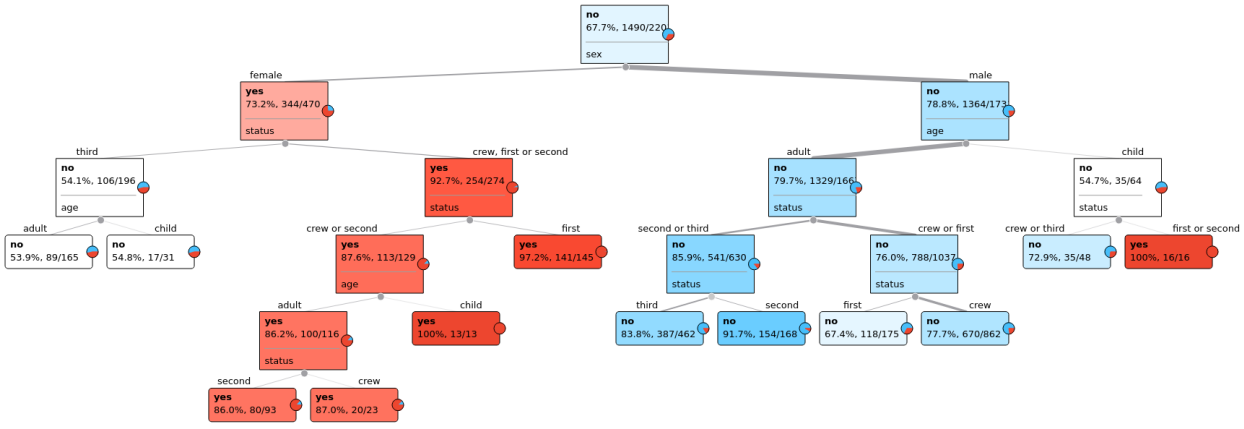

Figure 1: Arbre de décision de la survie d'un passager du Titanic.

1. Veuillez décrire les types des attributs du jeu de données: à quelle catégorie appartient la tâche de prédire la survie d'un passager? (1 pt)

2. Quelle est l'attribut le plus important pour prédire la survie des passagers selon l'arbre de décision? Justifier la réponse. (1 pt)

3. Écrire deux règles complètes de décision de l'arbre (de la racine à la feuille) de niveaux de "confiance" très différents. (1 pt)

4. Quel autre modèle pourrait être utilisé pour prédire la survie des passagers? Quels sont les avantages des différents modèles? (1 pt)

Solution

1. La variable prédite (si un passager a survécu ou pas au naufrage du Titanic) est catégorielle, binaire. Prédire cette variable est donc une tâche de classification.

2. Le sexe (genre) est l'attribut le plus important pour prédire la survie des passagers puisque c'est le premier choisi par l'arbre pour la classification.

3. Si sex = female et status = third et age = adulte, alors survie = no pour 53.9 % des personnes concernées
   Si sex = male et age = child et status = first or second, alors survie = yes pour 100 % des personnes concernées

4. Un autre modèle pouvant faire de la classification est un classifieur bayésien naïf, avec l'avantage qu'il fonctionne directement avec des attributs catégoriels.

Exercice 2 (analyse et fouille de données) 35 min (/4.5 pts)

On effectue une analyse de segmentation afin de mieux comprendre le contenu du fichier de données "cars". La méthode des k-moyennes est appliquée pour trois groupes. Le résultat est sauvé dans la base de données `cars.db` (disponible sur Moodle), avec la colonne Cluster contenant l'identifiant du groupe auquel chaque observation est assignée. Les centroïdes de chaque groupe sont présentés dans le tableau 1.

Cluster	N	MPG	Weight	Drive_Ratio	Horsepower	Displacement	Cylinders
C1	10	21.69	2.8865	3.378	110.7	156.2	5.5
C2	10	17.98	3.7359	2.482	128.4	296.8	7.4
C3	17	31.01	2.2474	3.307	77.6	109.2	4

Tableau 1: Valeur des centroïdes de chaque groupe.

1. Expliquer comment les centroïdes sont calculés (définition mathématique) et la requête SQL qui peut générer le tableau 1. (1.5 pts)

2. Décrire à l'aide du fichier et du tableau les trois groupes en termes des valeurs des attributs (caractéristiques des véhicules dans chaque groupe et en quoi sont-elles différentes des autres groupes)? (1.5 pts)

3. Utiliser un test statistique pour comparer le poids (variable "Weight") entre les différents groupes (un logiciel peut être utilisé): présenter les résultat du test et tirer la conclusion. (1 pt)

4. Proposer une méthode de fouille de données vue en cours pour analyser les groupes. (0.5 pt)

Solution

1. Les centroïdes sont le centre des éléments (vecteurs) de chaque groupe : la valeur de chaque attribut du centroïde est la moyenne arithmétique des attributs des éléments du groupe. La requête pour calculer les attributs des centroïdes est la suivante:
   ```
   SELECT Cluster, COUNT(*) AS N, AVG(MPG), AVG(Weight), AVG(Drive_Ratio),
   AVG(Horsepower), AVG(Displacement), AVG(Cylinders) FROM cars GROUP BY Cluster
   ```
2. test ANOVA
3. On voit clairement d'après les attributs de poids, consommation et puissance du moteur que le groupe C3 est le groupe des véhicules les plus efficaces et légers et à l'opposé, le groupe C2 le groupe des véhicules les plus consommateurs de carburant et les plus lourds. Le groupe C1 est intermédiaire, hormis pour driver ratio où il est proche des véhicules les plus lourds.
4. On peut utiliser un arbre de décision pour prédire le groupe d'appartenance des autos à partir de leurs attributs (y compris le pays d'origine): les variables choisies par l'algorithme de construction de l'arbre sont les plus "utiles" pour prédire la variable, donc ont une certaine corrélation avec l'appartenance au groupe.

Exercice 3 (modèle de régression) 20 min (/3 pts)

Les résultats d'un modèle de choix discret de type logit sont présentés dans le tableau 2. La variable à prédire est le choix de l'avion pour un déplacement (les alternatives sont le train, le bus et la voiture). Les variables explicatives du modèle sont les suivantes:

- invt: temps de déplacement (en min)
- hinc: revenu du ménage (1000$)
- psize: nombre de personnes se déplaçant ensemble

Dep. Variable:	car		**No. Observations:**		210	
Model:	Logit		**Df Residuals:**		206	
Method:	MLE		**Df Model:**		3	
Date:			**Pseudo R-squ.:**		0.8240	
Time:			**Log-Likelihood:**		-21.775	
converged:	True		**LL-Null:**		-123.76	
	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	9.7860	2.312	4.232	0.000	5.254	14.318
invt	-0.0469	0.009	-5.185	0.000	-0.065	-0.029
hinc	0.0291	0.020	1.460	0.144	-0.010	0.068
psize	-1.0656	0.536	-1.988	0.047	-2.116	-0.015

Tableau 2: Résultat d'un modèle logit

Veuillez répondre aux questions suivantes:

1. Décrire la qualité du modèle, s'il est significatif, et indiquer les variables significatives du modèle. Justifier. (1 pt)

2. Expliquer comment comparer les poids relatifs des différentes variables indépendantes. (0.5 pt)
3. Quel modèle permettrait d'étudier les facteurs associés au choix du mode de transport parmi au moins trois modes. (0.5 pt)
4. Discuter une méthode d'enquête permettant de recueillir de telles données (population de référence, type d'enquête et technique d'enquête). (1 pt)

Exercice 4 (données spatiales) 35 min (/4.5 pts)

Trois couches de données spatiales au format shapefile sont fournies pour cet exercice dans l'archive `donnees-spatiales.zip` (sur Moodle), soit les limites des zones de l'enquête OD de la grande région de Montréal de 2013 et deux couches de points. Veuillez répondre aux questions suivantes:

1. Comparer les ensembles de points des deux couches à l'aide de mesures de centralité et de dispersion (2 pts)
2. Décrire visuellement si les points des deux couches sont répartis de façon similaire: commenter le ou lesquels pourraient avoir une structure spatiale totalement aléatoire. (1 pt)
3. Décrire une méthode permettant de caractériser l'intensité des ensembles de points. (0.5 pt)
4. Décrire une procédure de traitement spatial pour caractériser les points selon les communes. (1 pt)

Solution

1. Il faut calculer le centre des points (disponible dans QGIS) et l'écart-type selon chaque dimension ou l'écart-type de la distance au centre (aucun outil ne semble disponible dans QGIS, une façon de faire est d'exporter les données en fichier csv pour traitement dans un outil externe comme Excel).
2. Les points ne sont pas répartis de façon similaire dans les deux couches de points fournies (points1 et points2). Les points de la couche « points2 » semblent répartis de façon uniformes selon les deux dimensions, ces points pourraient donc avoir une structure spatiale totalement aléatoire.
3. L'intensité est le nombre de points par unité de surface : il faut donc définir une partition de l'espace, puis faire le calcul pour chaque élément. On a vu trois exemples dans le cours, les quadrats (on fait une partition, puis il faudrait calculer l'intentité par quadrat), les cartes de chaleurs (similaire aux quadrats dans le découpage de l'espace, mais avec une fonctions qui peut lisser le nombre moyens de points en chaque case) et le diagramme de Voronoi (par définition, il y a un point par zone du diagramme, donc l'intensité y est l'inverse de la surface de chaque zone).
4. Pour caractériser les points par commune, il faut faire une jointure spatiale qui identifie pour chaque point la zone (commune) dans laquelle il se situe.

Exercice 5 (analyse spatiale) 30 min (/4 pts)

La figure 2 présente le nuage de points de Moran pour le nombre d'accidents par zone (arrondissement ou municipalité) de l'Île de Montréal pour l'année 2018 selon le critère de proximité de la tour ("Rook"). L'indice I de Moran global vaut 0.362 avec une pseudo valeur p de 0.001.

1. Est-ce que la variable du nombre d'accidents par zone présente une auto-corrélation spatiale? Justifier la significativité. Est-ce surprenant? (1.5 pt)
2. Décrire les quatre quadrants du nuage de points de Moran, en particulier lesquels démontrent une autocorrélation positive, et le lien avec l'indice I de Moran. (1 pt)

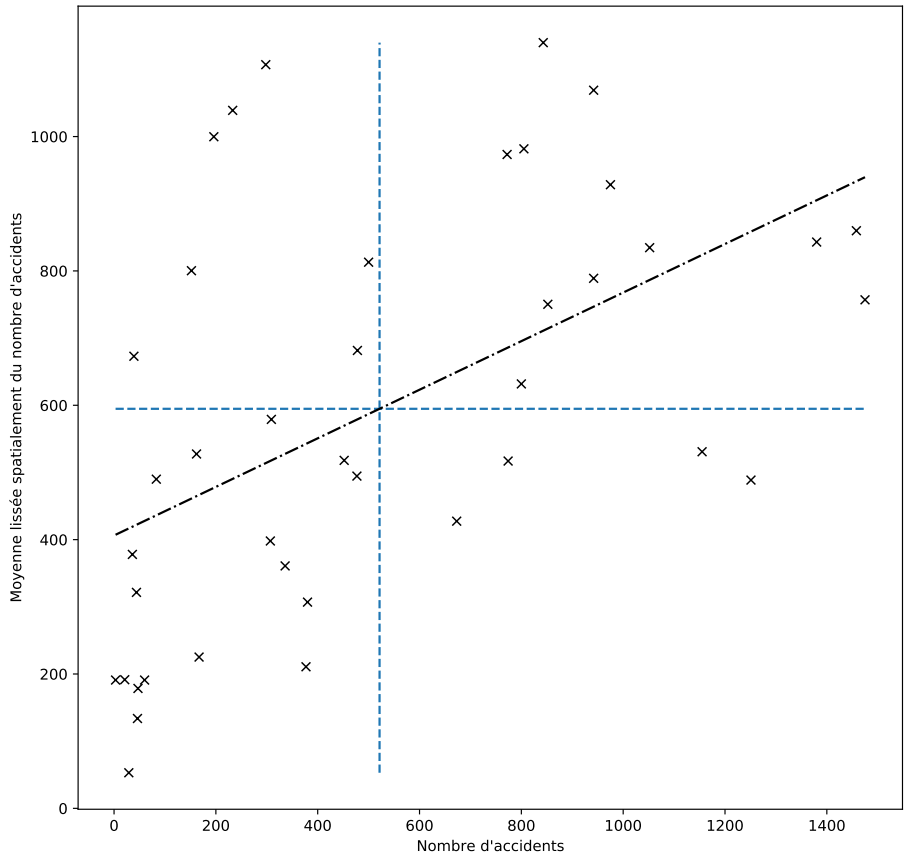

Figure 2: Nuage de points de Moran du nombre d'accident par zone de l'Île de Montréal.

3. Quelle est la différence avec le critère de proximité de la reine ("Queen")? Pensez-vous que l'indice I de Moran sera différent avec le critère de la reine? Justifier. (1 pt)

4. Quelle mesure permettrait de déterminer les zones avec des similarités ou dissimilarités particulières? (0.5 pt)

5. Expliquer comment la pseudo valeur p est obtenue (0.5 pt bonus)

Solution

1. Oui. La valeur p de 0.001 indique un petit risque de se tromper en rejetant l'hypothèse nulle d'absence d'auto-corrélation spatiale des résidus. Ce n'est pas surprenant puisque les accidents sont liés à des facteurs spatiaux, comme les aménagements routiers (limites de vitesse, longueur d'autoroutes, mesures d'apaisement de la circulation), qui sont aussi corrélés spatialement.

2. les quadrants en haut à droite et en bas à gauche montre une auto-corrélation positives (respectivement des zones avec un nombre d'accidents et un nombre lissé spatialement plus grands ou

plus bas que les moyennes respectives). Les deux autres quadrant dénotent des zones qui se démarquent de leurs voisins (nombre d'accidents plus élevé que la moyenne et nombre lissé plus bas (quadrant bas droite) ou réciproquement (quadrant haut gauche)). L'indice I de Moran est lié à la pente de la droite de régression du nuage de points de Moran.

3. les zones "en coin" seront aussi considérées commes voisines
   peu de différence attendue avec le critère queen (vu en classe)

4. L'indice de Moran I Local

5. Par simulation de Monte Carlo en faisant des permutations des valeurs du nombre d'accidents parmi toutes les zones
   Si les étudiants mentionnent que c'est obtenu par simulation, on pourrait donner le demi point