IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Fouille de données

Examen final– 21/02/2017

Nom : .....
Prénom : .....

Durée : 75mn.
Aucun document autorisé.
Aucun appareil (téléphone, calculatrice, etc.) autorisé.

Répondre directement sur les feuilles (il y a 9 pages et 20 questions ; **1 point** par question). Il sera tenu compte de la précision des explications ainsi que des justifications apportées lorsque nécessaire (restez concis). Écrivez lisiblement en évitant les ratures, etc.
Pour chacune des questions à choix multiples (partie I.), plusieurs réponses proposées peuvent être exactes. Vous devez cocher la ou les réponse(s) exacte(s) sans justification. Une bonne réponse complète rapporte **1 point**. Une réponse mauvaise, ou partiellement incomplète, enlève $\frac{1}{2}$ **point**. L'absence de réponse ne rapporte aucun point et n'enlève aucun point.

Corrigé

I. Questions à Choix Multiples

Q. 1 – Indiquez si les affirmations suivantes de votre ami ont du sens ou non.

J'ai terminé le premier exercice en 30 secondes et toi en 60 secondes. Cela t'a donc pris deux fois plus de temps que pour moi.

√ **oui**

◯ non

Solution: Peu importe ici où le temps 0 est fixé, il "existe" ne serait-ce qu'au début de l'examen, donc une échelle de ratio reste candidate. Par ailleurs les secondes sont formellement définies sur une échelle où la différence et les ratios font sens. Donc votre ami a raison.

Je suis donc deux fois plus intelligent que toi.

◯ oui

√ **non**

Solution: Là ... où est le 0 ? ... Donc affirmation fausse bien qu'utilisée dans le langage courant.

Dans la salle d'examen il faisait deux fois plus chaud qu'hier.

◯ oui.

√ **non.**

Solution: Là ... où est le 0 ... encore une fois. Cf. les différentes échelles de température. Donc votre ami a tord.

Q. 2 – Les étudiants étrangers de l'IMT Atlantique sont classés par le département Langue et Culture Internationale en : français (0) non lecteur, (1) lecteur débutant, (2) lecteur intermédiaire, (3) lecteur avancé. La classification est réalisée afin de placer les étudiants dans des groupes de niveau. L'échelle de mesure utilisée est :

Solution: Clairement il n'y a pas de niveau 0 bien q'un groupe soit nommé (0). Il n'y a pas non plus de moyen de dire que le groupe (1) est deux fois moins bon que le groupe (2). Les différents groupes sont en revanche organisés en niveaux de compétence en lecture du français. Le département Langue et Culture Internationale utilise une échelle ordinale.

◯ nominale

√ **ordinale**

◯ d'intervalle

◯ de ratio

Q. 3 – La Maisel veut connaître les chambres les plus populaires. Pour cela elle compte le nombre de demandes pour chaque chambre et classe les chambres en fonction du nombre de demandes. L'échelle de mesure utilisée est :

◯ nominale

√ **ordinale**

◯ de peintre

◯ d'intervalle

◯ de ratio

Solution: La Maisel veut connaître les chambres les plus populaires. Pour cela elle compte le nombre de demandes pour chaque chambre et **classe** les chambres en **fonction du nombre demandes**. On pourrait croire que derrière il y a une échelle de ratio (le 0 sur les demandes est clair) mais rien ne l'indique dans la procédure de la Maisel qui pourrait considérer le classement suivant : très faible (0-3 demandes), faible (4-6 demandes), normal (7-10 demandes), sur-demande (11-12 demandes), excellente demande (demandes >12). La Maisel utilise une échelle ordinale.

Q. 4 – La Classification Hiérarchique Ascendante

- √ **consiste à faire des regroupements d'individus qui se ressemblent selon leurs variables de description**
- ◯ est une méthode d'apprentissage supervisée qui regroupe les individus en classes homogènes
- ◯ nécessite que l'utilisateur fixe à l'avance le nombre de classes

Q. 5 – L'indice de la silhouette est un

- ◯ outil d'élagage des arbres de décision
- √ **indice de qualité d'un clustering**
- √ **moyen de déterminer le nombre de classes optimal dans un jeu de données**

Q. 6 – La technique des arbres de décision est une méthode d'apprentissage permettant de prévoir les valeurs prises par

- √ **une variable numérique**
- √ **une variable binaire**
- √ **une variable catégorielle**

> **Solution:** Méthodologiquement la méthode des arbres de décision divise de façon récursive les individus en fonction de la variable cible, peu importe qu'elle soit catégorielle ou continue. Les divisions sont évaluées selon la qualité des partitions obtenues (et selon différents critères). Les calculs ne sont pas faits sur les échelles des attributs prédictifs mais sur les partitions et donc les arbres de décision peuvent conceptuellement considérer tous types de variables cibles.

Q. 7 – À quoi servent les noyaux des SVM ?

- ◯ À simplifier les données pour n'en garder que les plus pertinentes
- ◯ À interpoler les données manquantes
- √ **À transformer les données pour que la séparation par hyperplan soit possible**
- ◯ À paralléliser les calculs en attaquant directement les noyaux des processeurs

Q. 8 – Dans les SVM, la notion de marge maximale renvoie à :

- √ **la maximisation de la distance entre la frontière de séparation et les individus les plus proches**
- ◯ une marge d'erreur sur la qualité de la solution fournie
- ◯ l'astuce du noyau

Q. 9 – Le taux d'erreur d'un modèle prédictif est toujours une bonne indication des performances du modèle

- ◯ oui
- √ **non**

Solution: On n'a pas précisé sur quel ensemble, mais sur l'ensemble d'apprentissage on sait que non (sans plus d'information c'est la réponse attendue), si vous précisez que c'est sur un ensemble de test alors la réponse oui pourrait être acceptable mais ce n'est pas la bonne car cela ne tient pas compte du déséquilibre de classe.

Q. 10 – La différence entre apprentissage supervisé et non supervisé se fonde sur une :

- ◯ hypothèse sur la distribution des données
- √ **connaissance ou non d'un label pour certains individus**
- ◯ hypothèses d'indépendance entre les individus

Q. 11 – Vous disposez de p variables qualitatives, chacune ayant 3 modalités, et de k variables continues. Vous devez appliquer les règles d'association sur des données ainsi décrites. Quelle(s) opérations devez-vous réaliser ?

- √ **discrétisation**
- ◯ ne rien faire
- √ **transformation disjonctive complète**

Q. 12 – Pour évaluer un modèle de risque du cancer du sein construit en population générale et dans laquelle les classes sont très déséquilibrées, quel type d'indicateur de performance vaut-il mieux utiliser ?

- √ **l'aire sous la courbe ROC, car l'indicateur est indépendant de la matrice de coût de mauvaise affectation**
- ◯ le taux d'erreur, car la matrice de coût de mauvaise affectation est unitaire
- √ **l'aire sous la courbe ROC, car l'indicateur est synthétique et aisément interprétable**

II. Questions / Réponses

Q. 13 – Quelle est la valeur de la mesure `confiance` à l'indépendance ?

Solution: La confiance d'une règle `A -> B` est p_{ab}/p_a. A l'indépendance $p_{ab} = p_a p_b$ donc la valeur à l'indépendance est p_p.

Q. 14 – Soient 10000 transactions dont 6000 avec paiement par `carte`, 6500 avec achat de `chocolat` et 3800 avec `carte` et `chocolat`.

Quelle est la confiance de la règle `carte=>chocolat` ?

Solution: La confiance de la règle `carte=>chocolat` est $3800/6000 = 63\%$.

Est-elle intéressante ? Justifiez votre réponse.

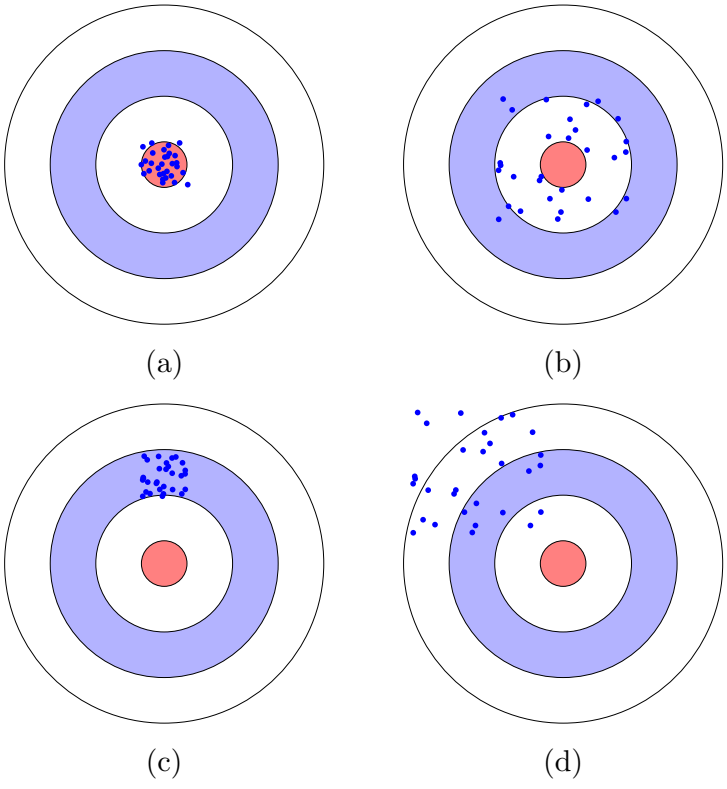
Solution:

A revoir car **La confiance de la règle** `carte=>chocolat` **est** $3800/6000 = 63\%$.. La fréquence de `chocolat` est $6500/10000 = 65\% > 63\%$.

La règle ne permet donc pas de "prédire" l'achat de chocolat (de ce point de vue pas intéressante) mais indique peut-être que le fait d'utiliser la carte joue négativement sur l'achat de chocolat (et donc de ce point de vue intéressante, bref un phénomène à creuser). Comme vous le voyez on peut répondre de deux façons, même si la 2ième réponse est préférable. C'est la cohérence de votre réponse et de la justification qui sera évaluée.

Q. 15 – On représente les performances d'un classifieur par le résultat d'un tir sur une cible. Le centre de la cible correspond à un classifieur parfait et plus on s'éloigne du centre plus ses performances sont mauvaises.

La figure suivante représente les résultats de 30 réalisations de quatre modèles –(a), (b), (c) et (d)– à partir de 30 ensembles d'apprentissage



Indiquez dans les crochets [ ] le classifieur correspondant à la description biais/variance.

Solution:

[(c)] Variance faible / biais élevé

[(b)] Variance élevée / biais faible

[(d)] Variance élevée / biais élevé

[(a)] Variance faible / biais faible

Justifiez votre réponse pour l'un des quatre classifieur au choix en précisant lequel.

Solution: Prenons le classifier (c) par exemple. Le tir est groupé donc peu dépendant à l'échantillon c'est-à-dire une variance faible. En revanche le tir est "loin" du centre (comparativement à (b), et (a) surtout) signe d'un biais élevé.

III. Solution(s) pour le RAK Mining

Le gérant du RAK vous demande de l'aider à comprendre les habitudes de consommation. Bref, il veut faire du RAK-Mining. Il met à votre disposition une base de données contenant 10 transactions.

	entrée	plat	accompagnement	dessert	café
1	carottes rapées	poisson	frites	glace	oui
2	carottes rapées	steak	frites	glace	non
3	salade tomates	steak	frites	gâteau	non
4	carottes rapées	steak	frites	glace	oui
5	carottes rapées	steak	frites	glace	non
6	carottes rapées	steak	frites	gâteau	non
7	salade tomates	steak	haricots verts	gâteau	non
8	salade tomates	steak	frites	gâteau	non
9	carottes rapées	poisson	riz	glace	oui
10	salade tomates	poisson	riz	glace	non

Q. 16 – Donnez les ensembles 1-fréquent, 2-fréquent, 3-fréquent, etc., ainsi que leur support pour un seuil minimal de support de $0,5$.

Solution:

Les 1-fréquents sont :

	items	support
1	accompagnement= frites	0.7
2	plat= steack	0.7
3	café= non	0.7
4	entrée=carottes rapées	0.6
5	dessert= glace	0.6

Les 2-fréquents sont :

	items	support
6	plat= steack,accompagnement= frites	0.6
7	plat= steack,café= non	0.6
8	entrée=carottes rapées,dessert= glace	0.5
9	entrée=carottes rapées,accompagnement= frites	0.5
10	accompagnement= frites,café= non	0.5

Les 3-fréquents sont :

	items	support
11	plat= steack,accompagnement= frites,café= non	0.5

Il y a donc 11 itemsets fréquents pour un seuil de support de 0.5.

Après lui avoir expliqué le principe de l'algorithme Apriori pour déterminer des règles d'association le gérant du RAK vous propose d'utiliser un seuil de confiance de 0,8.

On considère les deux règles suivantes :

R1 : `accompagnement = frites, café = non => plat = steak`

R2 : `entrée = carottes rapées => dessert= glace`

Q. 17 – Sont-elles des résultats du processus ? Justifiez votre réponse. Si oui donnez leur support, confiance et lift.

√ **oui.**

◯ non.

Solution:

R1 : `accompagnement = frites, café = non => plat = steak`
L'itemset qui supporte la règle est fréquent (cf. question précédente) de support 0.5. Sur les 5 transactions (0.5*10) contenant `accompagnement = frites, café = non` toutes contiennent également `plat = steak`. La confiance de la règle est donc 1.0. la règle R1 a donc un support et une confiance supérieurs ou égal aux seuils de support et de confiance fixés, c'est donc bien un résultat du processus. Son lift est $1.0/(7/10) \sim 1.42$.

R2 : `entrée = carottes rapées => dessert= glace`
Même raisonnement : support de 0.5, 5 transactions sur les 6 contenant `entrée = carottes rapées` contiennent également `dessert= glace`. La confiance de la règle est donc de 5/6 0.83. la règle R2 a donc un support et une confiance supérieurs ou égal aux seuils de support et de confiance fixés, c'est donc bien un résultat du processus. Son lift est $0.83/(6/10) \sim 1.38$.

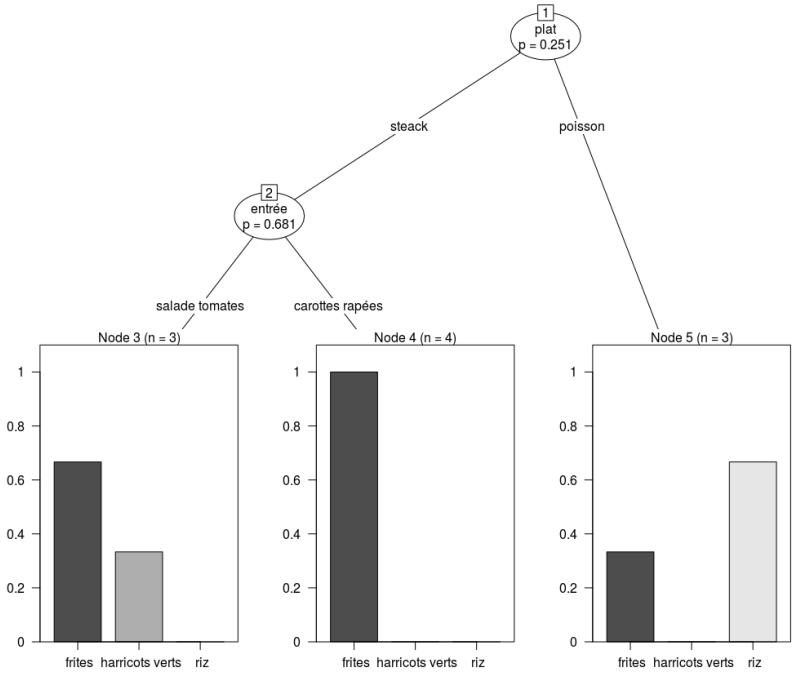
règle	support	confiance	lift
R1	0.5	1.0	1.42
R2	0.5	0.83	1.38

Q. 18 – Y a-t-il des règles ayant pour conséquent `café = non` ? Si oui lesquelles ? Justifiez votre réponse.

Solution: Il n'y a que 4 itemsets fréquents contenant `café=non` : `café= non`, `plat= steack,café= non`, `accompagnement= frites,café= non` et `plat= steack,accompagnement= frites,café= non`. Le premier, `café= non`, est un 1-itemset, il ne peut donc pas former de règle. Analysons donc les règles ayant pour conséquent `café=non` à partir des 3 itemsets restants :

- `accompagnement= frites,café= non` : seule la règle `accompagnement= frites => café= non` est candidate. Sa confiance est de 0.7 donc inférieure au seuil fixé et ce n'est pas une règle issue du processus.
- `plat= steack,café= non` : même raisonnement. La confiance de la règle `plat= steack => café= non` est de 0.85, c'est donc bien une règle issue du processus.
- `plat= steack,accompagnement= frites,café= non` : même raisonnement. La confiance de la règle `plat= steack,accompagnement= frites,café= non` est de 0.83, c'est donc bien une règle issue du processus.

Enfin le gérant est intéressé par prévoir l'`accompagnement`. Pour cela vous produisez l'arbre de décision suivant :



Q. 19 – Donnez la matrice de confusion de l'arbre en appliquant la règle majoritaire où chaque ligne correspond à la classe réelle de la variable `accompagnement`, chaque colonne à la classe prédite par l'arbre.

Solution:

- le nœud de gauche (n=3) classe en `frites` et il y a une erreur pour 1 exemple `haricots verts`

- le nœud du centre (n=4) classe en `frites` et il n'y aucune erreur
- le nœud de droite (n=3) classe en `riz` et il y a une erreur pour 1 exemple `frites`

	`frites`	`haricots verts`	`riz`
`frites`	6		1
`haricots verts`	1		
`riz`			2

Q. 20 – Le taux d'erreur que l'on peut calculer à partir de cette matrice est-il une bonne estimation du taux d'erreur réel ? Sinon que faudrait-il faire ?

Solution: Non, le taux d'erreur ainsi calculé sur l'ensemble d'apprentissage ne peut être une bonne estimation du taux d'erreur réel (votre réponse devrait être cohérente avec la question correspondante du QCM). Il faut mettre en place un protocole d'évaluation sur des données non vues (ensemble de test) par l'arbre ou bien une évaluation par validation croisée.

Q. 21 – À zéro point. Le cours de fouille de données a été (entourez la bomme réponse) :

Excellent – Très bien – Bien – Moyen – Nul.

Solution: Il n'y a pas de bomme réponse, au mieux une bonne réponse, mais pas vraiment une bonne réponse, plutôt la vôtre. Merci d'avoir répondu, en toute honnêteté on espère. Quelques commentaires qualitatifs viennent compléter les réponses quantitatives (sur une échelle ordinale). Merci. Il sera difficile de faire mieux cette année que ce mini-sondage et l'échange que nous avons eu en fin de cours. Mais qui sait...