ECUE21.3: Science des données **Examen corrigé**

1 juillet 2020.

---

1 Classification de tubes

Vous travaillez pour une entreprise qui fabrique des tubes de transport d'hydrocarbures. Ces équipements subissent un contrôle qualité rigoureux. Lors du contrôle, le tube est soumis à des tests physiques lors desquels de nombreuses mesures sont prises. Ces mesures sont ensuite analysées attentivement par des spécialistes. L'entreprise souhaite partiellement automatiser cette dernière étape.

Pour cela, on vous donne un jeu de données dans lequel 5 272 tubes sont décrits par 1 991 variables : 1 990 mesures prises lors des tests, et une variable binaire valant 1 si le tube passe le contrôle qualité et 0 sinon.

Votre tâche est d'utiliser ces données pour construire un modèle qui permettra de réduire le nombre de tubes qui devront être examinés par les spécialistes.

1. (1 point) De quel type de problème d'apprentissage statistique s'agit-il ?

   Solution: Apprentissage supervisé, classification binaire.

2. (2 points) Quel(s) critère(s) de performance allez-vous utiliser
   - pour choisir votre modèle ?
   - pour illustrer dans votre rapport final la performance que l'on peut attendre de votre modèle une fois en production?

   Justifiez vos réponses.

   Solution: Il est primordial d'avoir un minimum de faux positifs, les tubes étiquetés comme positifs ne passant pas de contrôle par un humain. Il serait donc raisonnable d'utiliser la *précision* (ou PPV) comme critère de sélection. Vous risquez cependant de choisir un modèle inutile qui prédit que tous les tubes sont positifs. Un score plus équilibré, par exemple le f1, est donc aussi une bonne idée. Pour votre rapport, la PPV sera encore une fois importante, mais vous pouvez vous permettre de rapporter plusieurs critères : vous pouvez mettre toute la table de confusion.

3. (2 points) Vous avez dressé avec vos collègues une liste des algorithmes d'apprentissage à essayer. Décrivez comment vous allez utiliser les données pour construire votre modèle et évaluer la performance que vous pouvez en attendre une fois déployé. Précisez le nombre d'éléments du jeu de données utilisé à chacune des étapes que vous décrivez.

Solution: Train/test split + validation croisée ou train/test de nouveau sur le train.

Par exemple 70/30 pour le train/test et 10-fold CV : 3690 observations dans le train, 1582 dans le test. Chaque étape de la validation croisée consiste à entrainer sur 3321 observations et évaluer sur 369.

4. (1 point) Le premier algorithme que vous testez est une régression logistique. Vous observez une bonne performance sur les observations que vous utilisez pour l'entraînement, mais quand vou appliquez votre modèle sur d'autres observations, cette performance chute de 15%. Est-ce surprenant ? Que se passe-t-il, et comment remédiez-vous à ce problème ?

Solution: Surapprentissage. Peu surprenant car beaucoup de variables, même si plus d'observations. Il faut régulariser.

5. (1 point) Vous essayez ensuite une forêt aléatoire. Sa performance est bien plus modeste que celle de la régression logistique. Que se passe-t-il, et comment remédiez-vous à ce problème ?

Solution: Sous-apprentissage. Augmenter le nombre d'arbres !

6. (3 points) Votre collègue, qui travaille aussi sur le problème, vous envoie ses résultats sous la forme de la figure 1. Suggérez trois façons d'améliorer cette représentation, en expliquant pourquoi vous faites ces suggestions.

Solution: Quelques suggestions...

- Légendes parlantes – quels sont les modèles ?
- Performance : quelle métrique ?
- Faire commencer les barres à 0 – proportional ink.
- Toutes les barres de la même couleur.
- Toutes les barres avec le même espacement.
- Axe des ordonnées suffisament grand pour que l'on voie la barre d'erreur en entier.
- Barres d'erreurs d'une couleur qui contraste avec les barres.

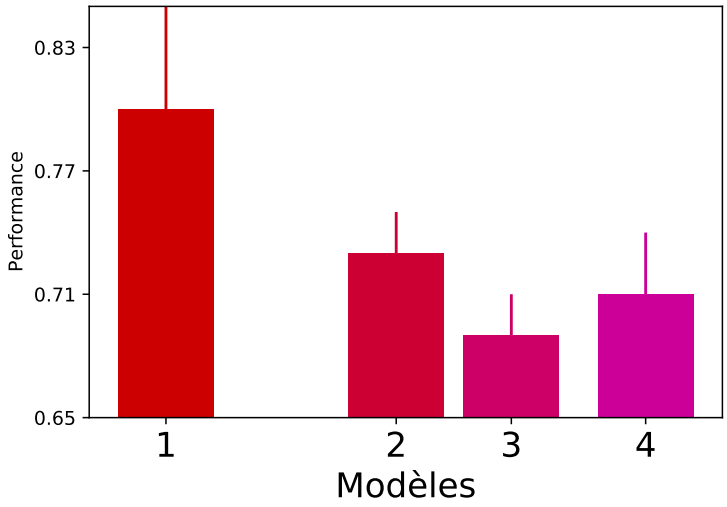


Figure 1: Résultats envoyés par votre collègue (Problème 1).

2 Reconnaissance faciale

Un des problèmes les plus étudiés dans la classification automatique des images de visages est celui de la reconnaissance du genre, généralement présenté comme un problème de classification binaire avec les étiquettes « homme » et « femme ». Une publication récente[1] présente les performances de 3 modèles commercialisés pour effectuer cette tâche.

La table 1, qui en est extraite, présente les performances de ces 3 modèles selon le genre et la couleur de peau des individus, décrite comme « claire » ou « foncée » en fonction de leur phototype[2]. Les modèles 1 et 3 sont commercialisés par des entreprises américaines et le modèle 2 par une entreprise chinoise.

Le jeu de données utilisé pour l'évaluation, appelé *Pilot Parliament Benchmark* (ou PPB) et construit à partir de photos de membres de parlements à travers le monde, contient 1 270 images. Sa composition est décrite dans la table 2. La composition du jeu de donnée *Adience*, utilisé depuis 2014 pour le benchmarking d'outils de classification automatique des images de visage par genre, est elle donnée dans la table 3.

Modèle	Total	DF	DM	LF	LM
1	78	55	18	5	0
2	123	89	2	29	3
3	154	94	38	21	1

Table 1: Nombre d'images mal classifiées pour 3 modèles de classification d'images de visage par genre, évalués sur le jeu de données PPB. Les performances sont données au global (« Total ») puis parmi les femmes à la peau foncée (« DF »), les hommes à la peau foncée (« DM »), les femmes à la peau claire (« LF »), les hommes à la peau claire (« LM »).

Total	DF	DM	LF	LM
1270	271	318	296	385

Table 2: Composition du jeu de données PPB : nombre de femmes à la peau foncée (« DF »), nombre d'hommes à la peau foncée (« DM »), nombre de femmes à la peau claire (« LF »), nombre d'hommes à la peau claire (« LM »).

Total	DF	DM	LF	LM
2194	162	140	979	913

Table 3: Composition du jeu de données Adience : nombre de femmes à la peau foncée (« DF »), nombre d'hommes à la peau foncée (« DM »), nombre de femmes à la peau claire (« LF »), nombre d'hommes à la peau claire (« LM »).

7. (1 point) Pourquoi chercher à construire un outil de classification automatique d'images de visages par genre ? Quelles applications pratiques y voyez-vous ?

Solution: Personalisation d'interactions/contenus. Une intelligence humaine perçoit le genre, pourquoi pas une IA. Première étape d'une reconnaissance faciale ?

8. (1 point) Comparez la performance globale des modèles sur le jeu de données PPB à leur performance sur chacun des sous-groupes DF, DM, LF et LM. Qu'observez-vous ? À quoi cela peut-il être dû ?

Solution: Taux d'erreurs :

1	6.3%	20.3%	5.7%	1.6%	0.0%
2	9.7%	32.8%	0.6%	9.8%	0.8%
3	12.1%	34.7%	11.9%	7.1%	0.3%

Le taux d'erreur est largement plus élevé chez les femmes noires que dans les autres catégories.

9. (1 point) En supposant que les taux d'erreurs sur chacun des sous-groupes DF, DM, LF et LM soient conservés, à quelle proportion d'erreurs globale vous attendez-vous sur Adience pour chacun des 3 modèles ? Comparez ces résultats à ceux obtenus sur PPB.

Solution:

- Modèle 1 : $(0.203 \times 162 + 0.057 \times 140 + 0.016 \times 979)/2194 = 2.6\%$

- Modèle 2 : $(0.328 \times 162 + 0.006 \times 140 + 0.098 \times 979 + 0.008 \times 913)/2194 = 7.1\%$
- Modèle 3 : $(0.347 \times 162 + 0.119 \times 140 + 0.071 \times 979 + 0.003 \times 913)/2194 = 6.6\%$

Les modèles semblent bien plus performants sur Adience que sur PPB.

10. (2 points) Le 8 juin 2020, IBM a annoncé se retirer du marché de la reconnaissance faciale avec ces mots : « IBM s'oppose fermement et ne tolérera pas l'utilisation de toute technologie, y compris la technologie de reconnaissance faciale proposée par d'autres fournisseurs, pour la surveillance de masse, le contrôle au faciès, les violations des droits de l'Homme et des libertés fondamentales ou tout autre objectif qui ne serait pas conforme à nos valeurs et à nos principes de confiance et de transparence. » Commentez cette décision au regard des informations données.

3 Réseau de neurones artificiel

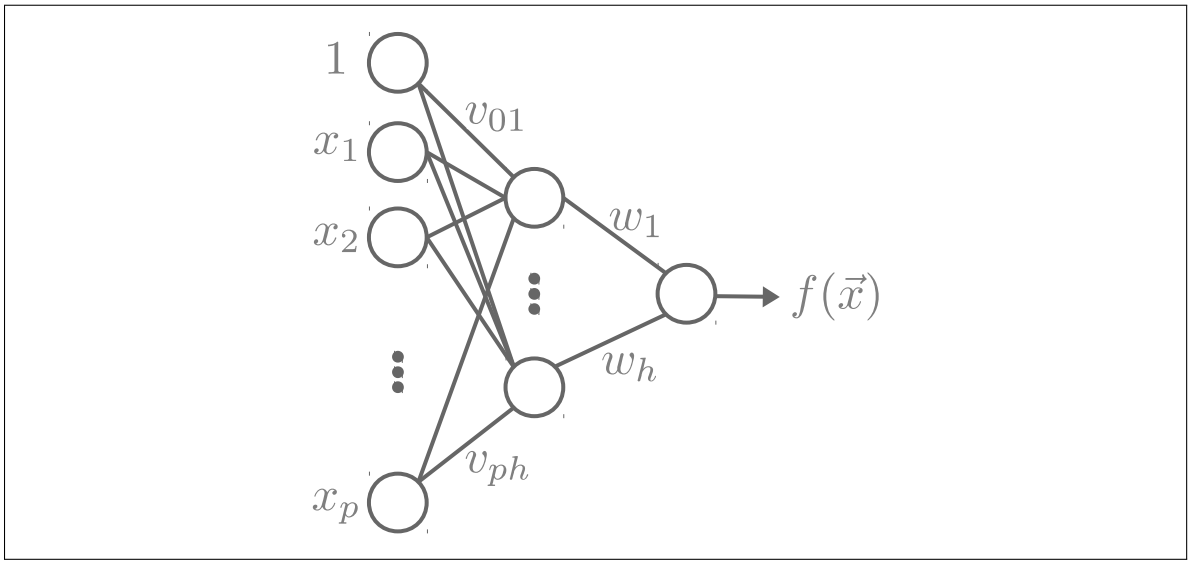
Soit $\mathcal{D} = \{(\vec{x}^{\,i}, y^i)\}_{i=1,\dots,n}$ un jeu d'apprentissage de de n observations en p dimensions, et leurs étiquettes réelles. Nous souhaitons entraîner un perceptron multi-couche à une couche cachée, contenant h neurones, sur ces données.

11. (1 point) Quelle fonction d'activation choisir pour l'unité de sortie et pourquoi ?

Solution: Il s'agit d'un problème de régression, donc l'identité.

12. (1 point) Dessinez l'architecture d'un tel réseau de neurones. Vous pouvez choisir d'ajouter une ou des unités de biais.

Solution: Exemple avec une unité de biais en entrée mais pas sur la couche intermédiaire.



13. (1 point) Quelle est l'espace des hypothèses $\mathcal{F}$ du réseau de neurones que vous venez de dessiner ? Vous vous fixerez (pour clarifier l'écriture, il n'y a pas de bon ou mauvais choix) une fonction d'activation pour les couches cachées parmi
    - la fonction logistique ;
    - la fonction tangente hyperbolique ;
    - la fonction dite ReLU (*Rectified Linear Unit*), à savoir $u \mapsto \max(u, 0)$.

Solution:

$$\mathcal{F} = \left\{ \vec{x} \mapsto \sum_{q=1}^{h} w_q \, a\left(\sum_{j=1}^{p} v_{jq} x_j + v_{oq} \right) ; w_1, \ldots, w_h, v_{01}, \ldots v_{0h}, v_{11}, \ldots, v_{ph} \in \mathbb{R} \right\},$$

en remplaçant a par la fonction d'activation choisie. (Ici pour une architecture avec unité de biais en entrée mais pas sur la couche intermédiaire.)

14. (1 point) Peut-on parler ici de modèle paramétrique ? Si oui, à combien de paramètres et quels sont-ils ?

Solution: Oui. Les paramètres sont $w_1 \ldots, w_h, v_{01}, \ldots v_{0h}, v_{11}, \ldots, v_{ph}$ et sont donc au nombre de $h + (p+1)h = (p+2)h$.

15. (1 point) En choisissant une fonction de coût raisonnable, formulez le choix d'un modèle dans $\mathcal{F}$ sous la forme d'un problème de minimisation du risque empirique.

Solution:

$$\underset{\vec{w}\in\mathbb{R}^h,\vec{v}\in\mathbb{R}^{h(p+1)}}{\arg\min} \sum_{i=1}^{n}\left(\sum_{q=1}^{h} w_q\, a\left(\sum_{j=1}^{p} v_{jq}x_j^i + v_{oq}\right) - y^i\right)^2 .$$

16. (2 points) Comment résoudre ce problème d'optimisation ? Quelles conséquences cela a-t-il sur le choix du modèle ? Pourquoi choisir néanmoins d'utiliser ce réseau de neurones ?

Solution: Le pb n'est pas convexe : pas de solution exacte et encore moins unique. On utilise des descentes de gradient mais aucune de garantie de converger vers le minimum global. Différentes initialisations, pas de descente etc donneront des solutions différentes. Mais plus grande puissance de modélisation.

4 Émission de composés organiques par des plantes

Lorsqu'elles sont endommagées, les plantes vertes émettent des composés organiques, appelés substances volatiles des feuilles (ou GLV pour *green leaf volatiles* en anglais). Ces molécules, qui sont responsables, entre autres, de l'odeur de l'herbe coupée, sont considérées comme un mode de communication entre plantes.

La figure 2 est tirée d'une publication récente[3]. Dans cette étude, les scientifiques ont cherché à observer l'effet des sécrétions orales de certaines chenilles sur les émissions de GLV. L'étude s'intéresse à 4 cas particuliers de GLV : les molécules (Z)-3-hexenal, (E)-2-hexenal, (Z)-3-hexenol, et (Z)-hexenyl acetate.

La figure 2 présente les quantités de ces 4 GLV émises par plante de maïs (*maize*) endommagée. L'exérience est conduite sur 4 plantes (les valeurs affichées sont donc moyennées) et sous 3 traitements:

- la plante a été exposée à une solution contrôle (*PBS*), supposée n'avoir aucun effet ;
- la plante a été exposée à des sécrétions orales de chenille (*Fresh Regurgitant*) ;
- la plante a été exposée à des sécrétions orales de chenille bouillies (*Boiled Regurgitant*).

Pour chacune des 4 molécules GLV et pour chacun des deux traitements par des sécrétions orales, un test t de comparaison de moyennes a été réalisé pour comparer les quantités de cette molécule émises sous ce traitement à celles émises sous le traitement contrôle (PBS). Les astérisques indiquent une différence statistiquement significative entre les deux valeurs.

17. (1 point) Considérons la comparaison des émissions de (E)-2-hexenal entre le contrôle (PBS) et l'exposition aux sécrétions orales de chenille non bouillies (Fresh Regurgitant). Quelle est l'hypothèse nulle du test t ?

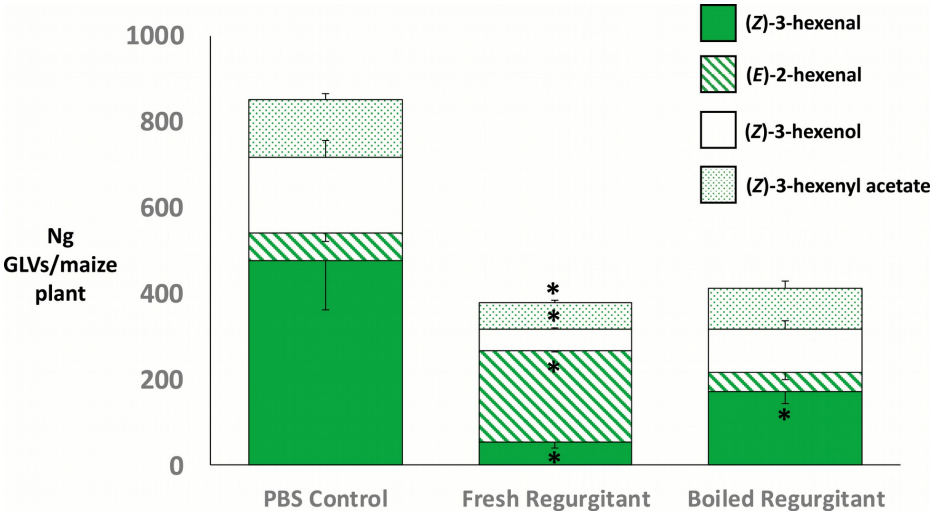


Figure 2: Émissions moyennes de (Z)-3-hexenal, (E)-2-hexenal, (Z)-3-hexenol, et (Z)-hexenyl acetate par des feuilles de maïs endommagées puis traitées avec une solution contrôle (PBS) et des sécrétions orales de chenille fraîches et bouillies.

Solution: Il n'y a pas de différence de moyenne entre la quantité de GLV émise dans l'expérience témoin et celle émise sous traitement.

18. (1 point) Quelle est l'hypothèse alternative de ce test ? S'agit-il d'un test unilatéral ou bilatéral et pourquoi ?

Solution: Il y a une différence de moyenne entre la quantité de GLV émise dans l'expérience témoin et celle émise sous traitement. Test bilatéral car on ne sait pas si on s'attend à une augmentation ou une diminution des émissions.

19. (1 point) Pourquoi avoir utilisé un test t plutôt qu'un test Z ?

Solution: Pas assez d'échantillons.

Considérons maintenant toujours uniquement le (E)-2-hexenal, mais dans les 3 conditions expérimentales.

20. (1 point) Quelle conclusion peut-on tirer de cette figure ?

Solution: La salive de chenille a un effet sur ce GLV uniquement quand elle n'est pas bouillie.