Campus Ynov Aix-en-Provence - M1 IA/Data

Module « Data science »

Février 2026

Nom : _____ Prénom : _____

Cochez les affirmations vraies. Il peut y avoir plusieurs réponses correcte pour une question. Pour info une des questions n'a aucune réponse correcte.

Cotation : +1 par réponse correcte, malus de -1 par réponse fausse avec un plafond de 25% des réponses. Interdit de cocher toutes les cases d'une question (malus supplémentaire de -1).

---

1. Parmi les éléments suivants, lesquels sont considérés comme des données personnelles au sens du RGPD ?

☐ A. Une adresse IP

☐ B. Un numéro de client interne

☐ C. La température moyenne d'une ville

☐ D. Une combinaison code postal + date de naissance + sexe

- Réponses : A, B, D

2. Un Data Scientist utilise un dataset public pour entraîner un modèle de scoring crédit. Le dataset contient l'âge, le code postal et le statut d'emploi. Aucune variable "origine ethnique" n'est présente. Quelle affirmation est correcte ?

☐ A. Le modèle ne peut pas être discriminatoire puisqu'aucune donnée sensible n'est utilisée

☐ B. Ce modèle risque d'induire un biais discriminatoire

☐ C. L'utilisation d'un dataset public dispense de toute obligation RGPD

☐ D. Seules les données directement identifiantes posent problème pour la conformité

- Réponses : B. (Proxy : on peut déduire une catégorie sociale depuis le code postal p. ex. ou la combinaison des variables)

3. Soit le noyau suivant :

$$K = \begin{vmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \end{vmatrix}$$

```
|  0  -1   0 |
```

Quel est l'effet de ce filtre ?

☐ A. Flou gaussien ☐ B. Détection de contours

☐ C. Accentuation de la netteté (sharpen) ☐ D. Effet de relief (emboss)

- Réponse C : coeff. négatifs autour d'un centre amplifié → augmentation des différences locales + somme = 1 → préservation de la luminosité

4. Quelle est la différence entre le top-hat et le black-hat en morphologie mathématique ?

☐ A. Le top-hat réalise l'opération (Original − Ouverture), le black-hat réalise l'opération (Fermeture − Original)

☐ B. Le top-hat utilise l'ouverture, le black-hat utilise la fermeture, mais les deux extraient des détails de même nature

☐ C. Le top-hat extrait les petits éléments clairs, le black-hat extrait les petits éléments sombres

☐ D. Ils sont identiques mais appliqués dans des ordres différents

- Réponses : A et C

5. Quelle est la principale différence entre un processus AR(p) et un processus MA(q) ?

☐ A. AR utilise les valeurs passées de la série, MA utilise les erreurs passées

☐ B. AR modélise la tendance, MA modélise la saisonnalité

☐ C. AR est toujours stationnaire, MA ne l'est jamais

☐ D. AR a une mémoire infinie, MA a une mémoire finie de q périodes

- Réponses : A car :
  - **AR(p)** : $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t$ utilise les **valeurs passées** Y_{t-1}, Y_{t-2}, etc.
  - **MA(q)** : $Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$ utilise les **erreurs passées** ε_{t-1}, ε_{t-2}, etc.
- D car :

- **AR(p)** : Mémoire **infinie** car les valeurs passées dépendent elles-mêmes de valeurs encore plus anciennes. Y_t dépend de Y*{t-1}, qui dépend de Y*{t-2}, qui dépend de Y_{t-3}, etc. D'où l'ACF décroît **exponentiellement** (jamais de coupure nette)

- **MA(q)** : Mémoire **finie de q périodes.** Un choc εt *affecte* Y_t, $Y\{t+1\}$, ..., $Y\_\{t+q\}$ Après q périodes, le choc disparaît complètement et l'ACF a une **coupure nette** au lag q

- B est fausse car ni AR ni MA ne modélisent spécifiquement tendance ou saisonnalité (c'est le rôle de d et de la partie saisonnière)

- C est fausse car AR peut être non-stationnaire si $|\varphi| \geq 1$, et MA est toujours stationnaire (si θ respecte la condition d'inversibilité)

6. Vous modélisez une série de ventes mensuelles et obtenez les résultats suivants :

Modèle	**AIC**	**P-value (Ljung-Box)**	**MAPE (forecast)**
ARIMA(1,1,1)	450	0.15	8.5%
SARIMA(0,1,1)(0,1,1)[12]	420	0.42	6.2%
SARIMA(1,1,1)(1,1,1)[12]	415	0.08	6.0%

Quel modèle choisissez-vous et pourquoi ?

☐ A. ARIMA(1,1,1) car il est le plus simple avec la meilleure prédiction

☐ B. SARIMA(0,1,1)(0,1,1)[12] car meilleur AIC et bon test Ljung-Box

☐ C. SARIMA(1,1,1)(1,1,1)[12] car meilleur AIC

☐ D. On ne peut pas décider avec ces informations

- Réponse : B car :

- SARIMA(0,1,1)(0,1,1)[12] a un AIC faible (même si pas le plus faible)

- SARIMA(0,1,1)(0,1,1)[12] a un test Ljung-Box clairement non significatif donc qualité des résidus ok (indicernables d'un bruit blanc)

- SARIMA(0,1,1)(0,1,1)[12] : 6.2% pas la pire en performance

- ENSEMBLE ces critères placent ce modèle en meilleure position (les autres ne cochent pas toutes ces cases ou sont moins bien placés)

7. Lors d'un entretien de cadrage, un responsable marketing vous dit : "On veut de l'IA pour mieux connaître nos clients." Après plusieurs questions, vous découvrez que son vrai problème est que les campagnes emailing ont un taux d'ouverture de 5% et qu'il est sous pression de sa direction. Quelle problématique data science est la plus pertinente ?

☐ A. Clustering de clients pour créer des personas marketing

☐ B. Système de recommandation produits personnalisé

☐ C. Modèle de scoring pour cibler les clients les plus susceptibles d'ouvrir les emails

☐ D. Dashboard de connaissance client 360°

- Réponse : C

8. Soit les vecteurs u = [2, -1, 3] et v = [1, 4, -2]. Quelle est la norme euclidienne de u + v ?

☐ A. √22 ☐ B. √26 ☐ C. 5 ☐ D. √30

- Réponse : √(2+1)**2 +(-1+4)**2 + (3-2)**2) =√(9 + 9 + 1) = √19 donc aucune proposition n'est correcte

9. L'inertie intra-cluster mesure :

☐ A. La distance entre les centroïdes de différents clusters

☐ B. La somme des distances au carré de chaque point à son centroïde de cluster

☐ C. Le nombre de points dans chaque cluster

☐ D. La variance entre les clusters

- Réponse

10. L'ACP possède les propriétés suivantes :

☐ A. L'ACP centre automatiquement les données

☐ B. Les composantes principales sont orthogonales entre elles

☐ C. L'ACP peut augmenter le nombre de dimensions

☐ D. L'ACP est une transformation linéaire

- Réponses : A (condition d'application de l'ACP), B (par déf.) et D (projection sur vecteurs propres). Et au contraire de C, on utilise l'ACP pour réduire les dimensions en négligeant les dimensions où la variance est réduite.

11. Dans la formule du seuillage d'Otsu, on cherche à minimiser :

☐ A. La variance inter-classe ☐ B. La variance intra-classe

☐ C. La moyenne des intensités ☐ D. Le nombre de pixels blancs

- Réponse B

12. Dans l'ACP, la première composante principale est :

☐ A. La variable avec la plus grande variance

☐ B. Le vecteur propre associé à la plus grande valeur propre de la matrice de covariance

☐ C. La moyenne des variables ☐ D. Le premier vecteur de la base canonique

- Réponse B

13. Parmi les données suivantes, lesquelles sont considérées comme des données sensibles (catégories particulières) au sens de l'article 9 du RGPD ?

☐ A. Le salaire d'un employé ☐ B. L'appartenance syndicale

☐ C. L'orientation sexuelle ☐ D. Le numéro de sécurité sociale

- Réponses : B et C

14. Soit A une matrice 3×2 et B une matrice 2×4. Quelle est la dimension du produit AB ?

☐ A. 3×4 ☐ B. 2×2 ☐ C. 3×2 ☐ D. Le produit n'est pas défini

- Réponse A. Pour AB : (m×n) × (n×p) = (m×p), ici : (3×2) × (2×4) = (3×4). Règle : le nombre de colonnes de A doit égaler le nombre de lignes de B

15. Quelle méthode utilise-t-on couramment pour choisir le nombre optimal de clusters k dans K-means ?

☐ A. Méthode du gradient ☐ B. Méthode du coude

☐ C. Validation croisée ☐ D. Test t de Student

- Réponse B

16. Soit un cluster contenant 3 points : [0, 0], [2, 0], [0, 2]. Quel est le centroïde de ce cluster ?

☐ A. [1, 1] ☐ B. [2/3, 2/3] ☐ C. [0, 0] ☐ D. [1, 0]

- Réponse B (([0+2+0]/3, [0+0+2]/3) = (2/3, 2/3))

17. Quelle affirmation est vraie concernant la relation entre ARIMA et ARMA ?

☐ A. ARIMA est toujours plus performant qu'ARMA

☐ B. ARIMA(p, 0, q) est mathématiquement équivalent à ARMA(p, q)

☐ C. ARMA ne peut pas modéliser de séries avec tendance, contrairement à ARIMA

☐ D. ARIMA nécessite toujours au moins une différenciation (d ≥ 1)

- Réponses : B (le paramètre du milieu correspond à la différenciation = 0 signifie pas de différenciation comme pour ARMA) et C : si pas de différenciation pour que la série soit stationnaire il ne faut pas qu'elle ait de tendance.

18. Dans l'égalisation d'histogramme, la LUT (Look-Up Table) est calculée à partir de :

☐ A. La moyenne des intensités ☐ B. L'histogramme cumulé (CDF)

☐ C. La dérivée de l'histogramme ☐ D. Le maximum de l'histogramme

- Réponse : B. La LUT est définie par f(i) = 255 × C(i) où C(i) est l'histogramme cumulé normalisé (CDF).

19. Quelle opération morphologique permet de supprimer les petits objets blancs (bruit) tout en préservant les grands objets ?

☐ A. Dilatation ☐ B. Érosion ☐ C. Ouverture ☐ D. Fermeture

- Réponse C. L'ouverture (érosion puis dilatation) supprime les petits objets blancs plus petits que l'élément structurant.

20. Une matrice de covariance Σ de dimension 4×4 contient :

☐ A. 4 variances et 6 covariances uniques ☐ B. 4 variances et 12 covariances

☐ C. 16 covariances ☐ D. 4 variances et 16 covariances

- Réponse A, techniquement la B est correcte aussi (pas de malus si seule A est sélectionnée)