Fiche d'auto-évaluation 12

April 20, 2022

1. Supposez de faire de la détection d'anomalies sur un jeu de données en utilisant l'algorithme des K-moyennes (K-means clustering). Pour choisir le nombre K de clusters, vous appliquez l'algorithme sur toutes les données, sans les diviser en données d'entraînement et de test, vous regardez le silouhette-score moyen et vous choisissez le K qui le maximise. Est-il légitime de faire cela? Ou aurait-on dû appliquer l'algorithme des K-moyennes sur les données d'entraînement seulement?

2. Écrivez la formule de la silhouette-score et de l'inertie (inertia or total within-cluster variation). Quel est l'avantage de l'une par rapport à l'autre?

3. Supposez de faire du clustering d'un jeu de données constitué des 4 points suivants: A=(1,1), B=(2,1), C=(2,2), D=(3,3). Représentez les points sur un graphe x-y et écrivez les étapes de l'algorithme des K-moyennes. Pour chaque étape, écrivez et représentez dans le graphe comment les centroides changent.

4. Écrivez la formule de l'importance des caractéristiques (feature importance) pour une forêt aléatoire (random forest).

5. Supposez d'utiliser une forêt aléatoire pour faire de la régression. Supposez de faire varier les nombres d'arbres et de mesurer l'erreur sur les données de test. Comment varie l'erreur par rapport au nombre des arbres? Dessinez une courbe hypothétique, qu'on pourrait observer avec des données réelles, avec x=nombre d'arbres, y=erreur.

6. Écrivez la formule de l'entropie, de l'index Gini d'impureté (Gini impurity index) et de gain d'information (information gain).

7. Si on entraîne un réseaux de neurones en Keras, avec plusieurs "seeds", obtient-on les mêmes poids? Pourquoi, exactement?

8. Pourquoi est-il important de répéter l'algorithme des K-moyennes, avec plusieurs initialisations des centroides? Après toutes les répétitions, quels sont les clusters qu'on sélectionne à la fin?

9. Quelle est la différence entre des technique de sélection de caractéristiques (feature selection - donnez un exemple de technique vue dans le cours) et de réduction de la dimensionnalité?

10. La réduction de la dimensionnalité, est-elle une technique supervisée ou non-supervisée?

11. À quoi peut-être utile la réduction de la dimensionnalité?

12. Comment peut-on utiliser la réduction de la dimensionnalité pour la détection des anomalies?

13. Quand on applique l'analyse en composantes principales, comment choisit-on le nombre de composantes à retenir?

14. Écrivez la formule de décomposition en valeurs singuliers (Singular Value Decomposition) d'une matrice de donnée $\mathbf{A}$. Écrivez les propriétés de chaque terme de la formule.

15. Si on fait une décomposition en valeurs singuliers (Singular Value Decomposition) sur une matrice de données et on met en graphique le nombre de composantes vs. la variance capturée par ces composantes, quel est la forme du graphe que vous obtenez? Dessinez-le, approximativement.

16. Prenez une matrice de données $\mathbf{A}$ et appliquez une décomposition en valeurs singuliers (Singular Value Decomposition). Supposons que $\mathbf{A}^{(k)}$ est la matrice réduite, obtenue en retenant seulement k composantes. A quoi est égale la différence entre $\mathbf{A}$ et $\mathbf{A}^{(k)}$, mesurée par la norme de Froebenius, c'est à dire $||\mathbf{A} - \mathbf{A}^{(k)}||_F$?

17. Considérez un jeu de données où un veut faire de la classification binaire. Seriez vous capable de générer un tel jeu de données de façon à ce que n'importe quel modèle de classification aie une précision très basse (pas plus que 50% environs)?