AIML 427 Big Data

Test – 2024 T1

Closed BOOK   Time allowed: 50 MINS

Question 1. Big Data Basics [4 marks]

1.1 (3 marks) Briefly define 3*Vs* in Big Data.

1.2 (1 marks) State the main difference between *a test set* and *a validation set*.

Question 2. Feature Manipulation [17 marks]

2.1 (3 mark) Briefly describe the "*curse of dimensionality*" and briefly explain why it can be addressed by *feature selection*.

2.2 (2 marks) Briefly state *two challenges* in achieving feature construction.

2.3 (3 marks) *Pearson Correlation Coefficient (PCC)* is a widely used measure for *filter feature selection*. Briefly explain one way of using *PCC* to evaluate the *relevance* and *redundancy* of a *feature subset*.

2.4 (4 marks) *Sequential forward selection (SFS)* is a common feature selection approach.

(1) Is *SFS* a *feature subset selection* or *feature ranking method?* Justify your answer.

(2) *Nesting problem* is the main limitation of SFS. Briefly describe the *Nesting problem* and briefly explain why *"Plus-L, minus-R" Selection* can address the *Nesting problem.*

2.5 (3 marks) Principal Component Analysis (PCA) is a feature construction technique.

(1) Is PCA a *filter*, *wrapper*, or *embedded* method? Justify your answer.

(2) State *one limitation* of PCA.

2.6 (2 marks) *Multi-tree GP* can be used to construct multiple features for a classification problem. *Multi-tree GP* has two main representations: *class-dependent* and *class-independent.* Briefly describe the two representations.

Question 3. Manifold Learning [8 marks]

3.1 (3 marks) Is *Geodesic distance* the same as *Euclidean distance*? Justify your answer.

3.2 (3 marks) *Multidimensional Scaling (MDS)* approaches can be divided into two main categories: *Metric MDS* and *Non-metric MDS.*

(1) State the main difference between *Metric MDS* and *Non-metric MDS.*

(2) Among *Metric MDS* and *Non-metric MDS,* which one is more sensitive to outliers? Justify your answer.

3.3 (2 marks) *t-SNE* is a common manifold learning algorithm. State *two main limitations* of *t-SNE.*

Question 4. Clustering [6 marks]

4.1 (2 mark) List two differences between *hierarchical* and *partition-based* clustering methods.

4.2 (2 marks) What is a *dendrogram*?

4.3 (2 mark) Briefly state *two limitations* of k-means clustering.

Question 5. Regression [15 marks]

5.1 (5 marks) *Lasso regression* aims to minimise the following function on a given set of data/observations:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

(1) Briefly state the *purpose* of the penalty term in Lasso.

(2) What will happen if $\lambda \rightarrow 0$? What will happen if $\lambda \rightarrow \infty$?

(3) Can Lasso regression do *feature selection*? Briefly justify your answer.

5.2 (2 marks) Briefly describe what the *collinearity issue* is and briefly state *two methods* to address the issue.

5.3 (2 marks) List one *regression splines* we discussed in the lectures. Briefly *explain* the term *knot* in the context of regression splines.

5.4 (2 marks) What is a *generalized additive model* (GAM), and briefly describe how do splines integrate into it?

5.5 (4 marks) *Model selection* aims to find a best balance between bias and variance.

(1) If a model *overfits* the training data, is it more likely to have *high bias* or *high variance*? Briefly justify your answer.

(2) If a model has *low variance* and *high bias*, is it more likely to *overly simple* or *overly complex*? Briefly justify your answer.

*********END*********