AIML427 Big Data: Assignment 1

This assignment has 100 marks and is due at **11:59 pm**, 25th March 2024. Please submit your answers as a single *.pdf* file. Make sure you read the Assessment section before writing the report. This assignment contributes 20% to your overall course grade.

1 Classification, Regression, and Clustering [15 marks]

Classification, regression, and clustering are three important tasks in big data, with many real-world applications. Read literature and/or online materials about these three tasks, and use *no more than 2 pages* (in total) to answer the following questions:

(i) Describe a real-world *big data classification* example.
Discuss why it is a *classification* problem and why it is a big data problem.

(ii) Describe a real-world *big data regression* example.
Discuss why it is a *regression* problem and why it is a big data problem.

(iii) Describe a real-world *big data clustering* example.
Discuss why it is a *clustering* problem and why it is a big data problem.

You should consider the definitions and the *5Vs* in big data in Lecture 1 when answering the above questions. For the question of *"why it is a big data problem"*, you should cover at least **3Vs out of the 5Vs**. You should cite the paper(s), the book(s) or website that you take idea from and write the report **in your own words**. Please be aware of The University's plagiarism policy when writing the report.

2 Feature Selection/Construction Methods [40 marks]

Based on our lecture notes and the suggested readings, answer the following questions using *no more than 2 pages*:

(i) What are the **major differences** between feature ranking/feature weighting and feature subset selection methods? Briefly describe **two** typical methods (name, main idea, and reference) for each type.

(ii) Define each of: filter, wrapper, and embedded feature selection/construction approaches. Briefly describe **two** typical methods (name, main idea, and reference) for each approach.

(iii) Compare and contrast the **performance** of supervised filter, wrapper, and embedded methods in terms of the classification accuracy, the computation cost, and the generality to different classification methods. Explain the reasons for these differences.

(iv) Describe the main idea of Correlation-based Feature Selection (CFS)[1], including the two search methods that can be used with it. Is CFS a feature ranking or feature subset selection method? Is it a filter, wrapper, or embedded method? Justify your answers.

[1] Hall, M. A. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the 7th International Conference on Machine Learning (2000), pp. 359–366.

3 Use KNIME to perform Feature Selection [45 marks]

Konstanz Information Miner (KNIME) is an open-source data mining and machine learning suite, which is very popular in both industry and academia. A getting-started guide is here. Examples of various workflows in KNIME are here.

You should use KNIME for feature selection on the *Ionosphere* dataset. There are three data files: Ionosphere.csv, Ionospheretrain.csv, and Ionospheretest.csv, which are the whole dataset, 70% of the data as the training set, and the other 30% of the data as the test set, respectively.

Answer the following questions using the given datasets. For all questions, the experiments should be run with the default settings in KNIME (although you are encouraged to tune the parameters to explore more beyond this assignment to get more practice!). You need to provide screen shots of the KNIME GUI including workflow and results whenever applicable. Your answers to all the following questions (i to vii) should *not exceed 8 pages*. You should:

(i) Use KNIME to:

(a) calculate the accuracy of the Naive Bayes (NB) classification algorithm using the given training set (Ionospheretrain.csv) to train the algorithm. Test the learned classifier on the given test set (Ionospheretest.csv). Report the training and testing accuracies;

(b) split the whole dataset (Ionosphere.csv), using a random seed of "202203" to 70% as the training set and the other 30% as the test set, and calculate the accuracy using NB. Report both the training and testing accuracies; and

(c) compare the results of (a) and (b), and discuss your findings.

(ii) Suppose Sam uses the **Forward Feature Selection** meta node in KNIME to select features on the whole dataset (Ionosphere.csv) with the default settings. He uses the forward feature selection strategy, NB as the wrapped classifier, and selects the features which lead to the best classification performance. What features are selected? Is there anything wrong during this feature selection process? If so, briefly discuss what and why.

(iii) Use KNIME to transform the given training set and the given test set by keeping only the features selected in (ii). Calculate the accuracy of NB on the transformed training and test sets. Report the accuracy and compare the results with that calculated in i(b). Analyse and discuss reasons for any differences.

(iv) Use KNIME to perform wrapper feature selection. Use the **Forward Feature Selection** meta node with NB as the wrapped classifier to perform feature selection on the given training set and test the selected features on the given test set (ensure you transform the test set based on the selected features!). You should:

(a) report the selected features, and the NB classification testing accuracy;

(b) compare the obtained accuracy with that obtained in i(a), analyse and discuss any differences; and

(c) compare the results (i.e. the selected features and the classification accuracy) with that obtained in (iii), and analyse and discuss any differences.

(v) Use the C4.5 Decision Tree to perform classification on the given training set and the given test set, report the generated tree (screenshot of "Decision Tree View (simple)"), the training accuracy, the test accuracy, and analyse how the decision tree can perform feature selection.

(vi) Use Principal Component Analysis (the **PCA Compute** node) in KNIME to perform dimensionality reduction: transform data using PCA on the given training set, and then:

(a) report the five best principal components;

(b) use the five best principal components to transform the training and test set, report the accuracy of C4.5 on the transformed training and test sets; and

(c) compare the training accuracy and testing accuracy with those obtained in (v), analyse and discuss any differences.

(vii) Use KNIME to perform Correlation-based Feature Selection method (CFS). You should:

(a) Use the **Rank Correlation** node with the Spearman's rank correlation to measure the feature-class and feature-feature correlations for CFS on the training set. You can create a new meta node (in the same way as with the **Forward Feature Selection** meta node).

(b) Test the selected features on the given test set using NB.

(c) Report the selected features and the classification accuracy, compare them with those obtained in (iv), and analyse and discuss any differences.

Assessment

Format: You can use any font to write the report, with a minimum of single spacing and 11 point size (hand writing is not permitted unless with approval from the lecturers). Reports exceeding the maximum page limit will be penalised. Any additional material such as code or figures/tables can be included in an appendix, which will not count towards the page limit.

Communication: a key skill required of a scientist is the ability to communicate effectively. No matter the scientific merit of a report, if it is illegible, grammatically incorrect, mispunctu- ated, ambiguous, or contains misspellings, it is less effective and marks will be deducted.

Marking Criteria: The final report will be submitted to Turnitin for a plagiarism check. Late submissions without a pre-arranged extension will be penalised as per the course outline. The usual mark checking procedures in place for all assessment apply to this report. The assessment of the reports will account of the understanding of big data, clarity and accuracy of answer, presentation, organisation, layout and referencing.

Submission: You are required to submit a single *.pdf* report through the web submission system from the AIML427 course website *by the due time*.