Stellenbosch University

Applied Machine Learning 874
prediction Post-block Assignment 1

Total: [401]
Deadline: 22 May 2022, 23:59

Instructions

1. The focus of this assignment is to test your understanding of the concepts covered in topics 2 and 3, and the influence of data quality issues on the machine learning modules covered in this module.
2. Make sure to complete all of the sections of this assignment.
3. You may use any programming language and appropriate libraries.
4. Submit your assignment as a pdf document. Please name this pdf document `???????PBA1_874.pdf`, where you replace the question marks with your student number. Please note that all documents submitted **must be pdf**. **No other formats will be accepted**. Also follow the above naming convention, because scripts will be used to automate the extraction of your documents from all of the submissions. If the above file name convention is not followed, your assignment will not be marked.
5. Please make sure that you do and submit your own work. Plagiarism will not be tolerated.
6. No AI tools are allowed to complete any aspect of this assignment.
7. Only submissions via SUNlearn/SUNonline will be accepted. Emailed submissions will not be evaluated.
8. Note that late submissions cannot be accepted and that no extensions to the deadline can be provided.

1 Theoretical Questions [191]

Answer each of the questions below:

1. Data quality issues refer to the presence of missing values, outliers, noise, irrelevant descriptive features, features with ranges of values that differ in scale, and skew/imbalanced class distributions (with the latter only applicable to calssification problems). Complete the following table to indicate for each listed machine learning algorithm (please answer, using a table as illustrated below)

   - if it is robust to each of these data quality issues;
   - if the algorithm is robust, discuss why it is robust.
   - if the algorithm is not robust, why is it not robust?

   For each of the machine learning algorithms, consider standard implementations with no additions or changes to address any data quality issues. (126)

Machine Learning Algorithm	Data Quality Aspect	Is it Robust?	Motivation
Classification Tree	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
	Skew class distribution		
Regression Tree	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
Model Tree	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
k-NearestNeighbour (for classification)	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
	Skew class distribution		
k-NearestNeighbour (for regression)	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
Neural Network	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
	Skew class distribution		
Support Vector Machine	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
	Skew class distribution		
Support Vector Regression	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
AdaBoost (for classification)	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
	Skew class distribution		
Random Forest (for classification)	Missing values		
	Outliers		
	Irrelevant features		
	Data in different ranges		
	Skew class distribution		

2. The table below shows a set of eight numbers.

   (a) What is the entropy of the numbers in this set? Show all your calculations. (2)

   (b) What would be the reduction in entropy (i.e. the information gain) if these letters are split into two sets, one containing the even numbers and the other containing the odd numbers? Show all your calculations. (4)

3	2	4	6	5	8	7	6

3. Consider the data set given below, and assume that information gain is used to decide on splits. Which

descriptive feature will be split upon in the root of the classification tree? Show all of your calculations. Note that the last column represents the target feature. (20)

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

4. Classification trees induced using information gain have four inductive biases. Name these biases, discuss the consequences of each, and how these biases can be addressed. (12)

5. What is the inductive bias of the k-nearest neighbour algorithm? (2)

6. The next question focusses on cross-validation.

   (a) Explain why cross-validation is used to evaluate the performance of predictive models. (4)

   (b) In your own words, describe and illustrate how cross-validation is applied to quantify the performance of a predictive model. (5)

   (c) What are the limitations in cross-validation when applied to time series forecasting? (3)

   (d) Explain how cross-validation for classification problems can be adapted to be applicable to time series prediction. (5)

   (e) Describe how cross-validation should be applied to evaluate the performance of time series predictive models. In your discussion, first discuss the limitations of the standard cross-validation approach above for non-temporal data. Then explain the approach to cross-validation for time series, and the justify this approach. (8)

2 Decision Tree and k-Nearest Neighbour Sensitivity to Data Quality Problems [180]

For this assignment, you are going to evaluate the sensitivity of decision trees and the k-nearest neighbour algorithm to different data quality issues. Your sensitivity analysis needs to consider both classification and regression problems. Sensitivity to the following data quality issues have to be explored:

- Outliers
- Noise
- Missing values
- Irrelevant features
- For continuous-valued features, the effect of value ranges that differ in order of magnitude
- For classification problems, the effect of skew class distributions

For each of the issues above, you have to carefully think about the process that you will follow. This includes creating appropriate datasets and selecting sensible performance measures.

Your report should provide a detailed description of the algorithms used, and a discussion of your expectations about sensitivity towards each of the above data quality issues. The approach followed towards each of the data quality issues is described in the methodology section of your report. The empirical process provides information about control parameters, performance measures, and data sets. All detail to reproduce your experiments have to be provided. The results are provided in the results section, and are used to provide a conclusion about sensitivity with respect to each data quality issue. Comment on whether the empirical observations correlate

with your expectations as discussed in the background section, and as indicated in your answers to the theory questions above.

Mark rubric as follows:

Aspect		**Mark**
Algorithm description		10
Expectations		60
Empirical approach	Quality issues	30
	Control parameters	10
	Performance measures	5
	Data sets	5
Results & discussion		60

3 Time Series Modelling using k-Nearest Neighbour Algorithm [30]

Describe in detail how the k-nearest neighbour algorithm can be used in time series modelling. In your discussion, you must consider the following two approaches to time series modelling:

- Prediction of the next value in the time-series, i.e. it is a regression problem. (10)
- Prediction of trends, where trends are `Up, Down` and `No Change`, i.e. it is a classification problem. (10)

In your discussion, consider the impact of the number of neighbours. (10)