Applied Machine Learning 874
Post-block Assignment 1
Department of Industrial Engineering

Deadline: 23 May 2021, 23:59
Total: 215

Instructions

1. The focus of this assignment is to test your understanding of the concepts covered in topics 0, 1, 2, and 3.
2. The first part of the assignment has a theoretical focus. Make sure that you answer each of the questions.
3. The second and third parts of the assignment are application-based, where you will implement some of the investigated machine learning models.
4. You may use any programming language and appropriate libraries.
5. Submit your assignment as a pdf document. Please name this pdf document `???????PBA1.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.
6. Please make sure that you do and submit your own work. Plagiarism will not be tolerated.
7. Note that late submissions cannot be accepted and that no extensions to the deadline can be provided.

1 Theoretical Questions [80]

Answer each of the questions below:

1. Some machine learning algorithms can work on only numerical input features. This requires that categorical features have to be transformed to numerical representations. Consider the following two transformation options:

   A1 Numerical encoding where each discrete value of a categorical feature is simply assigned a numerical value. The encoding therefore creates a single new numerical feature to represent the categorical feature, with each unique value of the categorical feature assigned a numerical value. This approach is described on slide 518.

   A2 Create one additional binary-valued feature for each of the possible discrete values of the categorical feature, and use a binary encoding scheme to represent each of the discrete values as illustrated on slide 519.

   Considering a classification problem, why is approach A2 better than approach A1? (5)

2. Data quality issues refer to the presence of missing values, outliers, noise, and skew/imbalanced class distributions. With reference to classification problems, complete the following table to indicate for each machine learning algorithm

   - 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 and what can be done to cope with that data quality issue?

(64)

Machine Learning Algorithm	**Data Quality Aspect**	**Is it Robust?**	**Motivation**
Classification Tree	Missing values		
	Noise		
	Outliers		
	Skew class distribution		
Regresion Tree	Missing values		
	Noise		
	Outliers		
Model Tree	Missing values		
	Noise		
	Outliers		
k-Nearest Neighbour for classification problems	Missing values		
	Noise		
	Outliers		
	Skew class distribution		
k-Nearest Neighbour for regression problems	Missing values		
	Noise		
	Outliers		

3. Discuss why feature selection is important for machine learning algorithms. (5)
4. 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

2 Classification Predictive Model for Breast Cancer [90]

For this assignment you will explore predictive models to predict breast tumors as malignant (M) or benign (B). The target feature is `diagnosis`. The dataset contains the following descriptive features: `id, diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean`, `concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst, gender` and `Bratio`. Provided to you are two dataset, i.e. the training set, `breastCancerTrain.csv`, containing 390 instances and the test set, `breastCancerTest.csv`, containing 179 instances.

Complete this assignment by doing the following:

1. Explore the datasets, and identify general data quality issues. (6)
2. The first predictive model that you will develop is a classification tree. Do the following:
   (a) Discuss how you have addressed the above data quality issues for classification tree induction. Provide clear motivations for what you have done. (12)
   (b) Discuss any other data transformations that you have applied, with clear motivations. (5)
   (c) Induce a classification tree on the training set to overfit the training set. In your pdf document describe the classification tree approach, stating the approach used to select optimal splits. Provide the confusion matrix for the training set and the test set. (15)

(d) Now apply a post-pruning process on the induced tree, and provide the confusion matrices for the training and test sets. Describe the post-pruning process used. (15)

(e) How did pruning help to provide a more accurate predictive model? (5)

3. For the last part of the assignment, you have to determine which of a k-nearest neighbor algorithm and your classification tree above provided best performance.

   (a) Discuss how you have addressed the above data quality issues for the k-nearest neighbor algorithm. Provide clear motivations for what you have done. (12)

   (b) Discuss any other data transformations that you have applied, with clear motivations. (5)

   (c) Now, conclude on which approach performed best, and provide clear motivations. (15)

3 Seoul Bike Sharing Predictive Model [45]

For this assignment, you have to develop a predictive model to predict the number of bikes rented per hour. The dataset is available on SUNlearn as `SeoulBikeData.csv`. More information about the datset and the problem can be found at `https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand`.

From the machine learning approaches covered in topics 2 and 3 (i.e. k-nearest neighbour and decision trees), you have to determine which model performs best in the prediction of hourly bike rental. In your pdf document, report on

1. Any data quality issues and data transformations done to construct the different predictive models. Provide clear motivations. (10)

2. In short, describe the predictive models used. (15)

3. Provide a conclusion on which approach is best, providing information on the process followed, performance measures used, and clear motivations for your decision. (20)