Statistical Methods for Data Science

DATA7202

Semester 1, 2021

Lab 1

Objectives

On completion of this laboratory session you should be able to understand and implement model selection for both the regression and the classification setting. Specifically, the following should be achieved.

1. Understand and implement the train/validate/test partition approach.
2. Understand and implement the cross-validation mechanism.

1. In this exercise, we will create synthetic datasets for regression and classification problems. Use `make_regression` and `make_classification` functions from the `sklearn.datasets` library to create two datasets (for regression and classification) with 5 explanatory variables.

2. Consider a regression problem from the `Lab1_regression_data` folder. The data was partitioned to training, validation, and test sets.

   (a) Load and explore the data using the `pandas` library.

   (b) In this exercise, we will use the k nearest neighbor regression (KNNR). Find the appropriate function in the `sklearn` library.

   (c) Train the KNNR learner with the training portion of the data, and determine the error using the validation set for $k = 1, \ldots, 50$.

   (d) Plot the validation error as a function of the number of neighbors k, and determine the best k. Use the `matplotlib` library.

   (e) Deliver the generalization error using the final test data-set. What is your conclusion?

3. Consider the S&P500 dataset in `stock.csv`. Our objective is to predict the direction of the stock market.

   (a) Load the dataset, set the direction to be a categorical variable.

   (b) Fit the Logistic Regression model and print the corresponding misclassification % and confusion matrix.

   (c) Are the results too good to be true? Can you identify the problem?

(d) Split the data to train and test sets (the test set corresponds to year 2005). Fit the model and discuss the corresponding misclassification % and confusion matrix.

(e) Try to predict the direction using "Lag5","Lag4","Lag3",and "Volume" variables only.

4. We consider the Credit Approval Data Set from `http://archive.ics.uci.edu/ml/datasets/credit+approval`. This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes – continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values.

(a) Load and explore the data (`crx.data.csv`). Explore the frequency of $+/-$ instances, and find missing values.

(b) Prepare the data for analysis. The attribute information is as follows.

```
A1: b, a.
A2: continuous.
A3: continuous.
A4: u, y, l, t.
A5: g, p, gg.
A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
A7: v, h, bb, j, n, z, dd, ff, o.
A8: continuous.
A9: t, f.
A10: t, f.
A11: continuous.
A12: t, f.
A13: g, p, s.
A14: continuous.
A15: continuous.
A16: +,- (class attribute)
```

i. Set categorical/continuous variables.

ii. Transform categorical variables to numbers.

iii. Eliminate rows with missing values.

(c) In this exercise, we will consider several classification algorithms and test their performance (zero-one loss), via 10-fold cross validation.

i. Write a function that takes 3 parameters: X, Y, and a model, and returns the 10-fold cross validation zero-one loss estimator.

ii. Write a function that implements several classifiers (Multilayer perceptron ,K-Neighbors-Classifier, Support Vector Machine, Random Forest, and Logistic regression). The function will receive X and Y and return the 10-fold cross validation zero-one loss for all classifiers.

(d) Use the above functions to identify the best classifier.

(e) Scale the data and repeat the classifier evaluations. Identify the best classifier.

(f) Use Principal Component Regression (PCR) and repeat the classifier evaluations. Identify the best classifier. Do not worry if you did not study PCA yet, think about this as a dimensionality reduction technique (this will be covered in class).

(g) What is your conclusion? Compare the obtained results with the paper `Simplifying decision trees.pdf`.

5. Consider a unit square and consider the area in the low left quarter of this square. Write a Crude Monte Carlo algorithm that uses $N = 1000$ sample size and estimates the are of the low left quarter of the unit square.