COMP9517

Lab 4, S1 2018

Pattern Recognition and Performance Metrics

The goal of this lab is to become familiar with the pattern recognition / machine learning algorithms and performance metrics.

This lab requires Scikit-learn and numpy libraries.

Preliminaries

The following experiments will be based on datasets from sklearn.datasets. These datasets are derived from the UCI benchmark. Notice that each dataset is designed for a specific machine learning task, either classification or regression. You may use any appropriate dataset for each question. Some general information follows:

- Data sets
  - iris datasets: two-class classification
  - digits datasets: multi-class classification
- Basic steps for experiments include:
  - import packages
  - import datasets from sklearn datasets
    - import datasets
    - split datasets into training and test sets using testing ratio 0.2
  - initialize learning model
  - fit the model using training sets
  - evaluate the model using metrics for the learning task on the test data set.

1. Decision Tree

- Train a decision tree for two-class classification
- Test the performance of the trained model using metrics
  - Accuracy
  - Sensitivity
  - Specificity
  - Confusion metrics
- Explain the confusion matrix

2. **Logistic Regression**

- Train a two-class classifier using logistic regression
- Test the performance of the trained model using the metric Area Under ROC Curve (AUC). Note that probabilistic classification results are required to generate TOC curve
- Plot the ROC Curve
- Explain how the ROC curve is generated.

3. **K-nearest Neighbour**

- Predict the label for instances in a dataset using kNN and Euclidean distance
- Test the performance of the trained model using the metric Accuracy
- Test for k between [1,3,5]

Evaluation Question: Show results for k= 1, 3, 5. Explain how kNN generates the label for a sample.