CE802 - Machine Learning and Data Mining

Assignment: Design and Application of a Machine Learning System for a Practical Problem

University of Essex

School of Computer Science and Electronic Engineering

PILOT STUDY

Submitted by
Arun Ravi
Registration number: 2009918
MSc – Artificial Intelligence & its Applications

Word count: 529

Overview:

This is a pilot study for application of machine learning techniques to classify the customer of a supermarket chain into two categories. Based on the information about the customers featuring age, average income, nationality and their previous shopping history they can be classified into two main classes: People who usually buy few expensive products and people who buy many cheap products. All these features can directly or indirectly affect the range of customer's shopping habits. Using the available data, I have been assigned to identify the class that a future customer will belong to.

Task:

A machine learning model is supposed to be used in this problem. This supervised learning comprises of labelled information features and ideal target variable. When data being used to predict categorical variable, this managed learning is called classification. This task is allowed to classified customers into two labels, hence known as binary classification. Wherein, regression predictive model is used to predict continuous values rather than fixed ones.

Proposal

Information of previous customers related to the features in variables that can be used to classify them into a category is the core data to build an algorithm. This data should be in a reasonable size without any errors and inaccuracies to train the model and obtain best possible result.

Since the problem is more related to classification of the data, this model will be based on classification algorithms. To select the best algorithm, the factors needed to be considered are:

1. Training data size
2. Time taken to train
3. Number of features
4. Linearity

Learning Procedures:

Logistic regression algorithm: To find different methods to regularize the model even when we are not sure of the features that are correlated. This algorithm has a good probabilistic interpretation and can be updated and adopted to new sets of data which is difficult with Decision Trees and Support Vector Machines.

SVM - Support Vector Machines: High accuracy and performance characteristics of SVMs are its advantages. It can guarantee overfitting and flexible selection of kernels in data that are not linearly separable. This algorithm exists with high dimensional spaces and popular in text classifications.

Decision Tree: This algorithm is simple to interpret and explain. This does not require any distribution and also non-parametric. Decision tree can effectively handle outliers whether or not the datasets can be separated linearly. Also, decision trees are good for few categories variables.

Evaluation:

Most of the binary classification algorithms can be evaluated by a prediction score. This score indicates the model’s confidence on how the result is categorised into a specific class. To conclude a decision of whether the observation is classified as Positive or Negative, we can interpret the score by selecting a classification threshold and comparing the probability against them. Observations with scores higher than the threshold are classified as positive and scores lower than the threshold are predicted to be negative.