Machine Learning Engineer Nanodegree

Capstone Proposal

Prakhar Dogra
February 13th, 2018

Proposal

Bank Marketing Classification

Domain Background

The algorithms stated above come under the Supervised Learning techniques that are used for classification purposes. There has been significant research on these techniques over different domains including Bank Marketing. The bank's marketing department can use machine learning to analyze customer datasets and develop statistically profiles of individual customer preference for product and service. In bank marketing domain, there are several techniques that can be used for classifying marketing service such as decision trees, naive Bayes classifier, support vector machine, etc. The research paper "A Comparison of Different Classification Techniques for Bank Direct Marketing" entails some of the above mentioned problems.

Problem Statement

As specified in the goal of the project, the main objective of this project is to classify if the client has subscribed for a term deposit depending on various bank marketing attributes. Different classification algorithms will be used and compared to see which performs best.

Datasets and Inputs

I plan to use the Bank Marketing Data set available at UCI ML repository. Following is the weblink to download the dataset:

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

There are two major datasets. First one is the older version that has 16 attributes. Second one is the new version that has 20 attributes. I will be using both and will see how results differ for the algorithms when using different datasets. Some important notes about dataset are:

- Some of the attributes are categorical and will be transformed using One Hot Encoding. Rest are continuous and will be normalized.
- There are over 40000 data points in each of the datasets.
- There are two classes ('yes' and 'no') in the dataset specifying whether the client has subscribed for a term deposit or not. Classes are slightly unbalanced.
- Dataset will be split into two main parts. One part (80%) will be for training & validation (will be used for cross validation) and other part (20%) will be for testing (to be used to measure accuracy and F1 score only after hyper parameter tuning).

Solution Statement

For the purpose of this project, I will be using different classification methods to see which algorithm reveals best results and in how much time. I will be measuring the methods on the basis of accuracy and F1-score. I will be using cross validation to get the best parameters. I also plan to use chi-square test and recursive feature elimination to filter some parameters and see how that improves the results. I will also see if using PCA improves the results. Following are some of the classification methods I will be using:

- Ada Boost
- Random Forest
- Support Vector Machine
- Naïve Bayes
- Multi-Layer Perceptron
- Decision Trees, etc.

Benchmark Model

I will be using Logistic Regression as the benchmark model and comparing it with different classification methods mentioned above.

Evaluation Metrics

I will be using accuracy and F1-score as evaluation metrics to compare the classification methods.

Project Design

Following depicts a step by step procedure for the project design:

- Load the dataset into memory.
- Convert categorical attributes into One-Hot Encoded attributes.
- Normalize continuous attributes.
- Apply classification methods one by one.
    - Logistic Regression
    - Ada Boost
    - Random Forest
    - Support Vector Machine
    - Naïve Bayes
    - Multi-Layer Perceptron
    - Decision Trees
- Compare accuracy and F1-scores.
- Apply PCA transformation and check if there is any improvement in any of the classification methods.
- Use chi-square test and recursive feature elimination on the original features and check if there is any improvement (just like PCA
- Apply the same set of steps on the second dataset (latest version) and see if there is any variation in trend.