Machine Learning Engineer Nanodegree

Capstone Proposal

Prakhar Dogra
February 13th, 2018

Proposal

Bank Marketing Classification

Main objective of this project is to classify if the client has subscribed for a term deposit depending on various bank marketing attributes. I will be comparing different classification methods such as:

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

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.

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.

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.

Benchmark Model

There have been multiple research papers on applying different classification techniques on Bank Marketing Dataset. Most have used Decision Trees, Support Vector Machines and Radial Basis Function Network. I will add some more classification methods and observe if the new methods work better than the ones proposed in the research papers.

Evaluation Metrics

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

Project Design

I will start by loading the first (older) dataset and apply classification methods one by one. After that I will use PCA transformation and check if there is any improvement in any of the classification methods. Then I will do a chi-square test on the original features and check if there is any improvement (just like PCA). Finally, I will use recursive feature elimination to check the same.

After that I will apply the same above stated procedure on the second dataset (latest version) and see if there is any difference.