

BlackFriday-Final Project Proposal Report

Xiangyao Deng,Bin Gui*
Professor: Yue Ning
Course: CS-559,Machine Learning
School: Stevens Institute of Technology

1 INTRODUCTION

Perhaps one of the most obvious applications of Machine Learning in retail is to predict how much a customer is likely to spend at a store and to classify different customers into different purchasing level group. Such predictor has a clear commercial value to the store owners as it would help with their financial planning, inventory management, marketing, and advertising.

2 BACKGROUND

Machine learning is a sub-discipline of artificial intelligence. It mainly studies the machine to learn from past experiences, model the uncertainty of data, and predict in the future.

2.1 Adaboost

AdaBoost, short for Adaptive Boosting, is a machine learning meta algorithm formulated by Yoav Freund and Robert Schapire. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

2.2 ANN

Artificial neural networks (ANN) or connectionist systems are computing systems inspired by the biological neural networks that constitute animal brains. The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.

3 DATASET

The dataset here is a sample of the transactions made in a retail store. The store wants to know better the customer purchase behavior against different products. Specifically, here the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

Classification problem can also be settled in this dataset since several variables are categorical, and some other approaches could be "Predicting the age of the consumer" or even "Predict the category of goods bought". This dataset is also particularly convenient for clustering and maybe find different clusters of consumers within it.

3.1 Acknowledgements

The dataset comes from a competition hosted by Analytics Vidhya.

3.2 About this file

Dataset of 550 000 observations about the black Friday in a retail store, it contains different kinds of variables either numerical or categorical. It contains missing values.

4 APPROACH

4.1 Random Forests

At first, discrete the Age and target value, purchase. Using multi-classifier for building random forests, which can classify with high accuracy different group of people into different level of purchasing ability.

4.2 Neural Network

Using Neural Network for analysis of regression, which predict the exactly money of that different people would spent in black day.

4.3 Evaluation Method

Using evaluation Metrics for classification, through comparing the true positive value and true negative value with false positive value and false negative value, we can evaluate the performance of model, then we can continue to improve the model.

- Recall= $tp/(tp+fn)$
- Precision= $tp/(tp+fp)$
- Accuracy= $(tp+tn)/(tp+tn+fp+fn)$

REFERENCES

- [1] "Artificial Neural Networks as Models of Neural Information Processing | Frontiers Research Topic". Retrieved 2018-02-20.
- [2] <https://www.cs.ucsb.edu/~william/papers/MachineLearningPresentAndFuture.pdf>

*Both authors contributed equally to this research.