CIS 520, Machine Learning, Project Proposal

Yixuan Meng, Zhaozheng Shen, Zhouyang Fang
Team: Finite State Machine

1 Motivation

In recent years, review text classification has attracted the attention of many researchers, which motivate us to come up with our predictive model to classify clothing rating with supervised learning. Based on the Women's E-Commerce Clothing Reviews Dataset , we plan to compare the performance of common traditional machine learning models and deep learning based methods.

2 Data set

We use the Women's E-Commerce Clothing Reviews Dataset[1] in this project. This dataset collect reviews written by customers. It contains 23486 rows, each row represents a customer's review. There are 10 data fields in total, including an ID for each cloth, customers' age, title and content of the review, whether the customer recommend the piece, number of other customers found the review positive, and three categorical features describing the division, department and class of the product.

3 Related Work

Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network[2]

This paper implemented a bidirectional recurrent neural network (RNN) with long-short term memory unit (LSTM) for recommendation and sentiment classification in this data set. Results have shown that a recommendation is a strong indicator of a positive sentiment score, and vice-versa. On the other hand, ratings in product reviews are fuzzy indicators of sentiment scores.

4 Problem Formulation

We have ten feature variables in this data set. We want to use different machine learning techniques to predict the rating using the rest nine features and analyze each feature's weight in the prediction. Basically, it is a typical supervised classification task.

5 Methods

We choose logistic regression as the baseline method with comparisons to SVM, Decision Tree, Random Forest, and Neural Network. They are all classic and popular classification algorithms. Considering some of the features, such as Clothing ID, may not have a strong correlation with label label Recommended IND, it is necessary to use dimensional reduction methods such as PCA first.

6 Evaluation

We select the feature Recommended IND as label, and the rest 9 features as observations. We separate the data into 20% test set, and use the rest for 5-fold cross validation, and prevent our models from overfitting by plotting the bias-variance trade-off graph. We plan to use cross entropy loss as the loss function and use F-1 score, AUC, Accuracy and Runtimes, to evaluate the performance of our methods.

7 Project Plan

- Week 11 (Yixuan Meng): Prepare data, finish pre-processing, implement baseline method.
- Week 12 (Zhouyang Fang): Implement SVM, decision tree, and random forest, and finish evaluation process.
- Week 13 (Zhaozheng Shen): Implement neural network, perform model performance comparison.
- Week 14 (All of us): Consolidate, write project report.

[1] https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews
[2] https://arxiv.org/pdf/1805.03687.pdf