Amazon Reviews Classification

Yue Dai

daiyue@usc.edu

Dataset and pre-processing

- The whole dataset is consist of 400,000 reviews of products collected from Amazon. In our program, due to the limit of time and computation space, we randomly pick 50,000 reviews, put every fifth review into the testing set(10,000 reviews), and collect the rest of the reviews to train the system(40,000 reviews)
- During the pre-processing we uses 6 different ways to get rid of noise data:
  - Delete stop words;
  - Extract words from string;
  - Change words to their stems;
  - Change words to lowercase;
  - Delete duplicate words in each review;
  - Check the spelling and get rid of symbols.

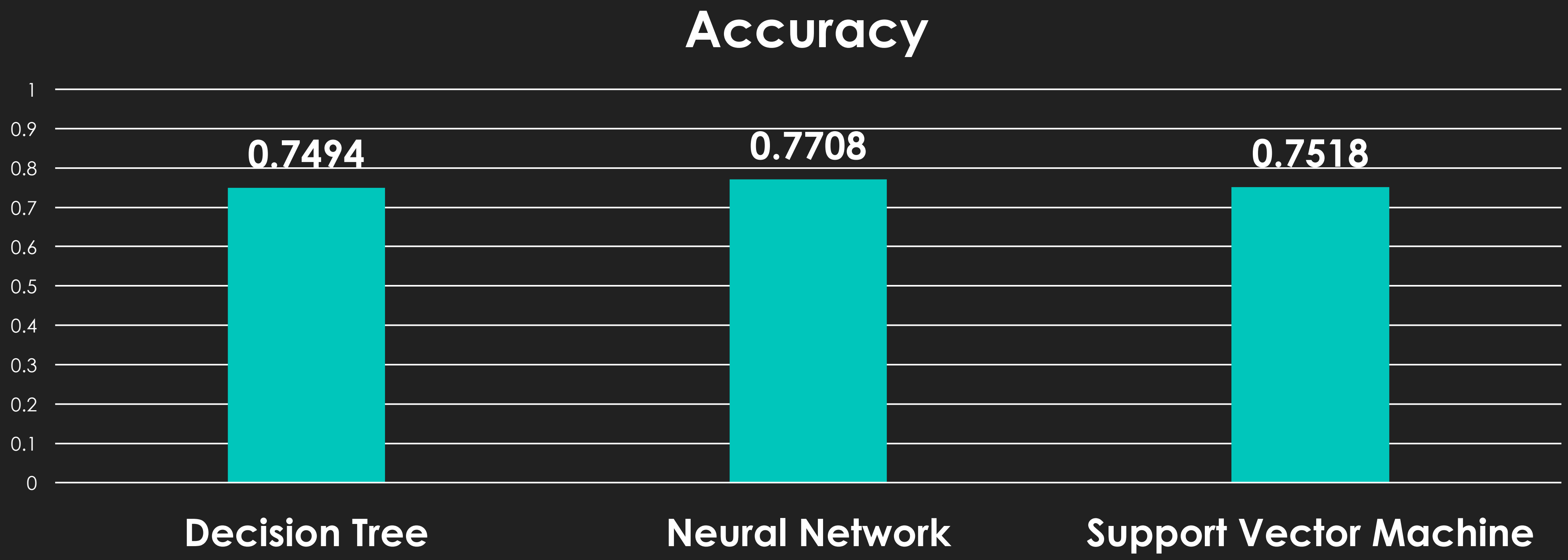
Feature extraction

- We use the term frequency–inverse document frequency(TF-IDF) method to extract the feature
  - If we analyzed the reviews in our dataset, we would end up with groups of documents about video', track', lyric', etc. We would gain a large amount of data about the types of reviews that had been written, but would not learn anything about what the users thought of those products.
  - diminishes the weight of the terms that occur in all the documents of corpus and similarly increases the weight of the terms that occur in rare documents across the corpus;
- We reduce the dimension of the features of each reviews
  - After applying FeatureExtraction to our data, each review has more than 1000 dimensions of feature, which makes it hard to apply classification algorithm to our training data and quiet a long time to train the system.
  - We have tried 4 different training data with 2 dimensions, 3 dimensions, 10 dimensions and 1000 dimensions, turns out 10 dimensions gives us almost the same accuracy as 1000 dimensions, but it is more efficient than 1000 dimensions.

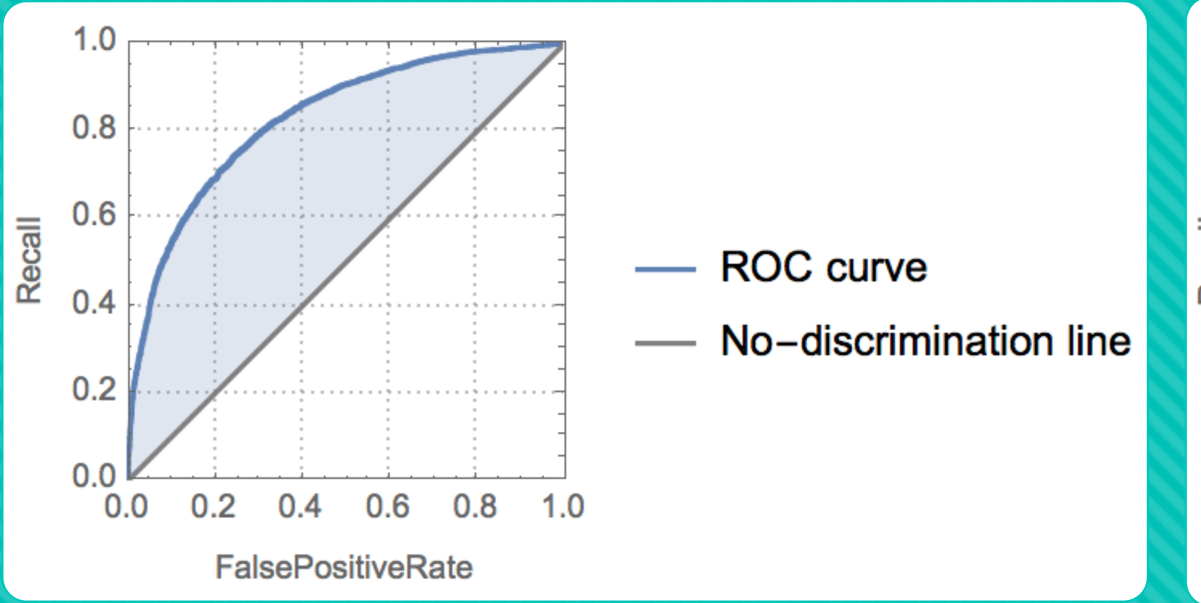
Classification Algorithms

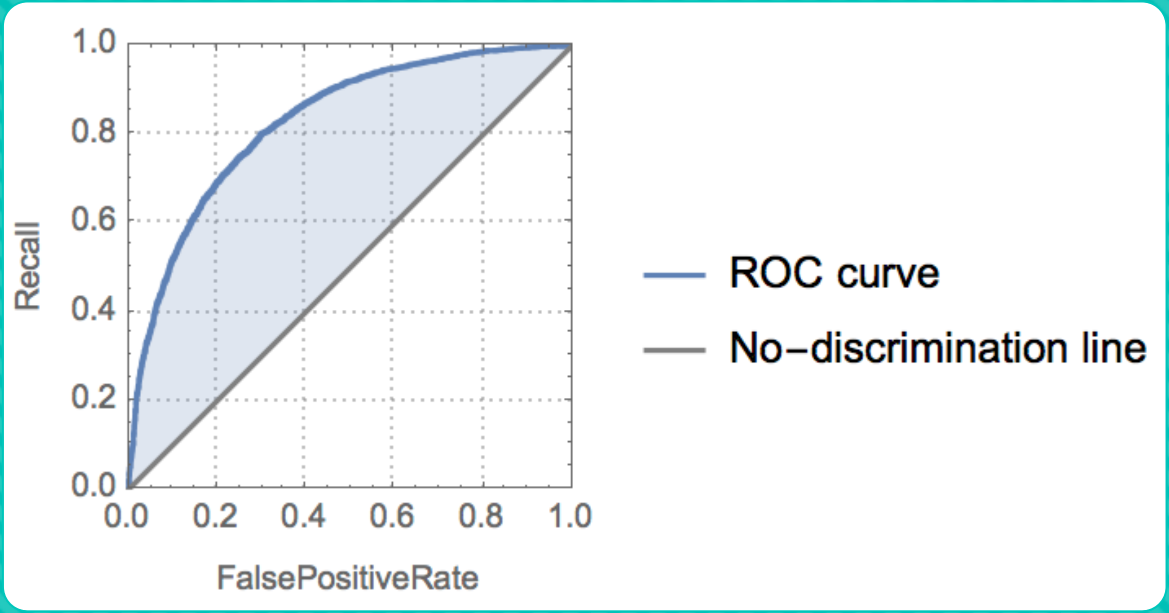
- Decision Tree Algorithm
  - **Easy to Understand**: Decision tree output is very easy to understand even for people from non-analytical background.
  - **Useful in Data exploration:** Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables.
  - **Less data cleaning required:** It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
- Neural Network Algorithm
  - Can approximate any function, regardless of its linearity
  - Great for complex/abstract problems like image recognition
  - Neural network models require less ford statistical training to develop
- Support Vector Machine Algorithm
  - SVM is a linear learning method that finds an optimal hyperplane to separate two classes
  - SVMs are a machine learning classification technique which use a function called a kernel to map a space of data points in which the data is not linearly separable onto a new space.
  - SVM have been proven as one of the most powerful learning algorithms for text categorization

Results and comparison

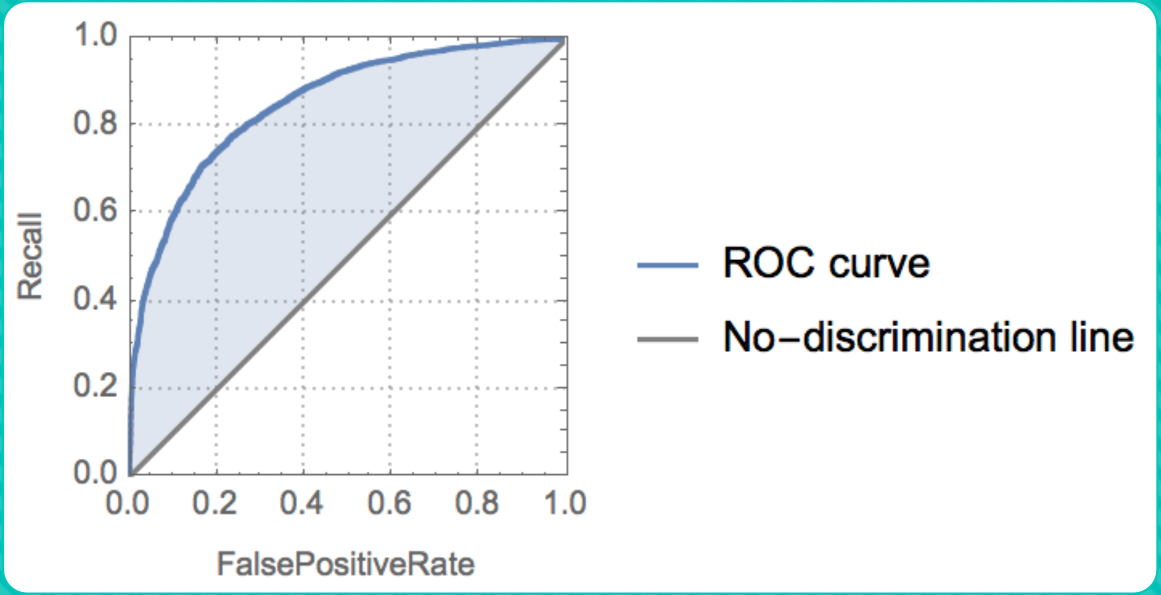


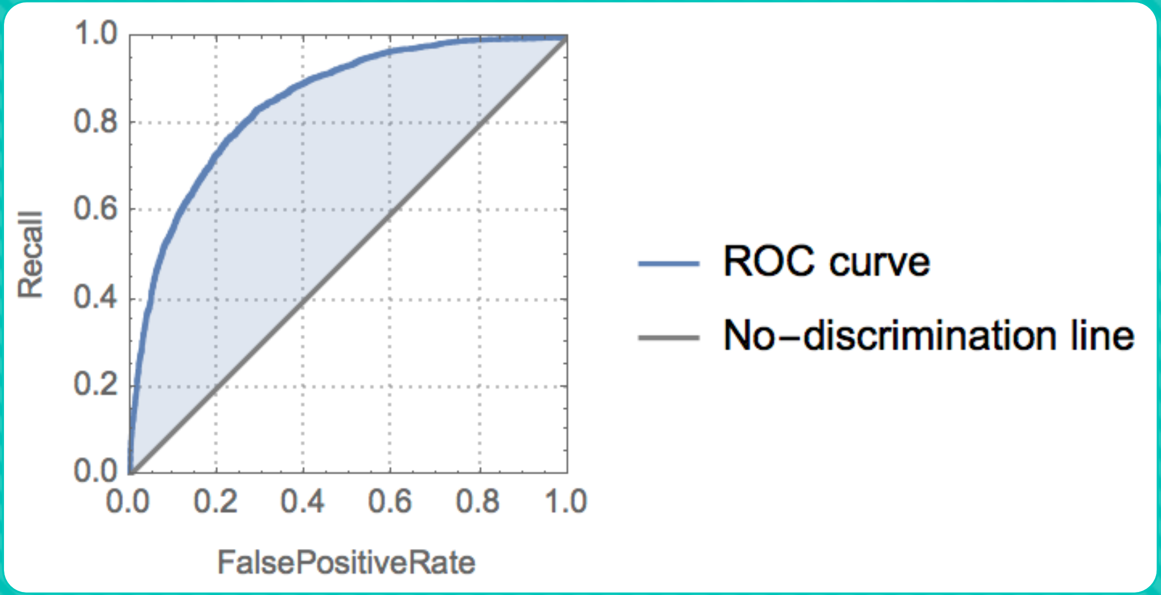
	Decision Tree		Neural Network		Support Vector Machine	
	Positive	Negative	Positive	Negative	Positive	Negative
Precision	0.74236	0.756276	0.773079	0.768715	0.749536	0.75393
Recall	0.748419	0.750343	0.75352	0.787409	0.741277	0.761914



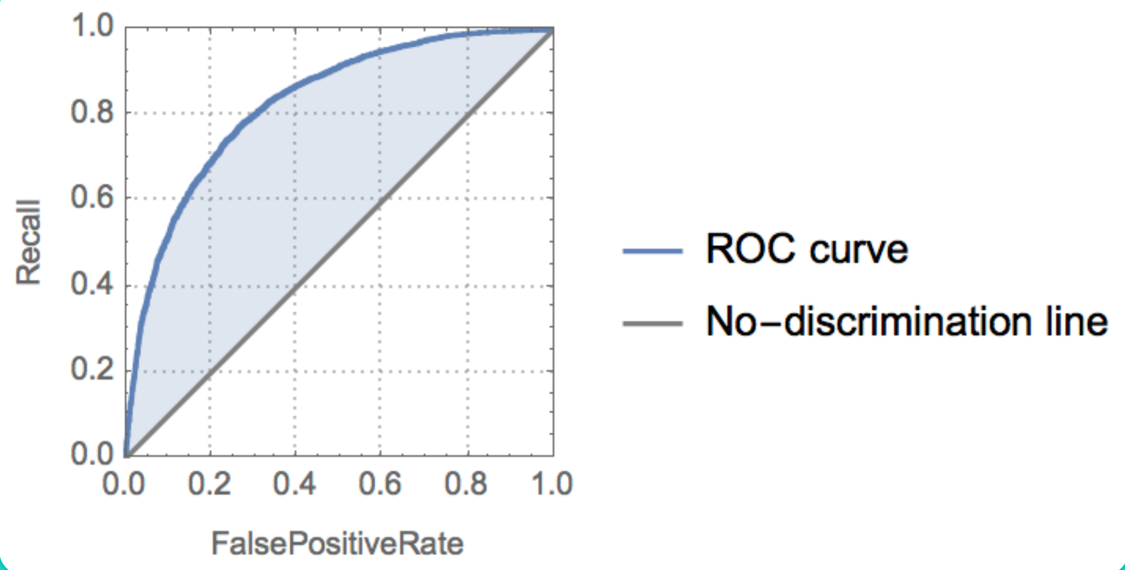


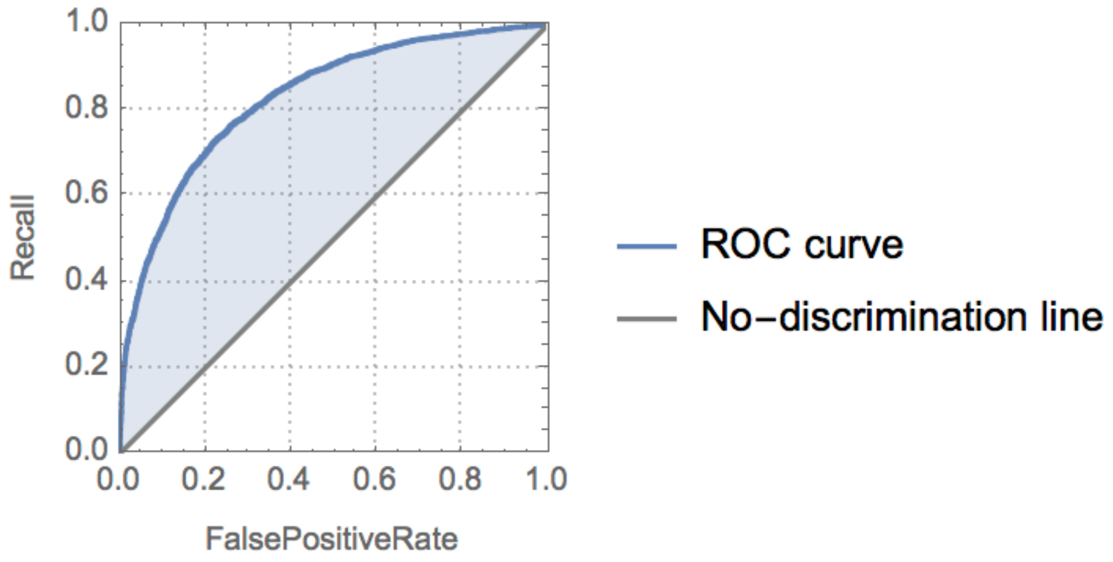
Decision Tree ROC





Neural Network ROC





Support Vector Machine ROC