CS 579 Presentation

Jiaqi Chen Xingtan Hu Xiaoyang Lu

Hypothesis

TCO price range can be predicted based on people's review.

Data Collection

- We collect people's reviews and Car's TCO price from Edmunds.com by using its API.
- Cars are classified into 12 main types, 230 cars, 896 styles.
- We have collected 3140 reviews, which has 11455 unique terms.
- Save all the data to .txt files for our program to read data for analysis.

Categorization

- Reading all the data from txt files.
- label the data with the TCO range, according to price ranges from Edmunds.com.

Tokenization

We tokenize the review data to terms for machine learning steps.

- We choose tf/idf vectorizer to get sparse matrix X.
- We use ngram and min_df to terms in different methods.

Method

In order to find the connection between them, we have tried different kind of machine learning method.

We use many built-in function in sklearn for machine learning.

Classification

We use Naive Bayes classifier for Classification.

- Gaussian Naive Bayes

  sklearn.naive_bayes.GaussianNB(X,y)

- Bernoulli models

  sklearn.naive_bayes.BernoulliNB(X,y)

Regression

We use Logistic Regression for Regression.

sklearn.linear_model.LogisticRegression(c)

We set the parameter c for regularization, it means inverse of regularization strength, smaller values specify stronger regularization.

Cross Validation

We use KFold for cross validation.

- Provides train/test indices to split data in train test sets.
- Split dataset into k consecutive folds. (without shuffling)

Experiments: Cross Validation Accuracy

	Cross Validation Accuracy
Logistic Regression (L2)	60.6%
Naive Bayes (Bernoulli Model)	56.0%
Naive Bayes (Gaussian)	51.8%

accuracy of most_frequent baseline : 48.1%

Future Work

- Try to find a more reasonable TCO price range classification method.
- Try sentiment analysis, but need to establish a new wordsets.
- Try taking other parameters into consideration.
  i.e. ratings for engines or ratings for reliability

Conclusion

Based on the accuracy and cross validation results, we can draw the conclusion that TCO price has some sort of connection with people's review which prove our hypothesis, although the connection is not so obvious. Future works needed to improve the results.

Thanks!

Any Questions?