Workshop 2 Answers

Team 18
Student Name: SHEN SHUTAO
Student ID: A0150148M

Question 1:

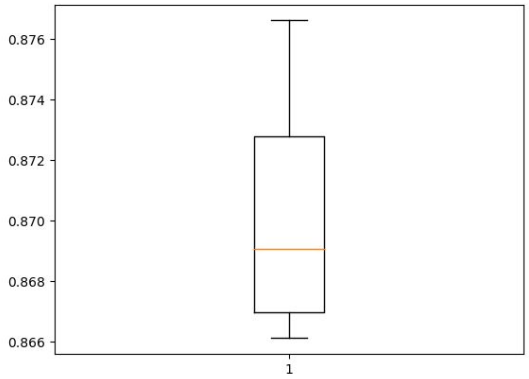
Step 1: choose best model by cross validation

Approach:

1. Random split the whole dataset into train & test set as 70%, 30%.
2. Train the model and do test.
3. Repeat step 1 & 2 for multiple times.
4. Summarize the accuracy result in a box plot.
5. Compare the results between different algorithms & choose the best algorithm.

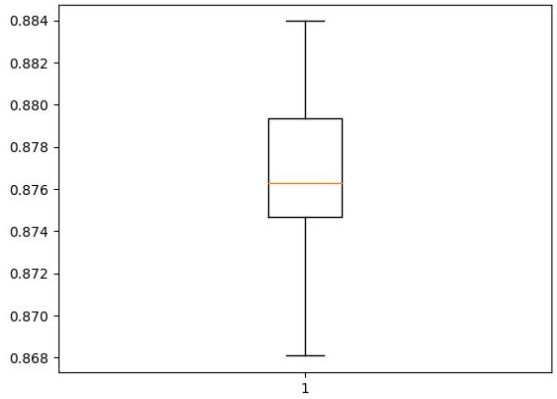
Performance of the models:

- Performance of SVM



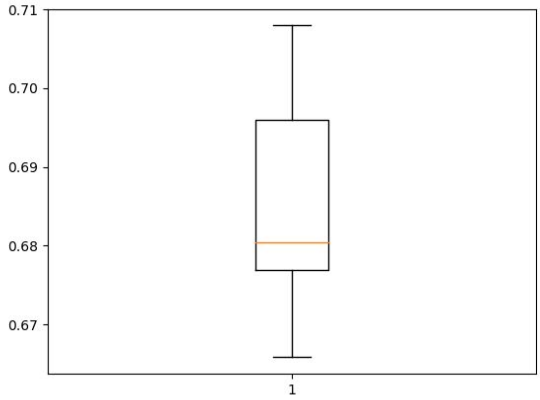
mean is 0.870114188214

- Performance of Naive Bayes (Multinomial Naive Bayes)



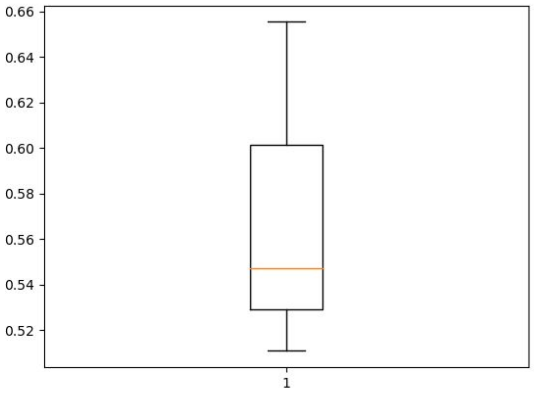
mean is 0.876692479328

- Performance of KNN:



mean is 0.68520803255

- Performance of Max Entropy:



mean is 0.691604322527

Step 2: retrain the best model with trainset data, test on testset data, get result

- Confusion Matrix:
  [[2684 382]
   [240 1760]]

- Classification Report:

	precision	recall	f1-score	support
negative	0.92	0.88	0.90	3066
positive	0.82	0.88	0.85	2000
avg / total	0.88	0.88	0.88	5066

- Accuracy Score:
  0.877220686932

Question 2:

Because in step 1, I'm using the Multinomial Naive Bayes, it gives me the most 5 coefficient features, and I got the biggest 5 feature prob for both positive & negative.

5 words most link to positive reviews:

- service
- great
- place
- good
- food

However this model is only give us the coefficient features for positive reviews.
so I use the nltk's naive bayes model to get the words most related to negative reviews.

5 words most link to negative reviews:

- NEG_returning
- flavorless
- ramen
- unacceptable
- nurse

Ps. which is interesting, as it is a binary classification.

- sklearn's multinomial naive bayes model regard the 'positive' class as binary 1
- nltk's naive bayes model regard the 'negative' class as binary 1.

It's not related to the class name, as I tried to replace their names with [1,0] or [0,1], or even reverse their class name.
It should not be consistent, I mean as the dataset change, this behavior should change. Need do further investigation.

But for now, it give us good result of the 5 words most link to positive / negative reviews.

Question 3:

Python file: workshop2q3.py
Classification model: classifier_nb.pk (python pipeline, include the Vectorizer)

The result is in the file ‘yelp_review.json’.