NLP Lab 9

Dixitha Kasturi
dkasturi@syr.edu

Topic: Cross Validation

General observation : The accuracy increased after cross validation as using folds ensures that the whole data is trained and tested on. The minimum accuracy with cross validation was higher than the accuracy that was achieved with 1 fold.

I performed both 5 fold and 10 fold cross validations on both bigrams and POS features. For 10 fold POS gave better mean accuracy whereas for 5 fold bigrams gave higher accuracy.

I. **For bigrams :** The mean accuracy for 5 folds and 10 folds was the same but the range of accuracies was higher in 10 fold, this could be because of the additional number of fold implying that the test sets are smaller in 10 fold compared to 5 fold. But the overall mean accuracy is the same because the whole data is trained and tested on. So essentially all records participate in training and testing at some point. The accuracy increased from 72.5% to 73.95% on 5 folds and 73.91% on 10 folds

   a. **1 fold accuracy**
   The accuracy was 0.725

```
# train a classifier and report accuracy
train_set, test_set = bigram_featuresets[1000:], bigram_featuresets[:1000]
classifier = nltk.NaiveBayesClassifier.train(train_set)
nltk.classify.accuracy(classifier, test_set)
```

```
0.725
```

   b. **with 5 fold cross validation :**
   max accuracy: 0.751
   min accuracy: 0.727
   mean accuracy : 0.7395

```
#for bigrams
num_folds = 5
cross_validation_accuracy(num_folds, bigram_featuresets)
```

```
Each fold size: 2132
0 0.7270168855534709
1 0.75187617260788
2 0.7298311444652908
3 0.7424953095684803
4 0.7467166979362101
mean accuracy 0.7395872420262665
```

c. with 10 fold cross validation:
max accuracy: 0.767
min accuracy: 0.720
mean accuracy : 0.7391

```
num_folds = 10
cross_validation_accuracy(num_folds, bigram_featuresets)
```

```
Each fold size: 1066
0 0.7204502814258912
1 0.7317073170731707
2 0.7607879924953096
3 0.7514071294559099
4 0.7317073170731707
5 0.726078799249531
6 0.7551594746716698
7 0.7204502814258912
8 0.7326454033771107
9 0.7607879924953096
mean accuracy 0.7391181988742964
```

II. **<u>For POS features :</u>** The mean accuracy for 10 folds was greater than 5 fold but the range of accuracies was higher in 10 fold just like the bigram features. The mean accuracy of 5 fold is lower than bigram feature set where as the mean for 10 fold is greater than 10 fold of bigram features.

a. 1 fold accuracy
The accuracy was 0.721

```
# train and test the classifier
train_set, test_set = POS_featuresets[1000:], POS_featuresets[:1000]
classifier = nltk.NaiveBayesClassifier.train(train_set)
nltk.classify.accuracy(classifier, test_set)
```

```
0.721
```

b. with 5 fold cross validation :
max accuracy: 0.754
min accuracy: 0.723
mean accuracy : 0.7389

```
num_folds = 5
cross_validation_accuracy(num_folds, POS_featuresets)
```

```
Each fold size: 2132
0 0.723264540337711
1 0.7542213883677298
2 0.7317073170731707
3 0.7378048780487805
4 0.7476547842401501
mean accuracy 0.7389305816135084
```

c. with 10 fold cross validation :
max accuracy: 0.759
min accuracy: 0.722
mean accuracy : 0.740

```
num_folds = 10
cross_validation_accuracy(num_folds, POS_featuresets)
```

```
Each fold size: 1066
0 0.7223264540337712
1 0.7373358348968105
2 0.7589118198874296
3 0.7598499061913696
4 0.7392120075046904
5 0.723264540337711
6 0.7467166979362101
7 0.7223264540337712
8 0.7335834896810507
9 0.7570356472795498
mean accuracy 0.7400562851782364
```