Project Progress Report 3

Status till report 2: I have trained below classifiers and had reported the cross validation score on the training data with a one particular set of parameters.

3. Experiments and Results:

Since the last report, I working on finding the performance of my classifiers on the validation data. I have tried many parameters for the classifiers to find out the best parameters settings for the classifier.

Below are the classifiers and the different parameters I tried for them:

1. Logistic Regression
   I tried all combinations of below 3 parameters exhaustively:
   a. C: 0.001, 0.01, 0.1, 1, 10, 100, 1000
   b. class_weight: “balanced”, None
   c. max_iters: 50, 100, 200, 300, 400

   These parameter settings worked the best with my validation data:

   C = 100 (irrespective of settings for class_weight and max_iters)

   Accuracy: 0.572525136154

2. Linear Support Vector Machine
   I tried all combinations of below 3 parameters exhaustively:
   a. C: 0.001, 0.01, 0.1, 1, 10, 100, 1000
   b. class_weight: “balanced”, None
   c. max_iters: 50, 100, 200, 300, 400

   These parameter settings worked the best with my validation data:

   C: 10, class_weight: “balanced” ()

   Accuracy: 0.572525136154

3. Neural Networks
   I tried all combinations of below 8 parameters exhaustively:
   a. Activation: “logistic”, “tanh”, “relu”
   b. solver: “lbfgs”, “sgd”
   c. max_iters: 50, 100, 200, 300, 400
   d. learning_rate: “constant”, “adaptive”
   e. early_stopping: True, False
   f. Shuffle: True, False
   g. Hidden_layers: (5,5), (5,10), (10,10), (10.15), (15,15), (15,20), (20,20)
   h. Alpha: 0.01, 0.1, 1, 10, 100

These parameter settings worked the best with my validation data:

Activation : logistic hidden layers : (5, 5) learning rate : constant early_stopping : True
Shuffle : False max_iter : 300 solver : lbfgs (irrespective of Alpha value)

Accuracy: 0.572525136154

Notes on the results:

My models are not performing very well. This is because I am using bi-grams models on the snippets, which are of length 42 (which also contain stop words that are removed during data preprocessing).

As bigrams not repeat much in the data, there is sparsity issue which is causing the classifier to not perform very well.

I am thinking to overcome this issue by using word2vec model, which I believe will perform better than just the bigram model, as these don't have sparsity issue and they also make equivalent vectors for similar words and also for the words those co-occur in similar context.