CSC411H1

Assignment 3

Jahyun Shin (998994888)

Submitted On:
December 6th, 2017

1 20 Newsgroups Predictions

1.1 Pre-sorting of Original Text

The following words were assumed to have no effect in overall text context or meaning and were eliminated from the lists of words to classify in order to reduce feature dimensions:

a) Words longer than 20 letters
b) Words shorter than 3 letters
c) Numbers
d) Special characters (i.e. punctuations; anything other than alphabets)
e) Words longer than 6 letters and ending with “ing” (present participle)
   **such words were not eliminated, but had “ing” taken out from the ending
f) Words longer than 6 letters and ending with “ed” (past tense)
   **such words were not eliminated, but had “ed” replaced with “e”
g) Any of: "the", "then", "Thanks", "Thank", "she", "but", "let", "for", "out", "whether", "not", "through", "couldnt", "cant", "wouldnt", "wont", "shouldnt", "how", "have", "has", "had", "havent", "hasnt", "hadnt", "are", "all", "was", "were", "anyone", "someone", "maybe", "probably", "please", "they", "you", "this", "that", "these", "those", "from", "with", "soon", "within", "sometimes", "often", "always", "why", "how", "whom", "who", "because", "also", "there", "here", "therefore", "since", "from", "until", "what", "hence", "and", "can", "could", "will", "would", "where", "there", "here", "should", "must", "might", "may", "shall"

After such pre-sorting, feature dimension was reduced by ~22% from 101,631 to 78,819 words.

1.2 Hyperparameter Selection

Decision Trees, Random Forest, and Neural Network MLP algorithms were selected for the 20 newsgroups classification. 10-fold cross validation on the training set was performed for each algorithm in order to pick the best hyperparameter for each model.

1.2.1 Decision Trees

The **maximum depth** to which the tree can grow was chosen as the hyperparameter. A set of 11 values of maximum depth, [500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000] was put under the test using 10-fold cross validation on the training set. The following plot displays the performance of each value of the hyperparameter on the validation set:

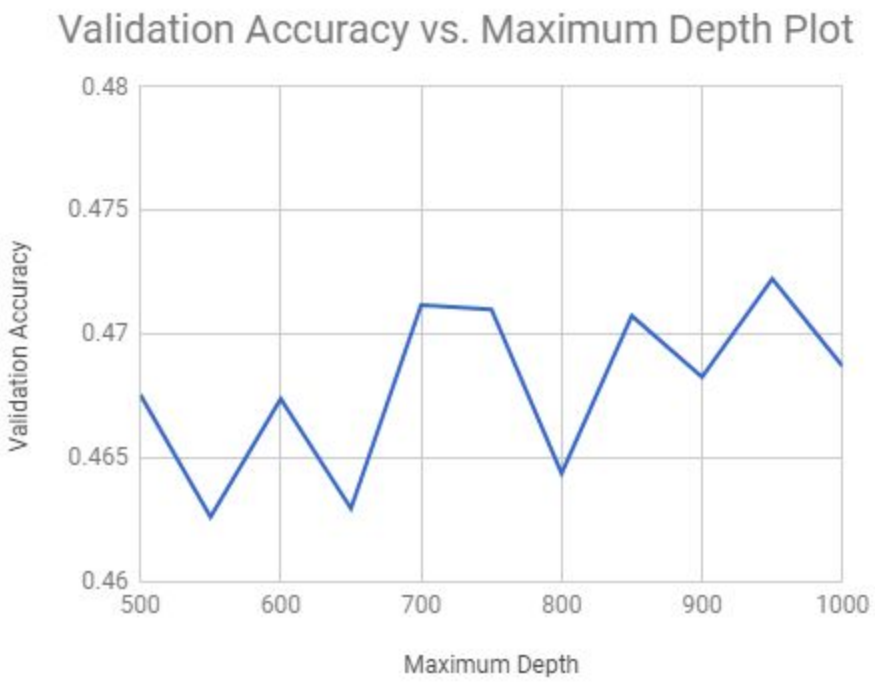

Figure 1. Validation Accuracy vs. Maximum Depth Plot for 10-fold Cross Validation on Decision Tree Algorithm

As observed from the plot, maximum depth of **950** showed the highest validation accuracy and was chosen as the optimal value.

1.2.2 Random Forest

The **number of trees** in the random forest was chosen as the hyperparameter. A set of 7 values of the number of trees, [100, 110, 120, 130, 140, 150, 160] was put under the test using 10-fold cross validation on the training set. The following plot displays the performance of each value of the hyperparameter on the validation set:

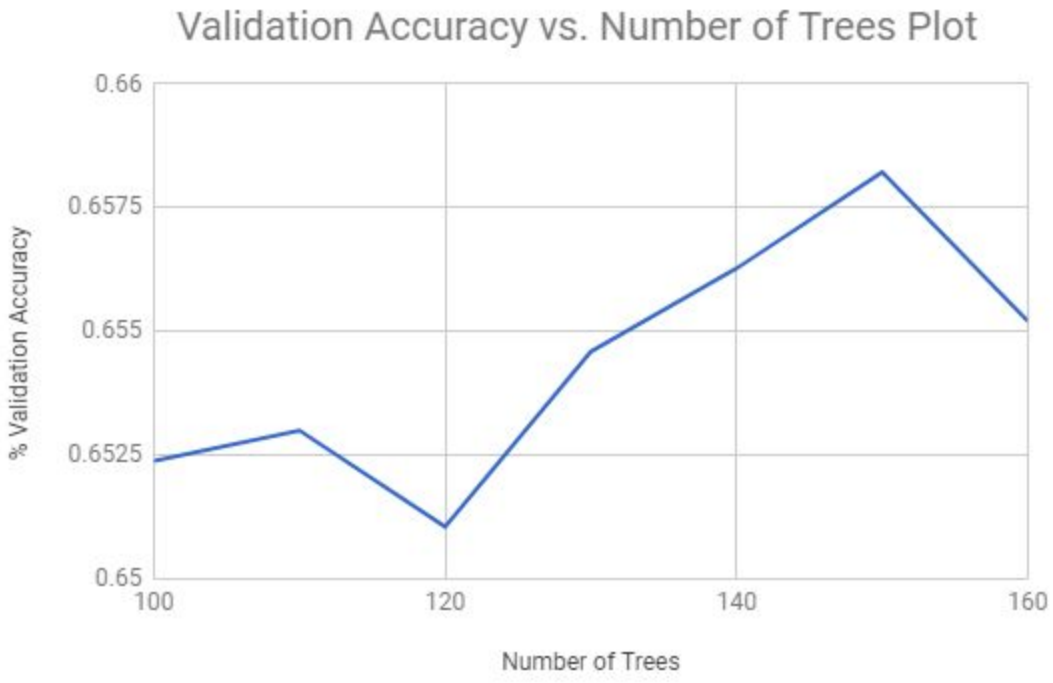

Figure 2. Validation Accuracy vs. Maximum Depth Plot for 10-fold Cross Validation on Random Forest Algorithm

As seen in the plot, having **150** trees showed the highest validation accuracy and was chosen as the optimal value.

1.2.3 Neural Network MLP

The **number of hidden layers** in the neural network was chosen as the hyperparameter. A set of 6 values of the number of hidden layers, [10, 20, 30, 40, 50, 60] was put under the test using 10-fold cross validation on the training set. The following plot displays the performance of each value of the hyperparameter on the validation set:

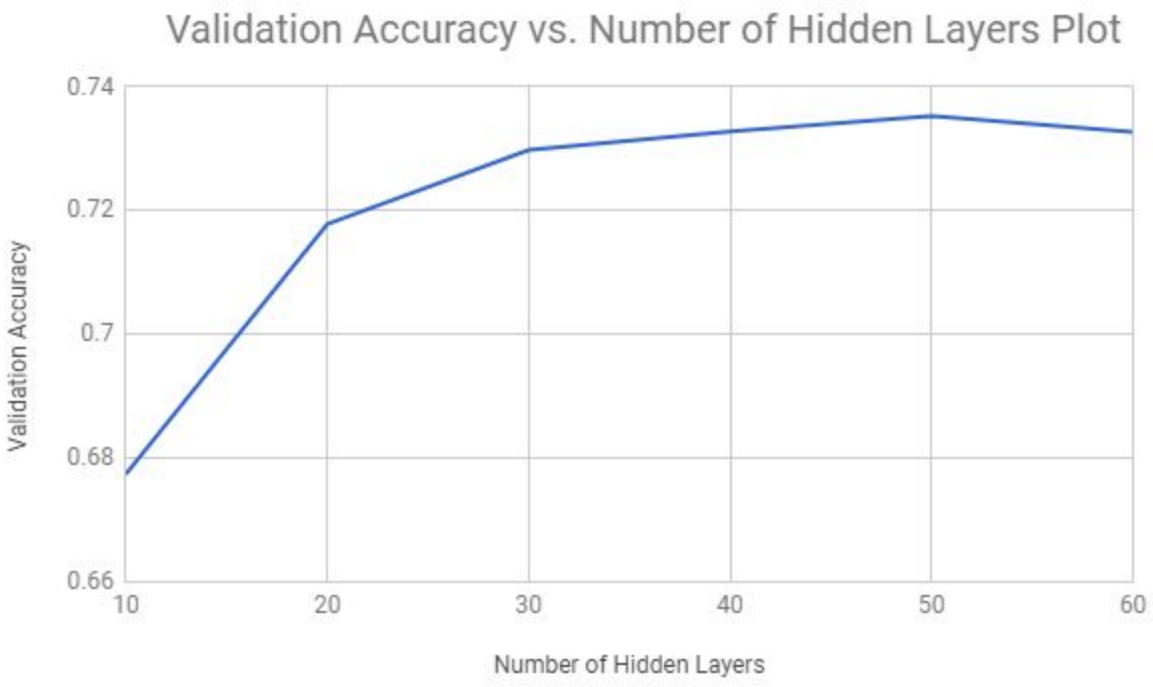

Figure 3. Validation Accuracy vs. Maximum Depth Plot for 10-fold Cross Validation on Neural Network MLP Algorithm

As observed from the plot, having **50** hidden layers in the neural network showed the highest validation accuracy and was chosen as the optimal value.

Additionally, learning rate of 0.05 and maximum iteration of 80 were used as suitable hyperparameters for convergence and acceptable running time. The cross-validated value of 50 hidden layers along with the two chosen values of learning rate and maximum iteration already outperformed the baseline. Since the scope of this project was to find an algorithm that outperforms the baseline, 10-fold cross validation was not performed on these parameters due to high running time (~10 hours for cross validation on the number of hidden layers only) of neural network MLP algorithm. To pursue the 20 newsgroups classification problem further in the future, cross-validation will be performed on those two hyperparameters as well.

1.3 Training and Test Accuracy of 4 Different Models

For developing the classification algorithm for 20 newsgroups dataset, tf_idf feature conversion was used to convert the original text. The following table summarizes the train and test accuracies on 20 newsgroups classification using 4 different models tuned with optimal hyperparameters:

Model	Hyperparameter	Value	Training Accuracy	Test Accuracy
Bernoulli Naive Bayes (baseline)	-	-	0.6506098638	0.4670738183
Decision Trees	Maximum Depth	950	0.9741912674	0.41011683483
Random Forest	Number of Trees	150	0.9741912674	0.59479553903
Neural Network MLP	Number of Hidden Layers	50	0.9741912674	0.66317047265

Table 1. train and test accuracies on 20 newsgroups classification using 4 different models with optimal hyperparameters

1.4 Observation on Performance

All three chosen models resulted in an equal training accuracy of ~97.4%. Neural Network MLP model with 50 hidden layers was found to be best performing with ~66.3% test accuracy. The next well-performing algorithm was Random Forest model with 150 trees with ~59.4% test accuracy. Unfortunately, Decision Trees model resulted in lower test accuracy of ~41.0% than the baseline Bernoulli Naive Bayes model with ~46.7%. This is quite expected since for within the 20 newsgroup data's huge feature dimensions, there must be a large amount of words that are very similar to each other (e.g. "song" and "songs; "insane" and "insanely") due to the linguistic nature of the dataset. Decision Tree model will split each of such similar words to a different leaf, which could easily lead to overfitting even if there is a limit in the maximum depth. Random Forest model reduces such concerns by considering a subset of features each time in building a tree and consequently performs much better than the Decision Trees model. Neural network MLP model must also have been able to group such similar words within its hidden layers, thus displaying a higher test accuracy then all three other models.

These three models were chosen because they were rather complex models suitable for multi-class classification problem. They also performed better than several other models, such as k-Nearest Neighbour algorithm whose test accuracy never reached greater than 15 percent.

1.5 Confusion Matrix of Neural Network MLP Algorithm

The confusion matrix C for the best-performing classifier, Neural Network MLP, was computed and displayed as below (*note: the confusion matrix itself is a 20 x 20 matrix without the highlighted class index from 1-20 on the first row and first column). Each element Cij represents the number of TEST examples truely belonging to Class j that were classified as Class i. The numbers at Cij where i equals j represent the number of correctly classified test examples.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	160	7	4	2	2	0	0	6	4	5	1	3	3	6	9	23	10	27	16	29
2	0	251	17	12	9	33	2	2	1	2	0	5	10	4	13	1	2	1	1	4
3	4	30	239	34	13	43	5	3	1	1	0	3	16	1	8	2	3	0	0	1
4	1	15	46	244	28	9	19	0	2	1	0	3	30	0	3	1	1	1	0	2
5	2	12	14	28	255	9	15	2	4	0	0	4	20	1	6	1	0	0	0	1
6	2	19	11	8	4	253	2	1	0	0	1	5	5	1	1	1	2	1	0	1
7	0	5	4	18	11	3	300	9	6	7	1	6	16	2	4	0	1	2	1	1
8	12	10	22	10	25	7	16	310	46	26	14	18	22	29	29	13	25	6	8	16
9	7	2	1	0	1	3	10	19	288	3	4	3	12	9	6	1	9	7	4	2
10	2	4	0	0	0	3	3	2	5	301	12	4	4	2	1	3	2	3	2	2
11	2	1	1	1	2	1	0	2	5	27	345	3	0	2	1	0	1	0	4	1
12	1	5	3	1	5	9	1	3	0	0	0	267	9	0	3	1	11	5	6	1
13	3	7	1	30	21	8	6	14	11	0	0	12	211	10	7	1	1	2	2	2
14	5	4	9	0	6	4	1	2	4	5	2	11	18	303	9	7	4	3	8	6
15	6	9	7	0	2	4	0	2	3	2	0	5	7	1	270	3	4	1	9	4
16	43	2	2	2	0	3	1	1	2	5	1	4	1	6	3	295	8	6	4	59
17	11	0	1	1	0	0	1	4	5	3	7	21	1	5	4	2	226	5	84	18
18	9	3	6	1	1	2	3	3	3	4	3	6	4	5	5	3	14	287	13	7
19	13	2	2	0	0	1	2	7	4	3	2	10	3	7	8	4	22	13	132	6
20	36	1	4	0	0	0	3	4	4	2	6	3	1	2	4	36	18	6	16	88

The classifier was most confused about **class 19 (“talk.politics.misc” subject)** and **class 20 (“talk.religion.misc” subject)**.