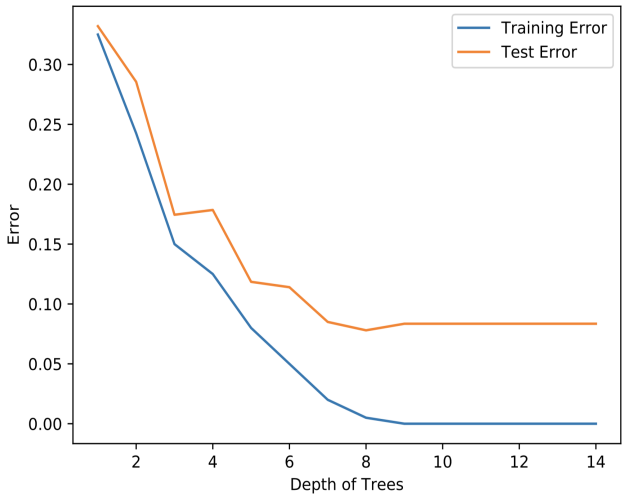
CPSC 340 Assignment 2 (due 2019-01-25 at 11:55pm)

Zitong Wang 40883150 k5j0b
Jingyuan Hu 41465155 z3o0b

January, 2019

1 Training and Testing

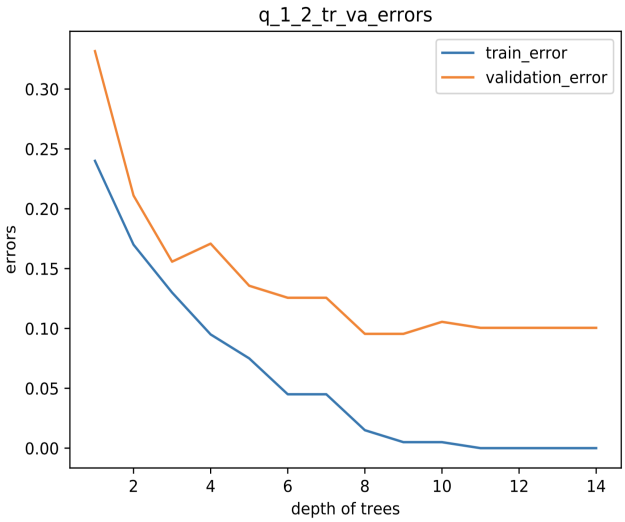
1.1 Training and Testing Error Curves



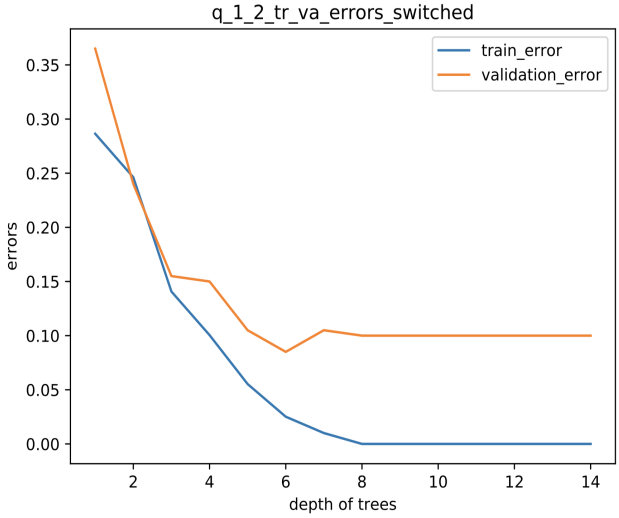
Observation: From the graph above, we can notice that the training error and test error decrease as depth of tree goes down. Around depth of tree = 8, there is a trend for test error to increase because of the overfiting. At the very last part of the plot, E_{train} stays at zero while E_{test} stays around 0.08.

1.2 Validation Set

This is for the original order(X_train, y_train first half, X_valid, y_valid rest half).



This is for the switched order:



Observation: For the original order, we would like to minimize the validation set error at depth of tree = 8.

After switching the order, there is a little bit difference between 2 plots; therefore, for the plot after order switched, we would like to minimize validation set error at depth of tree = 6.

To use more of our data, we would like to use K-fold cross validation to check the average best depth which can give a average best estimation.

2 Naive Bayes

2.1 Naive Bayes by Hand

2.1.1 Prior Probabilities

1. $p(spam) = \frac{3}{5}$

2. $p(notspam) = \frac{2}{5}$

2.1.2 Conditional Probabilities

1. $\frac{1}{6}$.
2. $\frac{5}{6}$.
3. $\frac{2}{6}$
4. 1.
5. $\frac{1}{4}$.
6. $\frac{3}{4}$

2.1.3 Prediction

For starters, test sample : $\hat{x} = [1\ 1\ 0]$.

$s.t$ for label "spam": $p(spam|\hat{x}) = p(spam = 1) * p(< yourname >= 1|spam) * p(pharmaceutical = 1|spam) * p(PayPal = 0|Spam) = \frac{3}{5} * \frac{1}{6} * \frac{5}{6} * \frac{2}{6} = \frac{1}{36}$

and, for label "not spam": $p(notspam|\hat{x}) = p(notspam = 1) * p(< yourname >= 1|notspam) * p(pharmaceutical = 1|notspam) * p(PayPal = 0|notSpam) = \frac{2}{5} * 1 * \frac{1}{4} * \frac{3}{4} = \frac{3}{40}$

$Therefore$, base on the calculation by given test sample, it is likely to label for "Not spam"

2.1.4 Laplace Smoothing

We can add four extra example in order to do lapalace smoothing:

$[0\ 0\ 0|spam], [0\ 0\ 0|not\ spam], [1\ 1\ 1|spam]$ *and* $[1\ 1\ 1|not\ spam]$.

2.2 Bag of Words

1. lunar
2. car fact gun video
3. talk

2.3 Naive Bayes Implementation

For code submission, please refer to the file naive_bayes.py in code folder

Naive Bayes (ours) validation error: 0.187

scikit-learn's implementation BernoulliNB validation error: 0.187

2.4 Runtime of Naive Bayes for Discrete Data

It should be $O(tdk)$.

Since $p(x_i)$ is calculated from the prior distribution sample, which is the training set and it's cost is $O(nd)$ thus we don't have to worry very much about the cost from this part, since we assume $k <= n$, therefore for $p(x_{ij}|y_i)$ we have to go through all t objects and d features from test set as well as classifying those by labels k, overall, the upper bound for the cost of classifying t test sample is $O(tdk)$.

3 K-Nearest Neighbours

1. Please refer to knn.py in code folder
2. KNN (ours) with k = 1 Training error: 0.000

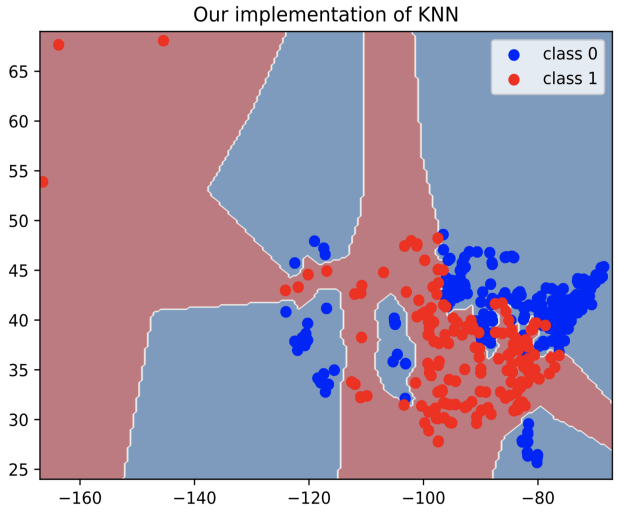
   KNN (ours) with k = 1 Test error: 0.065

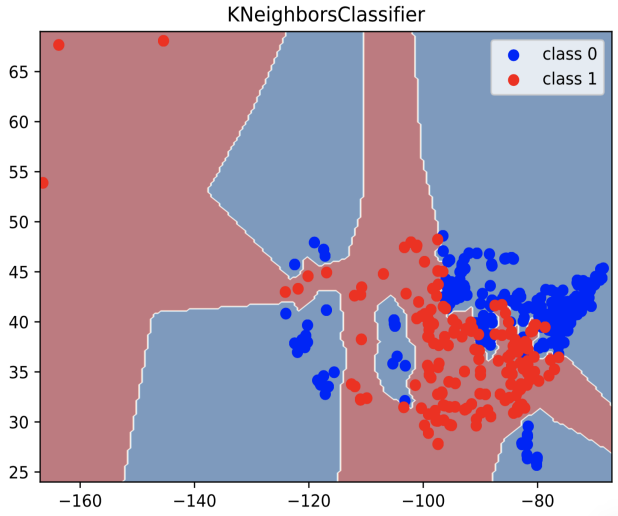
   KNN (ours) with k = 3 Training error: 0.028

   KNN (ours) with k = 3 Test error: 0.066

   KNN (ours) with k = 10 Training error: 0.072

   KNN (ours) with k = 10 Test error: 0.097

3. 



   Remark: both k are equal to 1

4. Because when choosing k = 1, every data point is considering itself as the closest neighbour, therefore the training error would be 0.
5. Firstly, we cannot choose k base on training set since k = 1 has lowest training error. In order to choose k, we can use cross validation by separating training data into training set and validation set

4 Random Forests

1. Because some of the data were selected as bootstrap sample such that would not be considered during training phrase.
2. Please refer to the file random_forest.py
3. **Decision tree info gain:**

   Training error: 0.000

   Testing error: 0.367

   Random Forest:

   Training error: 0.000

   Testing error: 0.178

   Random Tree:

   Training error: 0.125

   Testing error: 0.481

   Observation: Same as my expectation, testing error for Random Forest is smaller than a single decision tree or random tree.
4. **Random Forest:**

   Training error: 0.000

   Testing error: 0.182

   Runtime: 6.482173204421997

   RandomForestClassifier:

   Training error: 0.000

   Testing error: 0.152

   Runtime: 0.06684708595275879

   Observation: It is obvious that random forest classifier from skitlearn performs much more faster than than our implementation. But accuracy for 2 different implementation is floating among different selected data set. (Sometimes our is lower while sometimes is skitlearn's perform better)

5 Clustering

5.1 Selecting among k-means Initializations

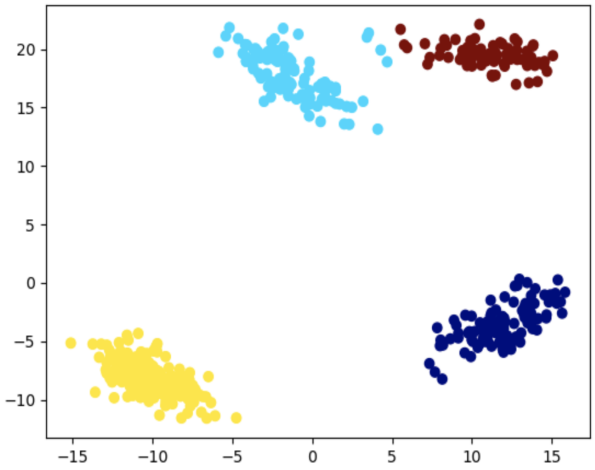
1. Please refer to the code file: kmeans.py
2. k = 1, 122972.69789150462

   k = 2, 45434.982570663095

   k = 3, 9909.520543791987

   k = 4, 3071.468052653853

3. The lowest error that reported is: 3071.468052653853



4. **n_cluster:** The number of clusters to form. (ie. number of means that selected in the first place)

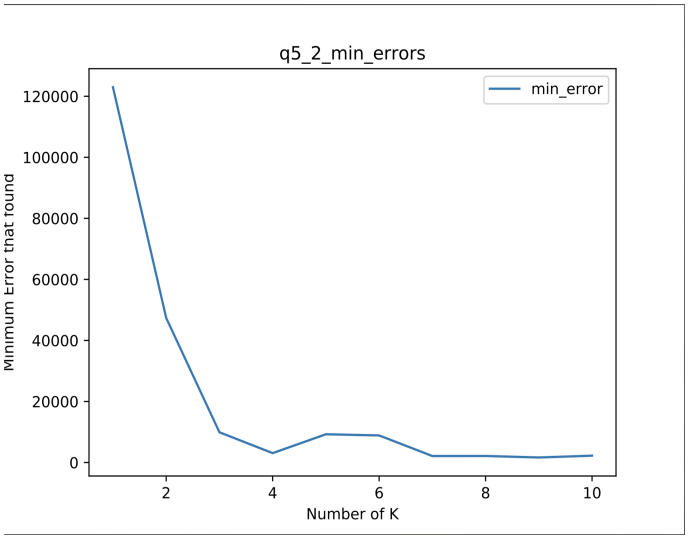
   init: An initialization method. Default setting is 'k-means++'. 'k-means++' can be used for speed up convergence selects initial cluster centers for k-mean clustering. 'random' can choose k initial means at random from data. When 'ndarray' is passed, it can ensure the shape is:(n_clusters, n_features) and gives initial center.

   n_init: number of time the k-means with different initially selected means. The final result will be the best output of n_init runs result.

   max_iter: The maximum number of iterations within each run of K-means

5.2 Selecting k in k-means

1. By choosing k in this way, we just simply memorize information gained from training data, once we switch to a new data set, that 'k' would not perform as good as the previous one.
2. K that chose from this way would be biased since we cannot use test data to choose our model.

3. 

4. From the graph above, it is reasonable to choose k = 3 since from k = 2 to k = 3 the absolute value of the slope of the plot seems to be the largest

5.3 Density-Based Clustering

Firstly, by setting min_sample = 2 (Selected point itself and one other) considered as core point. Then we can switch different esp values to check the outputs

1. Set esp = 2 or 3 will both produce 'true' clusters
2. Set esp = 5
3. Set esp = 15 and now is top and down 2 clusters
4. Set esp = 18 then there is only one cluster

6 Very-Short Answer Questions

1. Box plot can easily acknowledge, besides means and variance, but three different quantiles of the data (in order words data spread) and outliers of the data set.
2. The email system might be effected by forwarding or deleting previous emails such that leads potential uncertainty to independency.
3. Validation set is derived from the training data that used to ensure the accuracy while test set is used for provide a prediction of the model and it would not affect training phrase.
4. It may lead to an overfiting problem towards different data set since we are choosing the lowest error based on the training data.
5. Since we are using parametric models therefore the parameters that passed are fixed, therefore optimization bias(overfit to the validation set) will decrease fast with number of samples (n).
6. Advantage: For a large data set, the model that found would be more accurate than small k value. Disadvantage: run time would be extremely expensive.

7. For navie bayes, since we focusing on computing features conditioned on x_i such that we can treat $p(x_i)$ as constant then ignore.
8. Parameter, Parameter, hyperparameter.
9. When k is small, the training error would be small but approximation error would be large such that for k is big training error is large but approximation error would be small
10. Using k-means here would be effective and easy.
11. In supervised learning, we can tell whether the label y_i is good or not by testing how well the model can fit the data; however in clustering model, there is no absolute correct label for each cluster.
12. Not necessary, it might lead to overfitting or we would still choose the largest k for validation set.
13. Not necessary, since for k-means, the clusters were separated by means obtained (centroids) from each cluster while KNN separate clusters by chosen point and its kth nearest point.