CPSC 340 Assignment 2 (due 2018-09-28 at 11:55pm)

Instructions

Rubric: {mechanics:5}

The above points are allocated for compliance with the CPSC 340 homework submission instructions:
`https://github.ugrad.cs.ubc.ca/CPSC340-2018W-T1/home/blob/master/homework_instructions.md`

1 Training and Testing

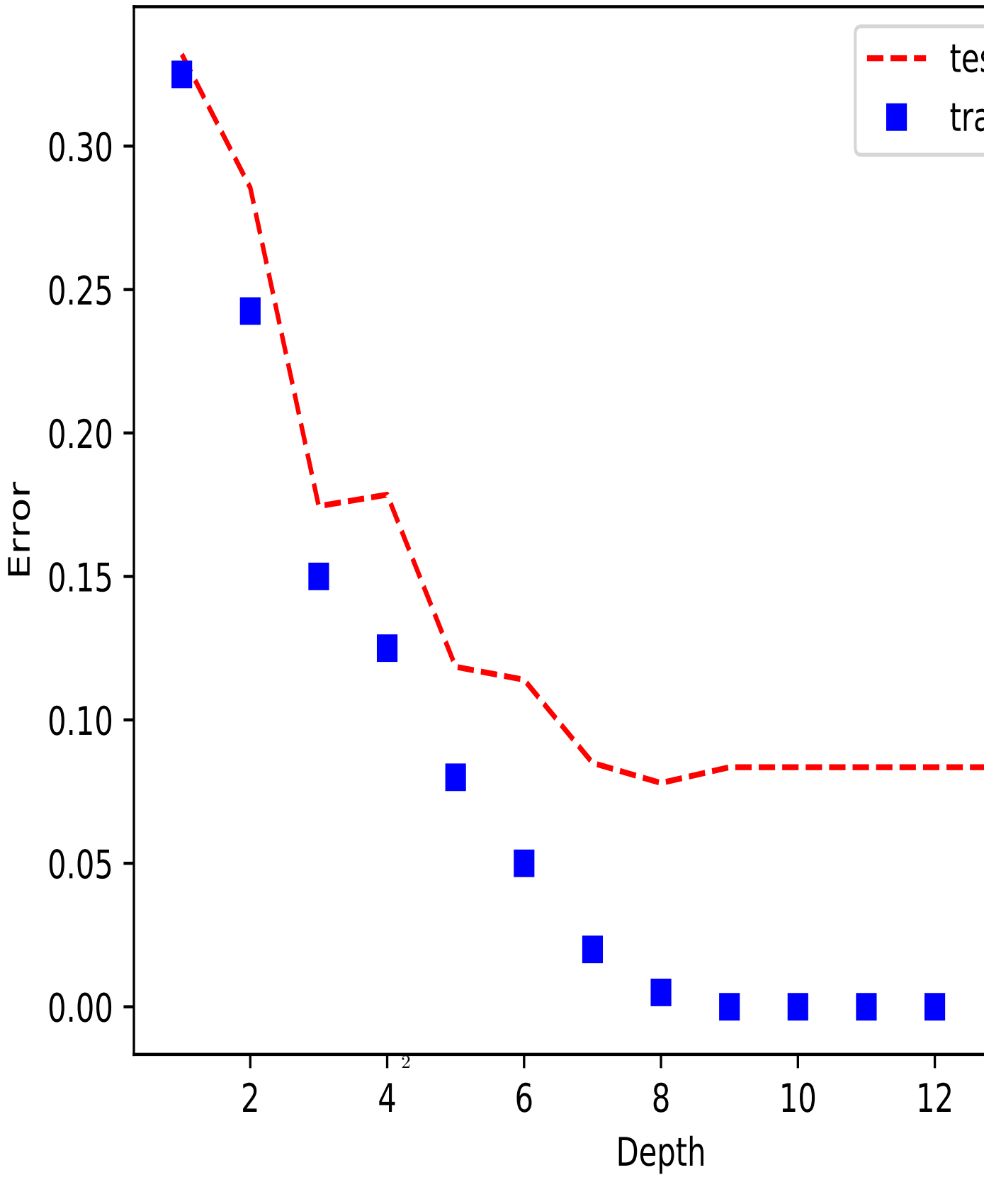
If you run `python main.py -q 1`, it will load the *citiesSmall.pkl* data set from Assignment 1. Note that this file contains not only training data, but also test data, `X_test` and `y_test`. After training a depth-2 decision tree with the information gain splitting rule, it will evaluate the performance of the classifier on the test data. With a depth-2 decision tree, the training and test error are fairly close, so the model hasn't overfit much.

1.1 Training and Testing Error Curves

Rubric: {reasoning:2}

Make a plot that contains the training error and testing error as you vary the depth from 1 through 15. How do each of these errors change with the decision tree depth?

Note: it's OK to reuse code from Assignment 1.


tes
tra
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Error
2
4
6
8
10
12
Depth

1.2 Validation Set

Rubric: {reasoning:3}

Suppose that we didn't have an explicit test set available. In this case, we might instead use a *validation* set. Split the training set into two equal-sized parts: use the first $n/2$ examples as a training set and the second $n/2$ examples as a validation set (we're assuming that the examples are already in a random order). What depth of decision tree would we pick to minimize the validation set error? Does the answer change if you switch the training and validation set? How could use more of our data to estimate the depth more reliably?

- We would pick N=9 which yields in a min error of 0.1 And switching the validation and testing set changes the result now the min error is 0.085. The optimal depth now is: N=6

  To increase the reliability of our depth choice, we could increase the size of our training set by collecting more data, or removing some elements from the validation set and adding them to the training set and selecting the k from this new k

2 Naive Bayes

In this section we'll implement naive Bayes, a very fast classification method that is often surprisingly accurate for text data with simple representations like bag of words.

2.1 Naive Bayes by Hand

Consider the dataset below, which has 10 training examples and 3 features:

$$X = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} \text{spam} \\ \text{spam} \\ \text{spam} \\ \text{spam} \\ \text{spam} \\ \text{spam} \\ \text{not spam} \\ \text{not spam} \\ \text{not spam} \\ \text{not spam} \end{bmatrix}.$$

The feature in the first column is <your name> (whether the e-mail contained your name), in the second column is "pharmaceutical" (whether the e-mail contained this word), and the third column is "PayPal" (whether the e-mail contained this word). Suppose you believe that a naive Bayes model would be appropriate for this dataset, and you want to classify the following test example:

$$\hat{x} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}.$$

2.1.1 Prior probabilities

Rubric: {reasoning:1} Compute the estimates of the class prior probabilities (you don't need to show any work):

- $p(\text{spam}). = \frac{6}{10}$

- $p(\text{not spam}). = \frac{4}{10}$

2.1.2 Conditional probabilities

Rubric: {reasoning:1}

Compute the estimates of the 6 conditional probabilities required by naive Bayes for this example (you don't need to show any work):

- $p(\text{<your name>} = 1 \mid \text{spam}). = \frac{1}{6}$
- $p(\text{pharmaceutical} = 1 \mid \text{spam}). = \frac{5}{6}$
- $p(\text{PayPal} = 0 \mid \text{spam}). = \frac{2}{6}$
- $p(\text{<your name>} = 1 \mid \text{not spam}). = 1$
- $p(\text{pharmaceutical} = 1 \mid \text{not spam}).= \frac{1}{4}$
- $p(\text{PayPal} = 0 \mid \text{not spam}). = \frac{3}{4}$

2.1.3 Prediction

Rubric: {reasoning:1}

Under the naive Bayes model and your estimates of the above probabilities, what is the most likely label for the test example? (Show your work.)

- All strings Kaan pharm and paypal below are normally in "" but I am going to skip them to facilitate typing it(except the first line)

  We are comparing $p(spam = 0|Kaan = 1, pharm = 1, paypal = 0$ to :

  $p(spam = 1|"Kaan" = 1, "pharm" = 1, "paypal" = 0) = \frac{p(kaan=1,pharm=1,paypal=0|spam=1)p(spam=1)}{p(kaan=1,pharm=1,paypal=0}$

  = p(kaan=1,pharm=1,paypal=0—spam=1) by multiplying each side by p(kaan=1,pharm=1,paypal=0) and using naive bayes we get: $p(kaan = 1|spam = 1)p(paypal = 0|spam = 1)p(pharm = 1|spam = 1)p(spam = 1)$

  which is equal to $\frac{1}{36}$

  the other side by a similar calculation is equal to $\frac{3}{40}$

  Since $\frac{3}{40} > \frac{1}{36}$ we conclude that the email is not a spam so label = 0

2.1.4 Laplace smoothing

Rubric: {reasoning:2}

One way to think of Laplace smoothing is that you're augmenting the training set with extra counts. Consider the estimates of the conditional probabilities in this dataset when we use Laplace smoothing (with $\beta = 1$). Give a set of extra training examples that we could add to the original training set that would make the basic estimates give us the estimates with Laplace smoothing (in other words give a set of extra training examples that, if they were included in the training set and we didn't use Laplace smoothing, would give the same estimates of the conditional probabilities as using the original dataset with Laplace smoothing). Present your answer in a reasonably easy-to-read format, for example the same format as the data set at the start of this question.

- We can add the following:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} \text{spam} \\ \text{spam} \\ \text{spam} \\ \text{not spam} \\ \text{not spam} \\ \text{not spam} \end{bmatrix}.$$

2.2 Bag of Words

Rubric: {reasoning:3}

If you run `python main.py -q 2.2`, it will load the following dataset:

1. X: A binary matrix. Each row corresponds to a newsgroup post, and each column corresponds to whether a particular word was used in the post. A value of 1 means that the word occured in the post.
2. *wordlist*: The set of words that correspond to each column.
3. y: A vector with values 0 through 3, with the value corresponding to the newsgroup that the post came from.
4. *groupnames*: The names of the four newsgroups.
5. $Xvalidate$ and $yvalidate$: the word lists and newsgroup labels for additional newsgroup posts.

Answer the following:

1. Which word corresponds to column 51 of X? (This is column 50 in Python.)
2. Which words are present in training example 501?
3. Which newsgroup name does training example 501 come from?

- 1)lunar

  2)bmw fact fans jews religion version

  3)talk.*

2.3 Naive Bayes Implementation

Rubric: {code:5}

If you run `python main.py -q 2.3` it will load the newsgroups dataset, fit a basic naive Bayes model and report the validation error.

The `predict()` function of the naive Bayes classifier is already implemented. However, in `fit()` the calculation of the variable `p_xy` is incorrect (right now, it just sets all values to 1/2). Modify this function so that `p_xy` correctly computes the conditional probabilities of these values based on the frequencies in the data set. Submit your code and the validation error that you obtain. Also, compare your validation error to what you obtain with scikit-learn's implementation, `BernoulliNB`.

- Naive Bayes (ours) validation error: 0.188 Scikit learn bernouilli nb validation error: 0.187

2.4 Runtime of Naive Bayes for Discrete Data

Rubric: {reasoning:3}

For a given training example i, the predict function in the provided code computes the quantity

$$p(y_i \mid x_i) \propto p(y_i) \prod_{j=1}^{d} p(x_{ij} \mid y_i),$$

for each class y_i (and where the proportionality constant is not relevant). For many problems, a lot of the $p(x_{ij} \mid y_i)$ values may be very small. This can cause the above product to underflow. The standard fix for this is to compute the logarithm of this quantity and use that $\log(ab) = \log(a) + \log(b)$,

$$\log p(y_i \mid x_i) = \log p(y_i) + \sum_{j=1}^{d} \log p(x_{ij} \mid y_i) + (\text{irrelevant propportionality constant}) .$$

This turns the multiplications into additions and thus typically would not underflow.

Assume you have the following setup:

- The training set has n objects each with d features.
- The test set has t objects with d features.
- Each feature can have up to c discrete values (you can assume $c \leq n$).
- There are k class labels (you can assume $k \leq n$)

You can implement the training phase of a naive Bayes classifier in this setup in $O(nd)$, since you only need to do a constant amount of work for each $X(i, j)$ value. (You do not have to actually implement it in this way for the previous question, but you should think about how this could be done.) What is the cost of classifying t test examples with the model and this way of computing the predictions?

- The runtime for classifying t example is $O(tdk)$. This is because we are trying to determine the cost of calculating $p(x_ij|y_i)$. In order to determine this we require three for loops. The first for loop going through every single label k, the second one going through every single feature d and the third label going through every single example d. as these are nested, the runtime can be found to be $O(tdk)$

3 K-Nearest Neighbours

Rubric: {code:3, reasoning:4}

In the *citiesSmall* dataset, nearby points tend to receive the same class label because they are part of the same U.S. state. For this problem, perhaps a k-nearest neighbours classifier might be a better choice than a decision tree. The file *knn.py* has implemented the training function for a k-nearest neighbour classifier (which is to just memorize the data).

Fill in the `predict` function in `knn.py` so that the model file implements the k-nearest neighbour prediction rule. You should Euclidean distance, and may numpy's `sort` and/or `argsort` functions useful. You can also use `utils.euclidean_dist_squared`, which computes the squared Euclidean distances between all pairs of points in two matrices.

1. Write the `predict` function.
2. Report the training and test error obtained on the *citiesSmall* dataset for $k = 1$, $k = 3$, and $k = 10$. How do these numbers compare to what you got with the decision tree?

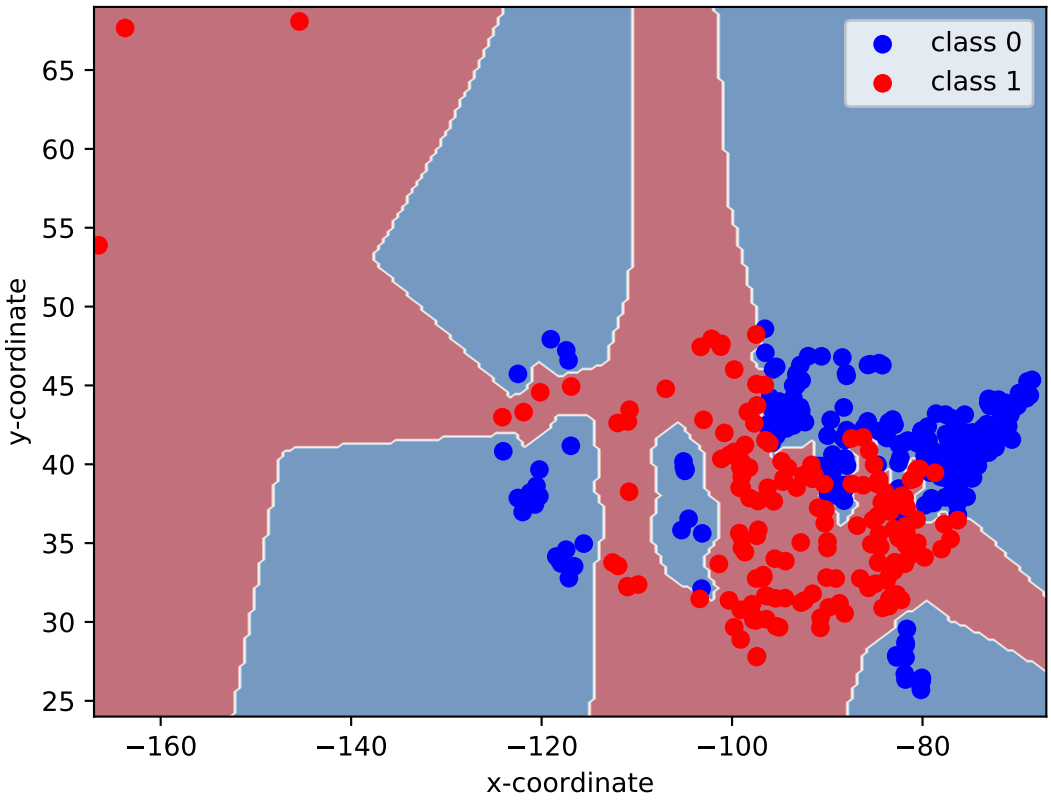
3. Hand in the plot generated by `utils.plotClassifier` on the *citiesSmall* dataset for $k = 1$, using both your implementation of KNN and the KNeighborsClassifier from scikit-learn.
4. Why is the training error 0 for $k = 1$?
5. If you didn't have an explicit test set, how would you choose k?

- 1) Code is submitted.

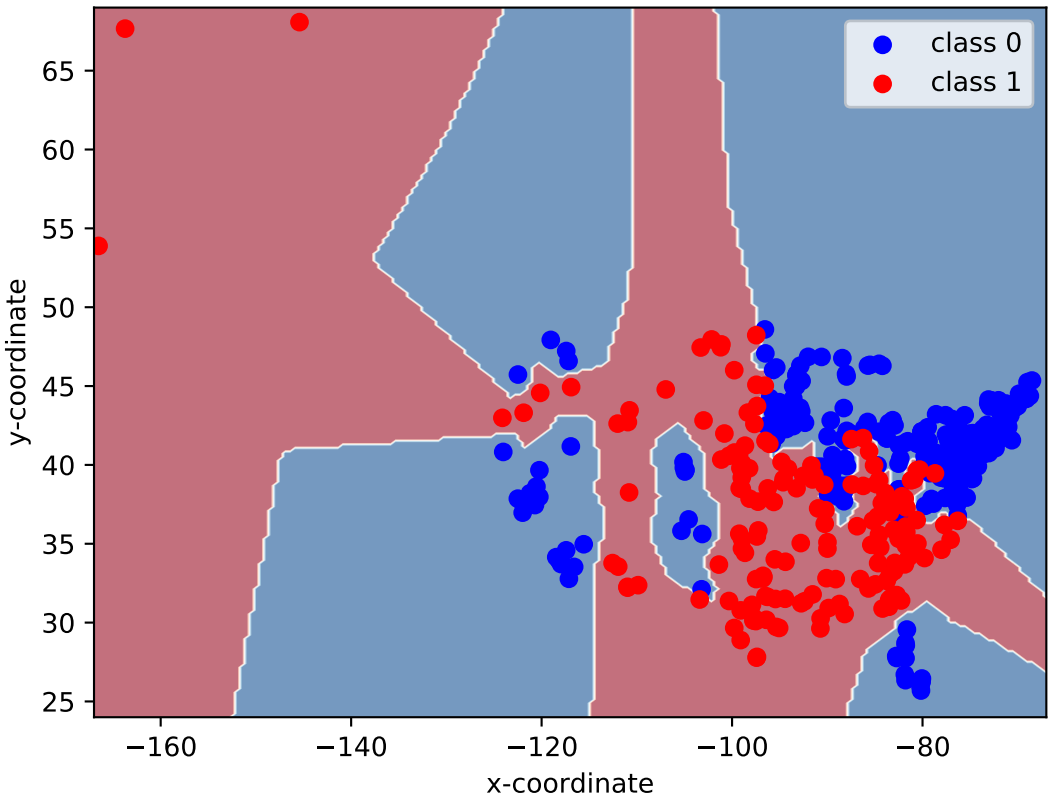
  2) N=1 test error: 0.065 Training error: 0.000 N=3 test error: 0.066 Training error: 0.028 N = 10 KNN test error: 0.097 Training error: 0.072

  The decision tree catches up with these tests errors after depth=8.Also, as k goes higher the training error goes to 0 with the decision tree but it does not become zero with KNN

  3) KNN with k=1



scikit-learn graph



4)

because the elements just get assigned to their own label so there is no room for error

5) I would eyeball the graph trying to see if I can select k by looking at how the data points are clustered and vary by location

4 Random Forests

4.1 Implementation

Rubric: {code:4,reasoning:3}

The file *vowels.pkl* contains a supervised learning dataset where we are trying to predict which of the 11 "steady-state" English vowels that a speaker is trying to pronounce.

You are provided with a `RandomStump` class that differs from `DecisionStumpInfoGain` in that it only considers $\lfloor \sqrt{d} \rfloor$ randomly-chosen features.[1] You are also provided with a `RandomTree` class that is exactly the same as `DecisionTree` except that it uses `RandomStump` instead of `DecisionStump` and it takes a bootstrap sample of the data before fitting. In other words, `RandomTree` is the entity we discussed in class, which makes up a random forest.

[1]The notation $\lfloor x \rfloor$ means the "floor" of x, or "x rounded down". You can compute this with `np.floor(x)` or `math.floor(x)`.

If you run `python main.py -q 4` it will fit a deep `DecisionTree` using the information gain splitting criterion. You will notice that the model overfits badly.

1. Why doesn't the random tree model have a training error of 0?
2. Create a class `RandomForest` in a file called `random_forest.py` that takes in hyperparameters `num_trees` and `max_depth` and fits `num_trees` random trees each with maximum depth `max_depth`. For prediction, have all trees predict and then take the mode.
3. Using 50 trees, and a max depth of ∞, report the training and testing error. Compare this to what we got with a single `DecisionTree` and with a single `RandomTree`. Are the results what you expected? Discuss.
4. Compare your implementation with scikit-learn's `RandomForestClassifier` for both speed and accuracy, and briefly discuss. You can use all default hyperparameters if you wish, or you can try changing them.

- 4)

  4.1] Because we use bootstrapping. the fact that we use infinite depth tree does not overfit since the data we overfit on is randomly generated.

  4.2] The code is submitted

  4.3] Single Random Tree: Training error: 0.148 Testing error: 0.515

  Decision Tree: Training error: 0.011 Testing error: 0.443

  Random Forest: Training error: 0.000 Testing error: 0.205

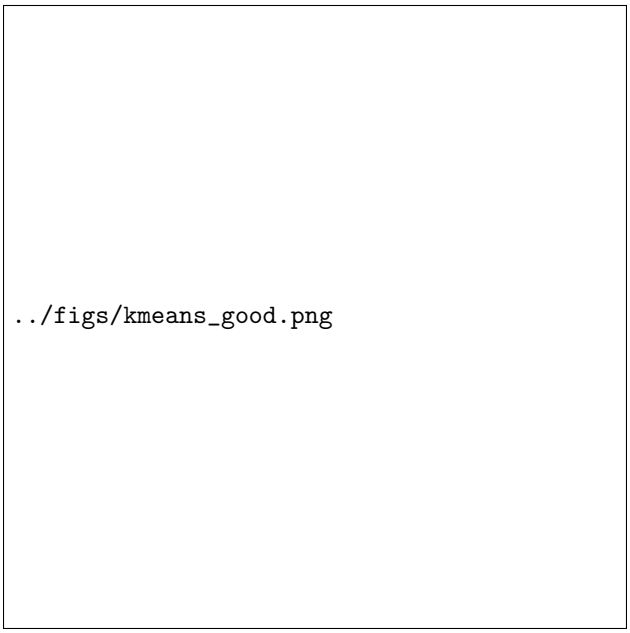
  The training errors are as I expected yet the difference between the decision tree and single random tree is unexpected. I would expect the single random tree to have a smaller testing error than the decision tree. Also, random forest does exceptionally well meaning very little testing error.

  4) Training error: 0.000 Testing error: 0.159 so the errors are the same but it is way faster than my implementation

5 Clustering

If you run `python main.py -q 5`, it will load a dataset with two features and a very obvious clustering structure. It will then apply the k-means algorithm with a random initialization. The result of applying the algorithm will thus depend on the randomization, but a typical run might look like this:

(Note that the colours are arbitrary – this is the label switching issue.) But the 'correct' clustering (that was used to make the data) is this:



5.1 Selecting among k-means Initializations

Rubric: {reasoning:5}

If you run the demo several times, it will find different clusterings. To select among clusterings for a *fixed* value of k, one strategy is to minimize the sum of squared distances between examples x_i and their means

w_{y_i},

$$f(w_1, w_2, \ldots, w_k, y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} \|x_i - w_{y_i}\|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - w_{y_i j})^2.$$

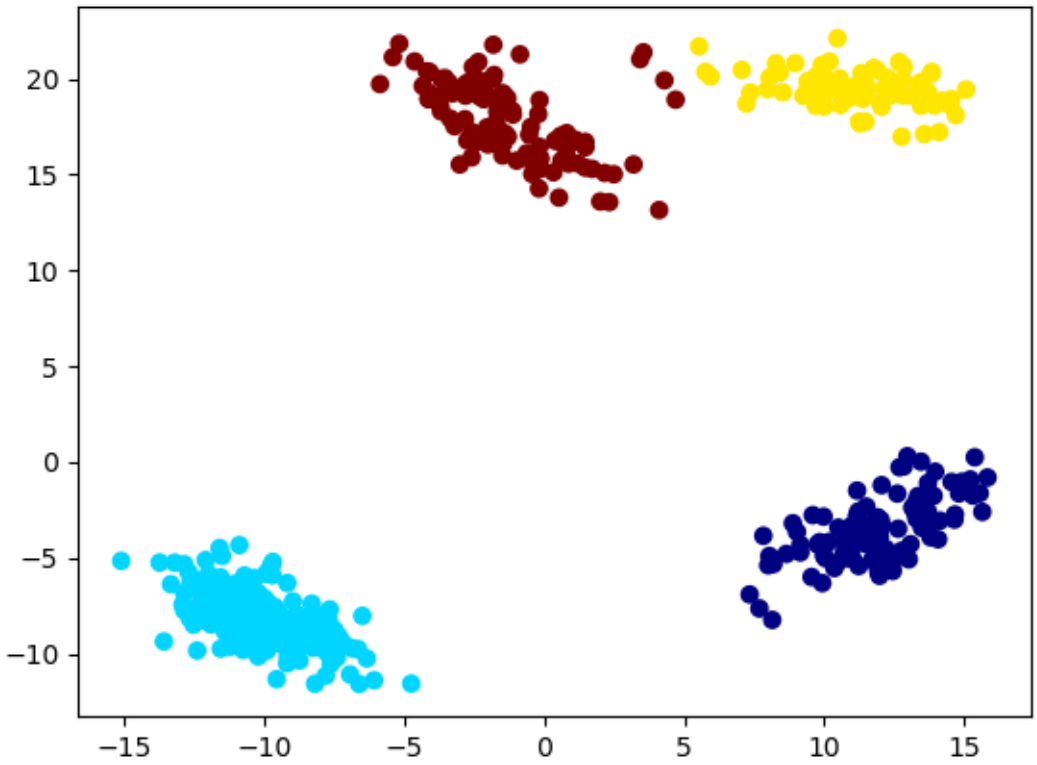
where y_i is the index of the closest mean to x_i. This is a natural criterion because the steps of k-means alternately optimize this objective function in terms of the w_c and the y_i values.

1. In the `kmeans.py` file, add a new function called `error` that takes the same input as the `predict` function but that returns the value of this above objective function.
2. What trend do you observe if you print the value of this error after each iteration of the k-means algorithm?
3. Using the `plot_2dclustering` function defined in `main.py`, output the clustering obtained by running k-means 50 times (with $k = 4$) and taking the one with the lowest error. Submit your plot.
4. Looking at the hyperparameters of scikit-learn's `KMeans`, explain the first four (`n_clusters`, `init`, `n_init`, `max_iter`) very briefly.

- 1) code is submitted

  2) The change in error is not uniform and it decreases and increases with the iterations; finally converging to some value after which the loop stops running

3)



4) n-clusters : The number of clusters to form as well as the number of means we are using to label

init : Method for initialization where k-means++ is the one in the bonus slides, random is what we did in the lecture and ndarray is the array of preselected means

n-init : how many times each mean gets updated

max-iter : how many times in total the algorithm iterates through the data

5.2 Selecting k in k-means

Rubric: {reasoning:5}

We now turn to the task of choosing the number of clusters k.

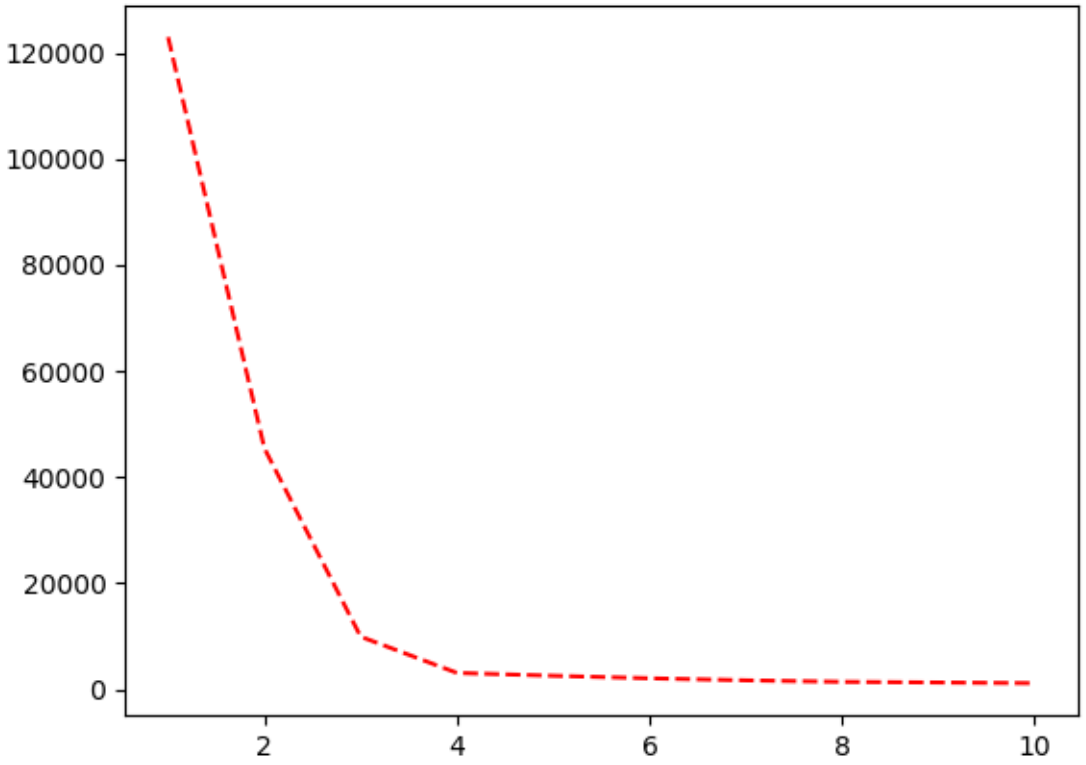
1. Explain why we should not choose k by taking the value that minimizes the `error` function.
2. Explain why even evaluating the `error` function on test data still wouldn't be a suitable approach to choosing k.
3. Hand in a plot of the minimum error found across 50 random initializations, as a function of k, taking k from 1 to 10.

4. The *elbow method* for choosing k consists of looking at the above plot and visually trying to choose the k that makes the sharpest "elbow" (the biggest change in slope). What values of k might be reasonable according to this method? Note: there is not a single correct answer here; it is somewhat open to interpretation and there is a range of reasonable answers.

- 1) Because error function is related to the training set. We want to minimize the test-error. Minimizing the training error might lead to overfitting

  2) This is the violation of the golden rule. We should ignore the test data until we tune our model and the hyper-parameters. If we use the k value that minimizes the test-error that is either violating the golden rule or treating the test set as training set; both of which are dubious methods.

  3)



  4)The biggest change in slope happens at k=3

  I think 3,4,5 could be reasonable guesses for the optimal k value because the higher the k value is the higher the risk for overfitting

5.3 Density-Based Clustering

Rubric: {reasoning:2}

If you run `python main.py -q 5.3`, it will apply the basic density-based clustering algorithm to the dataset from the previous part, but with some outliers added. The final output should look somewhat like this:

(The right plot is zoomed in to show the non-outlier part of the data.) Even though we know that each object was generated from one of four clusters (and we have 4 outliers), the algorithm finds 6 clusters and does not assign some of the original non-outlier objects to any cluster. However, the clusters will change if we change the parameters of the algorithm. Find and report values for the two parameters, `eps` (which we called the "radius" in class) and `minPts`, such that the density-based clustering method finds:

1. The 4 "true" clusters.
2. 3 clusters (merging the top two, which also seems like a reasonable interpretaition).
3. 2 clusters.
4. 1 cluster (consisting of the non-outlier points).

- 5.3.1] Eps= 2 gives the 4 true clusters

  5.3.2]Eps=5 gives 3 clusters

  5.3.3]Eps = 13 gives 2 clusters

  5.3.4]Eps = 20 and $\min_sample = 100 excludetheoutliersandgivesonecluster$

6 Very-Short Answer Questions

Rubric: {reasoning:13}

Write a short one or two sentence answer to each of the questions below. Make sure your answer is clear and concise.

1. What is an advantage of using a boxplot to visualize data rather than just computing its mean and variance?
2. What is a reason that the the data may not be IID in the email spam filtering example from lecture?
3. What is the difference between a validation set and a test set?
4. Why can't we (typically) use the training error to select a hyper-parameter?
5. What is the effect of n on the optimization bias (assuming we use a parametric model).
6. What is an advantage and a disadvantage of using a large k value in k-fold cross-validation.
7. Why can we ignore $p(x_i)$ when we use naive Bayes?
8. For each of the three values below in a naive Bayes model, say whether it's a parameter or a hyper-parameter:
   (a) Our estimate of $p(y_i)$ for some y_i.
   (b) Our estimate of $p(x_{ij} \mid y_i)$ for some x_{ij} and y_i.
   (c) The value β in Laplace smoothing.
9. What is the effect of k in KNN on the two parts (training error and approximation error) of the fundamental trade-off. Hint: think about the extreme values.
10. Suppose we want to classify whether segments of raw audio represent words or not. What is an easy way to make our classifier invariant to small translations of the raw audio?

11. Both supervised learning and clustering models take in an input x_i and produce a label y_i. What is the key difference?

12. Suppose you chose k in k-means clustering (using the squared distances to examples) from a validation set instead of a training set. Would this work better than using the training set (which just chooses the largest value of k)?

13. In k-means clustering the clusters are guaranteed to be convex regions. Are the areas that are given the same label by KNN also convex?

- 1. The obvious advantage is you can see the outliers visually. Also, boxplot shows data in clusters which makes it easier to see where the bulk of data is compared to mean which might not be the perfect measure for certain data type such as a list of 10 -10s and 10 10s, the mean is 0 but there is no element which is close to 0

  2. Assuming this question is asked about word-label pairs: meaning word-label pairings are not i.i.d. If it is the case that some group of words are used together in spam messages than not, the i.i.d. assumption is violated

  3. Validation set is part of our training set and we use it to check how well we could do in the test set. Whereas test set is not used until we are confident about our training set and it is not part of the training set it is a whole new set which we have not seen

  4. Because, we can overfit meaning try too many things and get one that works really well by chance or over-trying and get a training error of almost zero but it does not mean that we will do well in the test set.

  5. Optimization bias shrinks as the number of validation examples increases. So, given that we choose a fixed percentage of the training set (say 20

  6. Advantage could be that we now try so many different cross-validations sets that it is very likely that our estimate will be close to the actual test error assuming our data is i.i.d

  7. $p(y_i = spam|x_i)$ to $p(y_i = non-spam|x_i)$ When we use the Bayes rule both denominators are $p(x_i)$ then we can just simplify these two so no need to calculate $p(x_i)$. An exception would be when $p(x_i)$ is 0 but we can overcome this with laplace smoothing

  8. (a) parameter

  (b) parameter

  (c) hyperparameter

  9. As k increases, the training error decreases and approximation error increases.

  10. The blank background noise in the raw audios are likely to be the same across different segments of raw audio. So if we crop these parts out of all the segments we are only left with sounds that are different from silent parts or blank background noise. This accounts for the fact that there could be time translations. Also we can just add the translated versions of the raw audios to the training set so the model starts recognizing these

  11. In clustering models, we do not know the labels y_i

  12. Assuming the training sample is i.i.d, (training sample = training set+ validation set); the success of the model just depends on how big the training set and validation sets are. So, if they are equally split (50-50 split) then they would work just equally well. If the training set is bigger ,by the law of large numbers, the means are likely to be more accurate than the means that could be obtained by using the validation set.

  13. No, KNN areas can be concave, refer to the lecture demos