CM146, Fall 2020
Problem Set 1: Decision trees and k-Nearest Neighbors
Due Oct. 29, 2020 at 11:59 pm

Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- Please package your code (.py) for Problem 5 and submit it to CCLE.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

Parts of this assignment are adapted from course material by Andrea Danyluk (Williams), Tom Mitchell, Matt Gormley and Maria-Florina Balcan (CMU), Stuart Russell (UC Berkeley), Carlos Guestrin (UW), Dan Roth (UPenn) and Jessica Wu (Harvey Mudd).

1 Splitting Heuristic for Decision Trees [20 pts]

Recall that the ID3 algorithm iteratively grows a decision tree from the root downwards. On each iteration, the algorithm replaces one leaf node with an internal node that splits the data based on one decision attribute (or feature). In particular, the ID3 algorithm chooses the split that reduces the entropy the most, but there are other choices. For example, since our goal in the end is to have the lowest error, why not instead choose the split that reduces error the most? In this problem, we will explore one reason why reducing entropy is a better criterion.

Consider the following simple setting. Let us suppose each example is described by n boolean features: $X = \langle X_1, \ldots, X_n \rangle$, where $X_i \in \{0, 1\}$, and where $n \geq 4$. Furthermore, the target function to be learned is $f : X \to Y$, where $Y = X_1 \vee X_2 \vee X_3$. That is, $Y = 1$ if $X_1 = 1$ or $X_2 = 1$ or $X_3 = 1$, and $Y = 0$ otherwise. Suppose that your training data contains all of the 2^n possible examples, each labeled by f. For example, when $n = 4$, the data set would be

X_1	X_2	X_3	X_4	Y
0	0	0	0	0
1	0	0	0	1
0	1	0	0	1
1	1	0	0	1
0	0	1	0	1
1	0	1	0	1
0	1	1	0	1
1	1	1	0	1

X_1	X_2	X_3	X_4	Y
0	0	0	1	0
1	0	0	1	1
0	1	0	1	1
1	1	0	1	1
0	0	1	1	1
1	0	1	1	1
0	1	1	1	1
1	1	1	1	1

(a) **(5 pts)** How many mistakes does the best 1-leaf decision tree make over the 2^n training examples? (The 1-leaf decision tree does not split the data even once. Make sure you answer for the general case when $n \geq 4$.)

Solution: For the general case when $n \geq 4$, there are 2^{n-3} cases where $Y = 0$ and $2^n - 2^{n-3}$ cases where $Y = 1$. When $n \geq 4$, $2^n - 2^{n-3} > 2^{n-3}$, it will assign 1 to Y, so the best 1-leaf decision tree makes $\boxed{2^{n-3}}$ mistakes.

(b) **(5 pts)** Is there a split that reduces the number of mistakes by at least one? (That is, is there a decision tree with 1 internal node with fewer mistakes than your answer to part (a)?) Why or why not?

Solution: $\boxed{\text{No}}$

- If we split at X_i where $4 \leq i \leq n$, it won't make a difference since Y is independent of X_i where $4 \leq i \leq n$. Both splitting edges will assign value of 1 to Y since the majority of examples favor 1 under both edges. Number of mistakes remains at 2^{n-3}.
- If we split at X_i where $1 \leq i \leq 3$, for $X_i = 1$, it will assign 1 to Y since all the training data under this edge favors 1, for $X_i = 0$, there are 2^{n-3} cases where $Y = 0$ and $2^{n-1} - 2^{n-3}$ cases where $Y = 1$. When $n \geq 4$, $2^{n-1} - 2^{n-3} > 2^{n-3}$, so it will still assign 1 to Y under this edge. Number of mistakes remains at 2^{n-3}.

(c) **(5 pts)** What is the entropy of the output label Y for the 1-leaf decision tree (no splits at all)?

Solution:

$$\begin{aligned} H(S) &= -P_+ \log_2 P_+ - P_- \log_2 P_- \\ &= -\frac{2^n - 2^{n-3}}{2^n} \log_2 \frac{2^n - 2^{n-3}}{2^n} - \frac{2^{n-3}}{2^n} \log_2 \frac{2^{n-3}}{2^n} \\ &= -\frac{7}{8} \log_2 \frac{7}{8} - \frac{1}{8} \log_2 \frac{1}{8} \\ &\approx 0.54356 \end{aligned}$$

https://www.overleaf.com/project/5f8d8f5f3f10b50001ff23d6

(d) **(5 pts)** Is there a split that reduces the entropy of the output Y by a non-zero amount? If so, what is it, and what is the resulting conditional entropy of Y given this split?

Solution: Splitting at either X_1 or X_2 or X_3 will reduce the entropy of the output Y by a non-zero amount.

$$\begin{aligned} H(S|X_1 = 0) &= -\frac{2^{n-1} - 2^{n-3}}{2^{n-1}} \log_2 \frac{2^{n-1} - 2^{n-3}}{2^{n-1}} - \frac{2^{n-3}}{2^{n-1}} \log_2 \frac{2^{n-3}}{2^{n-1}} \\ &= -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \\ &\approx 0.811278 \\ H(S|X_1 = 1) &= -\frac{2^{n-1}}{2^{n-1}} \log_2 \frac{2^{n-1}}{2^{n-1}} - \frac{0}{2^{n-1}} \log_2 \frac{0}{2^{n-1}} = 0 \\ H(S|X_1) &= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0.811278 \approx 0.405639 \\ Gain(S, X_1) &= 0.54356 - 0.405639 = 0.137921 \end{aligned}$$

The resulting conditional entropy of Y given this split is $\boxed{0.405639}$, information gain is 0.137921.

2 Entropy and Information [5 pts]

The entropy of a Bernoulli (Boolean 0/1) random variable X with $p(X = 1) = q$ is given by

$$B(q) = -q \log q - (1 - q) \log(1 - q).$$

Suppose that a set S of examples contains p positive examples and n negative examples. The entropy of S is defined as $H(S) = B\left(\frac{p}{p+n}\right)$.

(a) **(5 pts)** Based on an attribute X_j, we split our examples into k disjoint subsets S_k, with p_k positive and n_k negative examples in each. If the ratio $\frac{p_k}{p_k+n_k}$ is the same for all k, show that the information gain of this attribute is 0.

Solution: Since the ratio $\frac{p_k}{p_k+n_k}$ is the same for all k, we know that:

$\frac{p_1}{p_1+n_1} = \frac{p_2}{p_2+n_2} = \frac{p_3}{p_3+n_3} = \ldots = \frac{p_k}{p_k+n_k} = \frac{p_1+p_2+\ldots+p_k}{p_1+n_1+p_2+n_2+\ldots+p_k+n_k} = \frac{p}{p+n}$

$Gain(S, X_j) = B(\frac{p}{p+n}) - \sum_{i=1}^{k} \frac{p_i+n_i}{p+n} B(\frac{p_i}{p_i+n_i}) = B(\frac{p}{p+n}) - \sum_{i=1}^{k} \frac{p_i+n_i}{p+n} B(\frac{p}{p+n})$

$= B(\frac{p}{p+n}) - B(\frac{p}{p+n}) \underbrace{\sum_{i=1}^{k} \frac{p_i+n_i}{p+n}}_{1} = B(\frac{p}{p+n}) - B(\frac{p}{p+n}) = 0$

so the information gain of this attribute is 0.

3 k-Nearest Neighbors and Cross-validation [10 pts]

In the following questions you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. Note that a point can be its own neighbor.

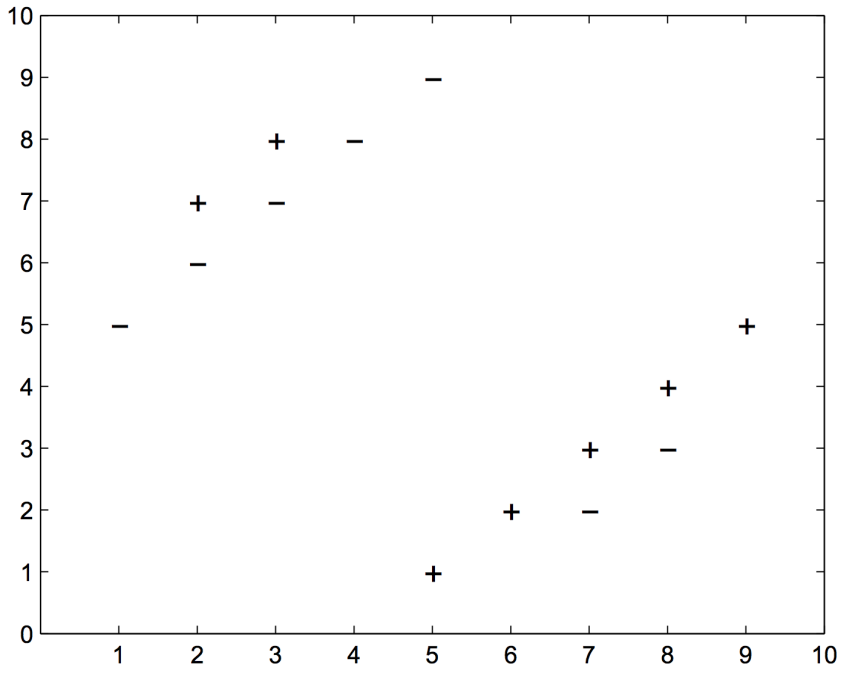


Figure 1: Dataset for KNN binary classification task.

(a) **(2.5 pts)** What will be the label of point (7,3) in Fig 1 using K-NN algorithm with majority voting when K=3?
Solution: Its three nearest neighbors are (7,3), (7,2), (8,3), two negative and one positive, so the label of point (7,3) will be negative, which is "-".

(b) **(2.5 pts)** What value of k minimizes the training set error for this dataset? What is the resulting training error?

Solution: Since a point can be its own neighbor, $\boxed{k = 1}$ can minimizes the training set error for this dataset. All the training data can be perfectly satisfied because it can be its own nearest neighbor. The resulting training error will be 0.

(c) **(2.5 pts)** Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?

Solution: Using too large values k might cause under-fitting of the training data. The model will be too simple and hence perform badly. Using too small values k might cause over-fitting of the training data. The model will be too complex and might perform perfectly on the training data but badly on the test data because there may be noise in the training data.

(d) **(2.5 pts)** What value of k minimizes leave-one-out cross-validation error for this dataset? What is the resulting error?

Solution: k=5. The resulting error is $\frac{4}{14} = 0.2857$

4 Decision Tree [15 pts]

In a binary classification problem, there are 4000 examples in the class with label 1 and 8000 examples in the class with label 0. Recall that the information gain for target label Y and feature X is defined as $Gain = H[Y] - H[Y|X]$, where $H[Y] = -E[\log_2 P(Y)]$ is the entropy.

(a) **(2 pts)** What is the entropy of the class variable Y?

Solution: $H[Y] = -\frac{4000}{4000+8000}\log_2\frac{4000}{4000+8000} - \frac{8000}{4000+8000}\log_2\frac{8000}{4000+8000} = 0.9183$

(b) **(5 pts)** Let's consider a binary feature A for this problem. In the negative class (with label 0), the number of instances that have $A = 1$ and $A = 0$ respectively: $(4000, 4000)$. In the positive class (with label 1), these numbers are: $(4000, 0)$. Write down conditional entropy and information gain of A relative to Y?

Solution:

$$H(S|A=0) = -\frac{0}{0+4000}\log_2\frac{0}{0+4000} - \frac{4000}{0+4000}\log_2\frac{4000}{0+4000} = 0$$
$$H(S|A=1) = -\frac{4000}{4000+4000}\log_2\frac{4000}{4000+4000} - \frac{4000}{4000+4000}\log_2\frac{4000}{4000+4000} = 1$$
$$H(S|A) = \frac{4000}{4000+8000}\cdot 0 + \frac{8000}{4000+8000}\cdot 1 = \frac{2}{3} = 0.6667$$
$$Gain(S, A) = 0.9183 - 0.6667 = 0.2516$$

Conditional entropy is 0.6667. Information gain is 0.2516.

(c) **(5 pts)** Let's consider another binary feature B. In the negative class (with label 0) , the number of instances that have $B = 0$ and $B = 1$ respectively are: $(6000, 2000)$. In the positive class (with label 1), these numbers are: $(3000, 1000)$. Write down conditional entropy and

information gain of B relative to Y?

Solution:

$$H(S|B=0) = -\frac{3000}{3000+6000}\log_2\frac{3000}{3000+6000} - \frac{6000}{3000+6000}\log_2\frac{6000}{3000+6000} = 0.9183$$
$$H(S|B=1) = -\frac{1000}{1000+2000}\log_2\frac{1000}{1000+2000} - \frac{2000}{1000+2000}\log_2\frac{2000}{1000+2000} = 0.9183$$
$$H(S|B) = \frac{9000}{12000}\cdot 0.9183 + \frac{3000}{12000}\cdot 0.9183 = 0.9183$$
$$Gain(S,A) = 0.9183 - 0.9183 = 0$$

Conditional entropy is 0.9183. Information gain is 0.

(d) **(3 pts)** Using information gain, which attribute will the ID3 decision tree learning algorithm choose at first?

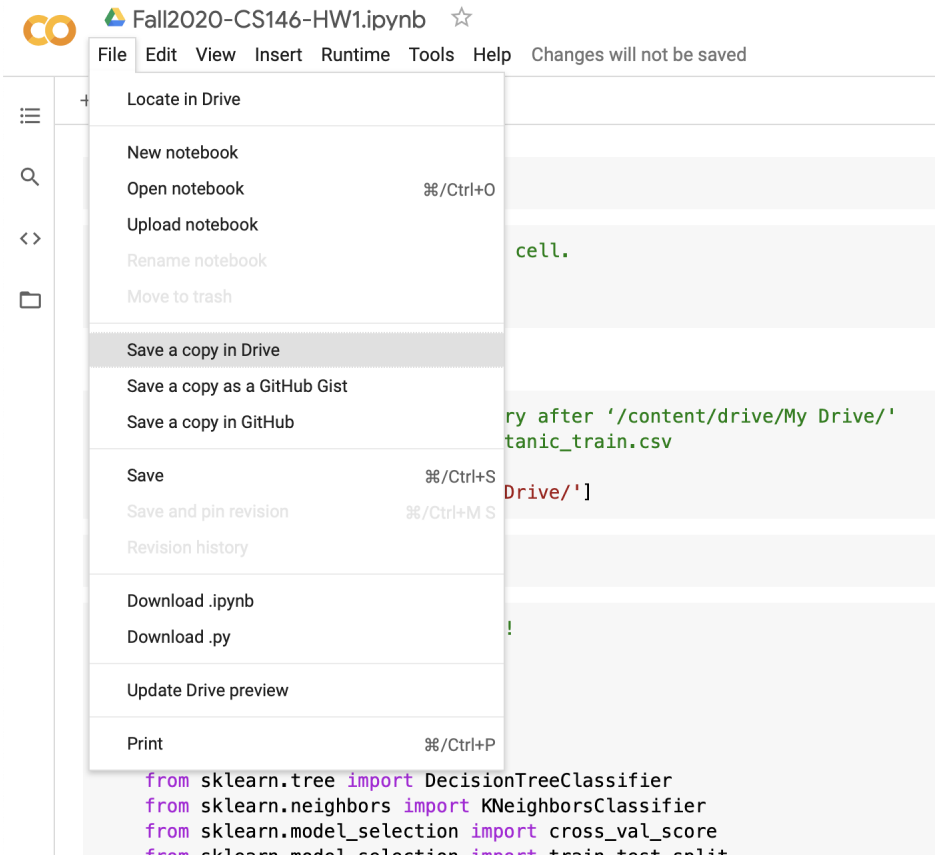
Solution: The ID3 decision tree learning algorithm will choose attribute A first because A has greater information gain relative to Y.

5 Programming exercise : Applying decision trees and k-nearest neighbors [50 pts]

To work on this HW: you need to download two files (i) nutil.py (ii) adult_subsample.csv from here. Then copy/upload them to you own Google drive.

Next, for all the coding, please refer to the following colab notebook Fall2020-CS146-HW1.ipynb.

Before executing or writing down any code, please make a copy of the notebook and save it to your own google drive by clicking the "File" → "Save a copy in Drive".



You will then be prompted to log into your google account. Please make sure all the work you implement is done on your own saved copy. You won't to able to make changes on the the original notebook shared for the entire class. Running the first two cells will further mount your own google drive so that your copy of the Colab notebook will have access to the two files (nutil.py and adult_subsample.csv) you've just uploaded.

The notebook has marked blocks where you need to code.

$========= TODO : START =========$

$========= TODO : END ==========$

Submission instructions for programming problems

- Please export the notebook to a `.py` file by clicking the "File" → "Download.py" and upload to CCLE.

  Your code should be commented appropriately. The most important things:
  - Your name and the assignment number should be at the top of each file.
  - Each class and method should have an appropriate doctsring.
  - If anything is complicated, it should include some comments.

  There are many possible ways to approach the programming portion of this assignment, which makes code style and comments very important so that staff can understand what you did. For this reason, you will lose points for poorly commented or poorly organized code.

- Please submit all the plots and the rest of the solutions (other than codes) to Gradescope

For the questions please read below.

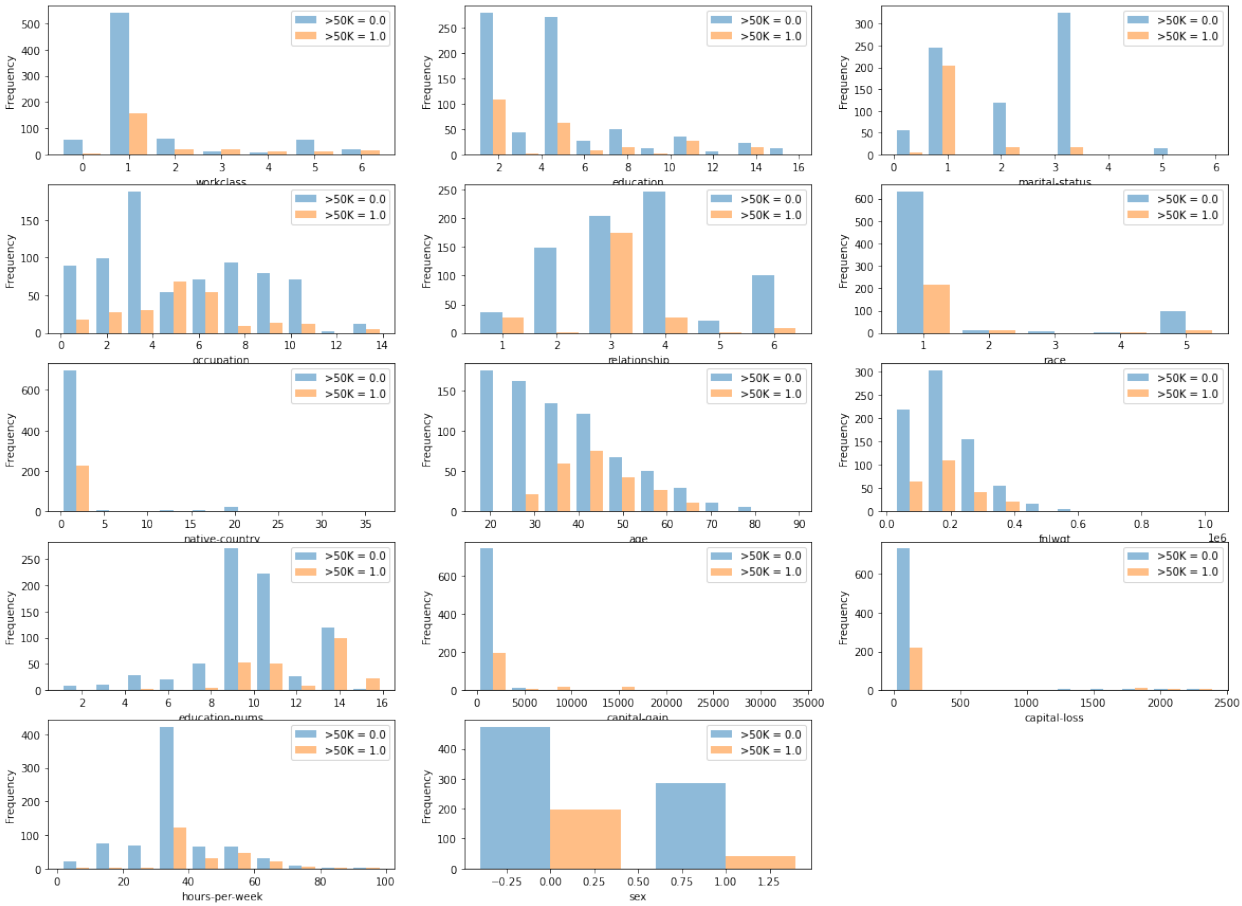
5.1 Visualization [5 pts]

One of the first things to do before trying any formal machine learning technique is to dive into the data. This can include looking for funny values in the data, looking for outliers, looking at the range of feature values, what features seem important, etc.

Note: We have already converted all the categorical features to numerical ones. The target column is the last one: ">50k", where 1 and 0 indicate >50k or $\leq$ 50k respectively. The feature "fnlwgt" describes the number of people the census believes the entry represents. All the other feature names should be self-explanatory. If you want to learn more about this data please click here

(a) **(5 pts)** Make histograms for each feature, separating the examples by class (e.g. income greater than 50k or smaller than or equal to 50k). This should produce fourteen plots, one for each feature, and each plot should have two overlapping histograms, with the color of the histogram indicating the class. For each feature, what trends do you observe in the data? (Please only describe the general trend. No need for more than two sentences per feature)

Solution:



- **Workclass**: most people are within workclass 1 and only a small amount of them can earn greater than 50k.
- **education**: most people are within education 2 and education 5 and less than half of them can earn greater than 50k.
- **marital-status**: most people who earn more than 50k are within marital-status 1 (Divorced).
- **occupation**: people who earn more than 50k are approximately in normal distribution under occupation feature.
- **relationship**: most people who earn more than 50k are within relationship 3 (Husband).
- **race**: most people are within race 1(white)and most people who earn more than 50k are also white.
- **native-country**: most people are within native-country 1 and most people who earn more than 50k are also from native-country 1.
- **age**: people who earn less than 50k are right skewed and people who earn more than 50k are approximately normal.
- **fnlwgt**: people who earn less than and more than 50k are both right skewed.
- **education-nums**: people who earn less than and more than 50k are both left skewed..
- **capital-gain**: most people are between capital-gain 0 and 5000 and most people who earn more than 50k are also between capital-gain 0 and 5000.

- **capital-loss**: most people are between capital-loss 0 and 500 and most people who earn more than 50k are also between capital-loss 0 and 500.
- **hours-per-week**: people who earn less than 50k are approximately normal and most people who earn more than 50k work more than 30 hours per week.
- **sex**: compared with the male ones, a larger proportion of female people can earn more than 50k.

5.2 Evaluation [45 pts]

Now, let's use `scikit-learn` to train a `DecisionTreeClassifier` and `KNeighborsClassifier` on the data.

Using the predictive capabilities of the `scikit-learn` package is very simple. In fact, it can be carried out in three simple steps: initializing the model, fitting it to the training data, and predicting new values.[1]

(b) **(0 pts)** Before trying out any classifier, it is often useful to establish a *baseline*. We have implemented one simple baseline classifier, `MajorityVoteClassifier`, that always predicts the majority class from the training set. Read through the `MajorityVoteClassifier` and its usage and make sure you understand how it works.

Your goal is to implement and evaluate another baseline classifier, `RandomClassifier`, that predicts a target class according to the distribution of classes in the training data set. For example, if 85% of the examples in the training set have `>50k = 0` and 15% have `>50k = 1`, then, when applied to a test set, `RandomClassifier` should randomly predict 85% of the examples as `>50k = 0` and 15% as `>50k = 1`.

Implement the missing portions of `RandomClassifier` according to the provided specifications. Then train your `RandomClassifier` on the entire training data set, and evaluate its training error. If you implemented everything correctly, you should have an error of 0.374.

(c) **(10 pts)** Now that we have a baseline, train and evaluate a `DecisionTreeClassifier` (using the class from `scikit-learn` and referring to the documentation as needed). Make sure you initialize your classifier with the appropriate parameters; in particular, use the 'entropy' criterion discussed in class. What is the training error of this classifier?

Solution: Training error: 0.000.

(d) **(5 pts)** Similar to the previous question, train and evaluate a `KNeighborsClassifier` (using the class from `scikit-learn` and referring to the documentation as needed). Use k=3, 5 and 7 as the number of neighbors and report the training error of this classifier.

Solution:

$k = 3$, training error: 0.153
$k = 5$, training error: 0.195
$k = 7$, training error: 0.213

(e) **(10 pts)** So far, we have looked only at training error, but as we learned in class, training error is a poor metric for evaluating classifiers. Let's use cross-validation instead.

Implement the missing portions of `error(...)` according to the provided specifications. You may find it helpful to use `StratifiedShuffleSplit(...)` from `scikit-learn`. To ensure that we always get the same splits across different runs (and thus can compare the classifier results), set the `random_state` parameter to be the same (e.g., 0).

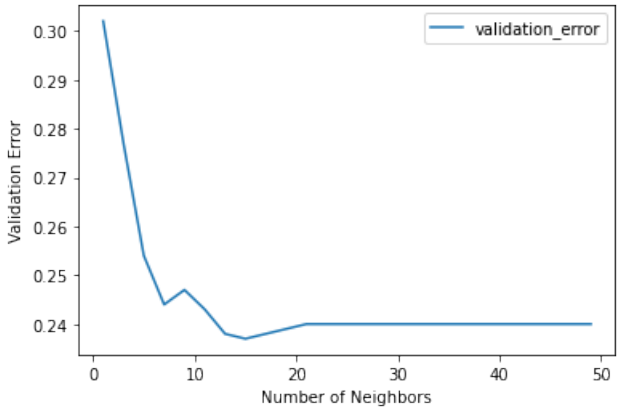
[1] Note that almost all of the model techniques in `scikit-learn` share a few common named functions, once they are initialized. You can always find out more about them in the documentation for each model. These are `some-model-name.fit(...)`, `some-model-name.predict(...)`, and `some-model-name.score(...)`.

Next, use your `error(...)` function to evaluate the training error and (cross-validation) test error and test micro averaged F1 Score (If you dont know what is F1, please click here) of each of your four models (for the `KNeighborsClassifier`, use k=5). To do this, generate a random 80/20 split of the training data, train each model on the 80% fraction, evaluate the error on either the 80% or the 20% fraction, and repeat this 100 times to get an average result. What are the average training and test error of each of your classifiers on the adult_subsample data set?

Solution:
Investigating RandomClassifier
– training error: 0.375
– test error: 0.382
– f1 score: 0.618
Investigating MajorityVoteClassifier
– training error: 0.240
– test error: 0.240
– f1 score: 0.760
Investigating DecisionTreeClassifier
– training error: 0.000
– test error: 0.205
– f1 score: 0.795
Investigating KNeighborsClassifier(k = 5)
– training error: 0.202
– test error: 0.259
– f1 score: 0.741

(f) **(5 pts)** One way to find out the best value of k for `KNeighborsClassifier` is n-fold cross validation. Find out the best value of k using 10-fold cross validation. You may find the `cross_val_score(...)` from `scikit-learn` helpful. Run 10-fold cross validation for all odd numbers ranging from 1 to 50 as the number of neighbors. Then plot the validation error against the number of neighbors, k. Include this plot in your writeup, and provide a 1-2 sentence description of your observations. What is the best value of k?

Solution:

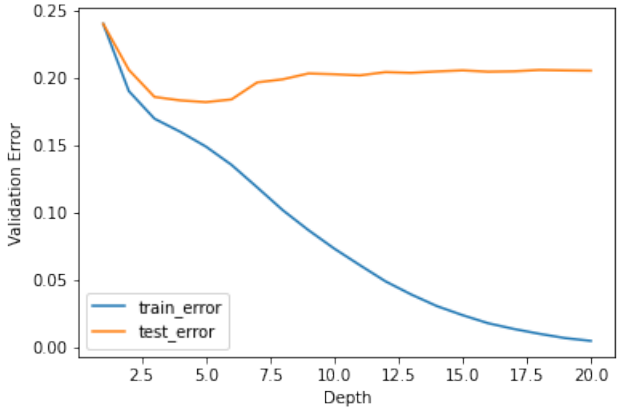


The best value of k is 15.

(g) **(5 pts)** One problem with decision trees is that they can *overfit* to training data, yielding complex classifiers that do not generalize well to new data. Let's see whether this is the case.

One way to prevent decision trees from overfitting is to limit their depth. Repeat your cross-validation experiments but for increasing depth limits, specifically, $1, 2, \ldots, 20$. Then plot the average training error and test error against the depth limit. Include this plot in your writeup, making sure to label all axes and include a legend for your classifiers. What is the best depth limit to use for this data? Do you see overfitting? Justify your answers using the plot.

Solution:

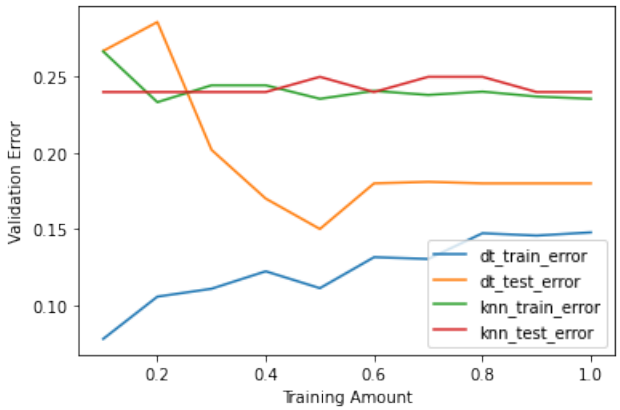


The best depth limit to use for this data is 5. Overfitting is obvious to see from this graph because as depth increases the gap between training error and test error increases. It tends to perform well on the training data but badly on the test data as depth increases due to overfitting.

(h) **(5 pts)** Another useful tool for evaluating classifiers is *learning curves*, which show how classifier performance (e.g. error) relates to experience (e.g. amount of training data). For this experiment, first generate a random 90/10 split of the training data and do the following experiments considering the 90% fraction as training and 10% for testing.

Run experiments for the decision tree and k-nearest neighbors classifier with the best depth limit and k value you found above. This time, vary the amount of training data by starting with splits of 0.10 (10% of the data from 90% fraction) and working up to full size 1.00 (100% of the data from 90% fraction) in increments of 0.10. Then plot the decision tree and k-nearest neighbors training and test error against the amount of training data. Include this plot in your writeup, and provide a 1-2 sentence description of your observations.

Solution:



As proportion of training data increases, decision-tree training errors increase and decision-tree test errors decrease. This is probably because more training data effectively prohibits overfitting. However, the KNN training and test errors only fluctuate by a little amount. Maybe increasing training data only has limited effects on improving KNN model with the best K given.

(i) **(5 pts)** Pre-process the data by standardizing it. See the sklearn.preprocessing.StandardScaler package for details. After performing the standardization such as normalization please run all previous steps part (b) to part (h) and report what difference you see in performance.
Solution: Answer in part(d) changes, training errors for k=3,5,7 decrease to 0.114, 0.129, 0.152.
Answer in part(e) changes, training error, test error for KNN with k=5 decrease to 0.133, 0.209. F1 score increases to 0.791.
The plot in part(f)(validation error against the number of neighbors, k) changes to a more symmetric shape, with the best value k changes from 15 to 27.
The plot in part(h) becomes more consistent with the same trend for decision-tree training and test error, but KNN training and test errors also decrease as training amount increases. Hence, pre-process the data by standardizing it seems to improve the behavior of the KNN model.