Fundamentals of Data Analysis and Decision Science
Duke University

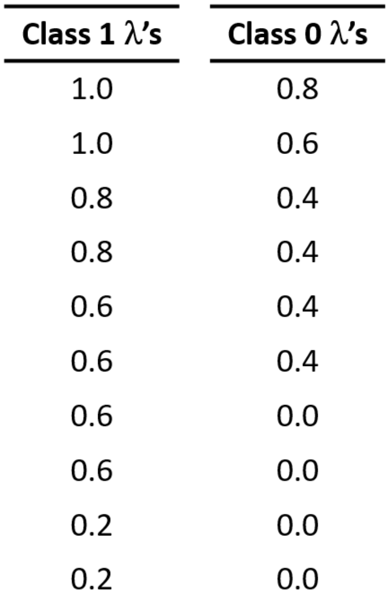
EGR/Math 238L
Spring 2021

K Nearest Neighbors and Binary Classification

Lab Assignment

By: Kenneth Marenco

1. A classifier produces the following decision statistics, λ, for a binary (two-class) classification problem. Recall that a decision statistic for a particular value of the feature x is the estimated probability $P(G = 1|X = x)$.



Class 1 λ's	Class 0 λ's
1.0	0.8
1.0	0.6
0.8	0.4
0.8	0.4
0.6	0.4
0.6	0.4
0.6	0.0
0.6	0.0
0.2	0.0
0.2	0.0

Find the confusion matrix when a decision threshold of 0.3 is applied, *clearly* indicating how the contents of each cell of the confusion matrix was found. Then draw the ROC curve, using all relevant decision thresholds.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	8 counts	2 counts
Predicted Negative (0)	6 counts	4 counts

By using the decision threshold of 0.3, any values above that in class 1 were deemed as a true positive with the others in the list being a false positive. The opposite is true in class 0 as values above 0.3 were false negatives with the others true negatives.

2. This question involves Figure 1 below. Suppose you have some training data with two-dimensional features X (the discs in the figure), and two possible class labels: $G = 1$ (filled-in discs) and $G = 0$ (striped discs). Imagine that you use a 3-NN method.

   What estimates would your method derive for the decision statistics $P(G = 1|X = A), P(G = 1|X = B), P(G = 1|X = C), P(G = 1|X = D)$, where A, B, C, D are as indicated in the figure?

P(G=1 | X=A) = 1/3 This is because 2 of the 3-NN are striped
P(G=1 | X=B) = 0 This is because all 3 of the NN are striped
P(G=1 | X=C) = 2/3 This is because 2 of the 3-NN are filled-in
P(G=1 | X=D) = 2/3 This is because 2 of the 3-NN are filled-in

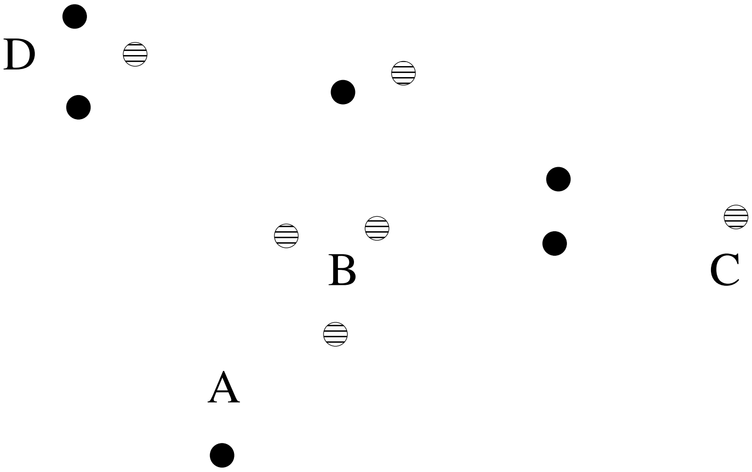


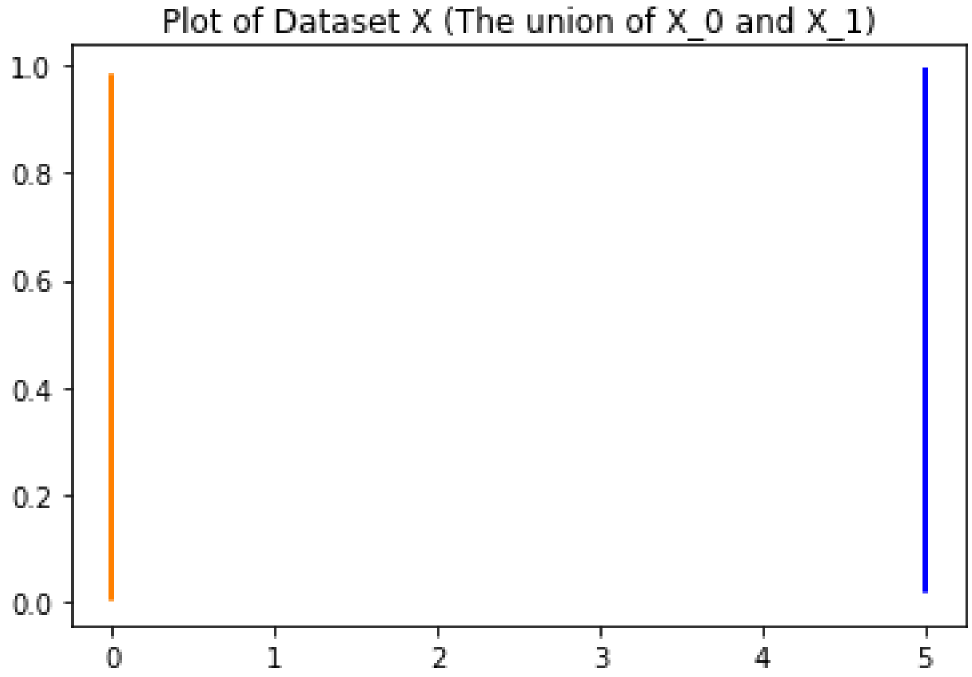
Figure 1: Twelve training points in a two-dimensional feature space, along with four test points.

3. Create a two-D dataset X which is the union of two subsets X_0 and X_1, $X = X_0 \cup X_1$.

   Let X_0 consist of 100 points, all of which have 0 as the first coordinate, and the second coordinate is drawn from a uniform distribution from 0 to 1. (`rand` generates uniform random numbers between 0 and 1.)

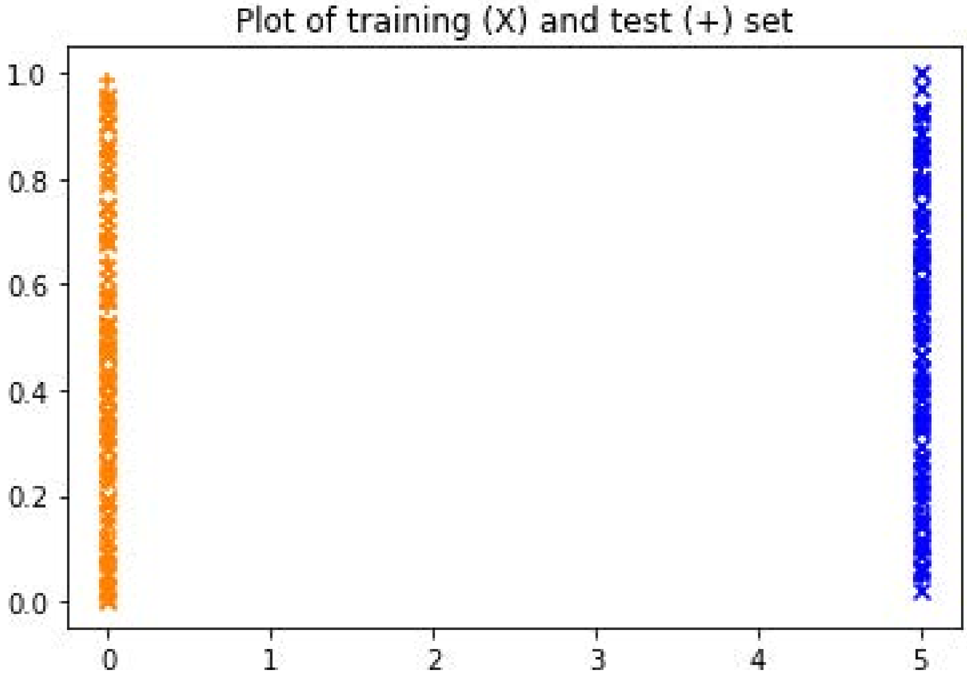
   Let X_1 consist of 100 points, all of which have 5 as the first coordinate, and the second coordinate is drawn from a uniform distribution between 0 and 1.

   (a) Plot the dataset X with X_0 in orange and X_1 in blue.[1]
   (The RGB triple for orange is $\{255\ 127\ 0\}$ or $\{1\ 0.5\ 0\}$.)



[1] It is good practice to attempt choose colors which may be more distinguishable by individuals who are colorblind.

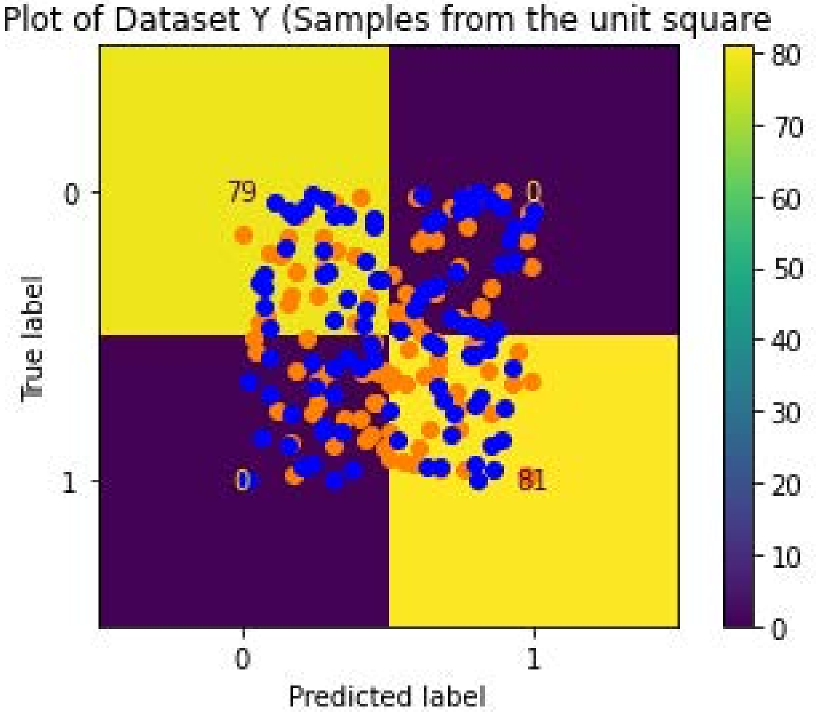
(b) Thinking of X as your wizard data, divide this wizard data into 5 disjoint folds, making sure to keep colors balanced. Consider the first four folds to be training data, and the last fold to be test data. Visualize this by using distinct symbols for training and for test data.



(c) Predict how well a 1-NN classifier will do using this training and this test set, and explain why the performance is what it is.

I believe the classifier will do decently well since the two clusters that it should be finding are those with the same x value. The largest changes between the second coordinate only varies between 0 and 1 meaning the 1st coordinate should lead to easy clustering.

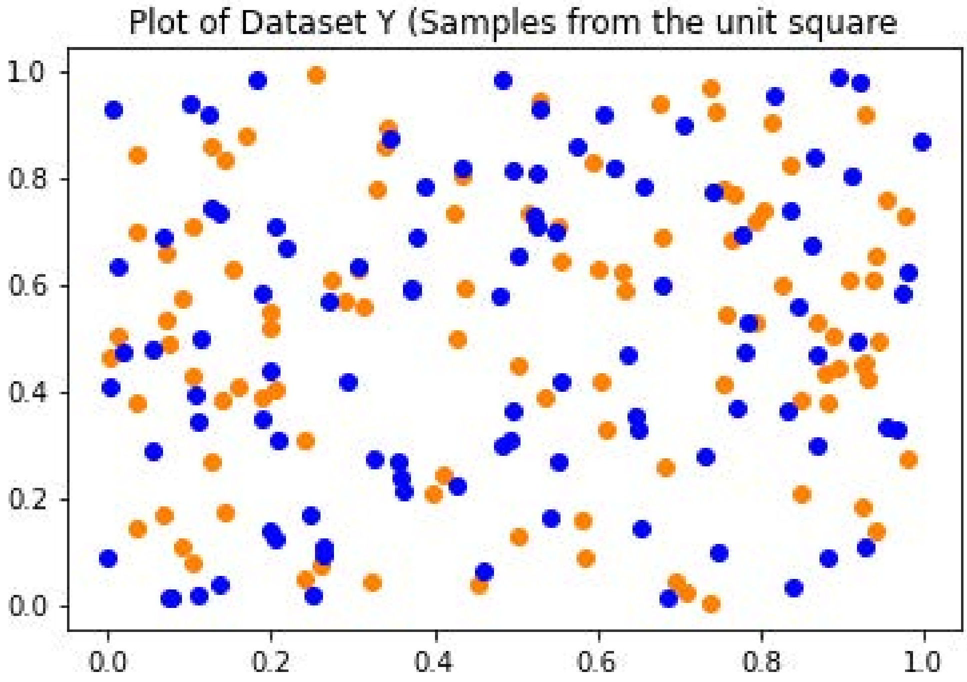
(d) Verify your prediction using software, and output your confusion matrix.



(e) Now do cross-validation by repeating the above procedure five times, each time using a different fold for the test set. Output statistical summaries of this process (e.g, for each cell in the confusion matrix, output the mean and the standard deviation over the five folds.)

4. Create a two-D dataset Y consisting of 200 points drawn uniformly at random from the unit square in the plane. Randomly assign half of the points to the subset Y_0 and the other half of the points to the subset Y_1.

   (a) Plot the dataset Y with Y_0 in orange and Y_1 in blue.[2]
   (The RGB triple for orange is $\{255\ 127\ 0\}$ or $\{1\ 0.5\ 0\}$.)



[2] It is good practice to attempt choose colors which may be more distinguishable by individuals who are colorblind.

(b) Repeat all steps from the previous problem, but now on this new dataset

For this dataset, the model has difficulties finding clear clusters because the seemingly overlap each other without cause.

5. This exercise concerns the 'Auto Data'. The goal is to determine which combinations of numerical variables (displacement, horsepower, weight, acceleration) best predict whether the car is 'old' or 'new'. For the purposes of this exercise, 'old' will mean model year 1974 or earlier, and 'new' will mean model year 1975 or later. [3]

   (a) Consider the variables displacement and horsepower.

   Create a 2D array of these two variables for all the cars (*e.g.* N rows and 2 columns). This array contains the features.

   Also create a logical column vector indicating if each car is old or new (*e.g.* old == 0 and new == 1). This vector contains the target variables; the n^{th} element of the target variable vector contains the target variable corresponding to the n^{th} set of features (the n^{th} row in the feature array).

   (b) Plot these data in the plane, using color to indicate new and old.[4]

[3] The first Clean Air Act was passed in 1970, and required a 90% reduction in emissions from new vehicles by 1975. `https://www.epa.gov/transportation-air-pollution-and-climate-change/timeline-major-accomplishments-transportation-air`

[4] It is good practice to attempt choose colors which may be more distinguishable by individuals who are colorblind.

(c) Run 1-NN with 5-fold cross-validation, and provide the resulting confusion matrix.

(d) Run 7-NN with 5-fold cross-validation, and provide the ROC curve along with the AUC (write your own function to find the AUC for an ROC). Interpret it as best you can.

(e) For all other combinations of possible predictors (singletons, pairs, triplets, all four),[5] run 7-NN with 5-fold cross-validation, and provide ROC curves along with the AUC, and discuss which combinations of predictors (features) seem to provide better/worse performance. If you can, try to justify why this would be so.

[5] There are 15 different features sets to consider: 4 different combinations of features taken one at a time, 6 different combinations of features taken two at a time, 4 different combinations of features taken three at a time, and 1 possible way to take all four features.