Jonathan Kaneshiro
861195520
CS 171 Assignment 1
Late days: 0
Total Late days used: 0
4/18/2018

Approaches taken are in commented code.

Question 1

3) [20%] For each distribution (using the histograms), under the figure you produce in your report also write whether it is 1) {mostly symmetric or mostly skewed}, and 2 {mostly unimodal, mostly bimodal/multimodal, mostly uniform}.

I assumed some graphs are mostly bimodal/multimodal because there are huge gaps between peaks especially when the bin sizes are large.

Iris-Setosa

- Sepal Length
  - Bin 5 Mostly unimodal Mostly symmetric
  - Bin 10 Mostly skewed Mostly bimodal/multimodal
  - Bin 50 Mostly bimodal/multimodal Mostly symmetric
  - Bin 100 Mostly bimodal/multimodal Mostly symmetric
- Sepal Width
  - Bin 5 Mostly symmetric Mostly unimodal
  - Bin 10 Mostly skewed Mostly bimodal/multimodal
  - Bin 50 Mostly skewed Mostly bimodal/multimodal
  - Bin 100 Mostly skewed Mostly bimodal/multimodal
- Petal Length
  - Bin 5 Mostly unimodal Mostly symmetric
  - Bin 10 Mostly symmetric Mostly unimodal
  - Bin 50 Mostly bimodal/multimodal Mostly symmetric
  - Bin 100 Mostly bimodal/multimodal Mostly symmetric
- Petal Width
  - Bin 5 Mostly skewed Mostly unimodal
  - Bin 10 Mostly skewed Mostly unimodal
  - Bin 50 Mostly skewed Mostly unimodal
  - Bin 100 Mostly unimodal Mostly skewed

Iris-Versicolor

- Sepal Length
  - Bin 5 Mostly skewed Mostly unimodal
  - Bin 10 Mostly skewed Mostly bimodal/multimodal
  - Bin 50 Mostly skewed Mostly bimodal/multimodal
  - Bin 100 Mostly skewed Mostly bimodal/multimodal
- Sepal Width
  - Bin 5 Mostly skewed Mostly unimodal
  - Bin 10 Mostly skewed Mostly bimodal/multimodal
  - Bin 50 Mostly skewed Mostly unimodal
  - Bin 100 Mostly skewed Mostly unimodal
- Petal Length
  - Bin 5 Mostly skewed Mostly unimodal
  - Bin 10 Mostly skewed Mostly bimodal/multimodal

Bin 50 Mostly skewed Mostly bimodal/multimodal
Bin 100 Mostly skewed Mostly bimodal/multimodal
Petal Width
Bin 5 Mostly skewed Mostly bimodal/multimodal
Bin 10 Mostly skewed Mostly bimodal/multimodal
Bin 50 Mostly skewed Mostly bimodal/multimodal
Bin 100 Mostly skewed Mostly bimodal/multimodal

Iris-Virginica
Sepal Length
Bin 5 Mostly symmetric Mostly unimodal
Bin 10 Mostly skewed Mostly bimodal/multimodal
Bin 50 Mostly skewed Mostly bimodal/multimodal
Bin 100 Mostly skewed Mostly bimodal/multimodal
Sepal Width
Bin 5 Mostly symmetric Mostly unimodal
Bin 10 Mostly skewed Mostly bimodal/multimodal
Bin 50 Mostly symmetric Mostly bimodal/multimodal
Bin 100 Mostly symmetric Mostly bimodal/multimodal
Petal Length
Bin 5 Mostly skewed Mostly unimodal
Bin 10 Mostly skewed Mostly bimodal/multimodal
Bin 50 Mostly skewed Mostly bimodal/multimodal
Bin 100 Mostly skewed Mostly bimodal/multimodal
Petal Width
Bin 5 Mostly skewed Mostly bimodal/multimodal
Bin 10 Mostly skewed Mostly bimodal/multimodal
Bin 50 Mostly skewed Mostly bimodal/multimodal
Bin 100 Mostly skewed Mostly bimodal/multimodal

Question 2

1c) What is the absolute minimum number of calls to the correlation(x,y) function in order to fill in this matrix?

The absolute minimum number of calls to fill in this matrix is half of the intended values of the matrix because since the matrix is symmetric, having the upper or lower matrix is sufficient enough to have all the calls necessary to fill in the matrix without losing data. The diagonal is also given as it will always produce 1.

1d) Do you observe any correlated features? How can this information be useful.

Yes, the correlated features are numbers that are close to 1. In the Iris dataset, we have sepal length correlated with sepal length, sepal width correlated with sepal width, petal length correlated with petal length, petal width correlated with petal width, petal length correlated with sepal length, petal width correlated with sepal length, and petal length correlated with petal width. In the wine dataset, alcohol is correlated with alcohol, malic acid is correlated with malic acid, and ash is correlated with ash. This information can be useful because the data collected can be cut in half by only keeping the upper or lower triangular matrix knowing that there are duplicate answers. The diagonal will always be 1 because an attribute is being compared to itself.

2b) Do you observe any pairs of features being discriminative? By “discriminative” we mean pairs of features that show good separation of the two classes in the 2D space defined by those features.

Iris setosa is discriminative because it does not overlap among the other two classes in almost all scatterplots except for sepal length comparing with sepal length and sepal width comparing with sepal width.

2c) Do you observe any pairs of features being non-discriminative? To what extent does this agree with the set of correlated features from #1?

Iris Virginica and Iris Versicolor are pairs being non-discriminative because they overlap in all scatterplots. It agrees with the set of correlated features because their correlation matrix and their representation in the scatterplots closely match with their corresponding correlation coefficients in the heat map. There is symmetry in the scatterplots and in the heat maps.

3c) What is the absolute minimum number of calls to the distance function you need to do to fill in this matrix?

The absolute minimum number of calls to fill in this matrix is half of the intended values of the matrix because since the matrix is symmetric, having the upper or lower matrix is sufficient enough to have all the calls necessary to fill in the matrix without losing data. The diagonal is also given as it will always produce 0 since the distance from one attribute to itself is always 0.

3e) For each data point, find its non-trivial nearest data point (e.g., the point that is not the same point). What is the label of the nearest point? Is it the same? Does the answer change for different values of p?

In iris, versicolor has 2 flowers that have the same minimum distance. The answer does not change for different values of p because as the labels stay the same, the distances get scaled the same amount so the answers are the same.

Also I don’t know how to save a .mat file into a .png file so if you run the nearest point code while commenting out everything, it will create a dist_iris.mat file where I get this information.

In wine, there are no same minimum distances. Same situation as above, however my file is dist_wine.mat.