Engineering and Applied Science Programs for Professionals
Whiting School of Engineering
Johns Hopkins University
685.621 Algorithms for Data Science
Homework 3
Assigned at the start of Module 5
Due at the end of Module 6

Total Points 100/100

Collaboration groups have been set up in Blackboard. Make sure your group starts an individual thread for each collaborative problem and subproblem. You are required to participate in each of the collaborative problem and subproblem. Do not directly post a complete solution, the goal is for the group develop a solution after everyone has participated.

Problems for Grading

1. **Problem 1**
   20 Points Total

   In this problem, develop code to analyze the Iris data sets using the test statistics listed in Table 1.

Table 1: Data Analysis Statistics

Test Statistics	**Statistical Function $\boldsymbol{F}(\cdot)$**
Minimum	$F_{\min}(\mathbf{x}) = \min(\mathbf{x}) = x_{min}$
Maximum	$F_{\max}(\mathbf{x}) = \max(\mathbf{x}) = x_{max}$
Mean	$F_{\mu}(\mathbf{x}) = \mu(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
Trimmed Mean	$F_{\mu_t}(\mathbf{x}) = \mu_t(\mathbf{x}) = \frac{1}{n-2p}\sum_{i=p+1}^{n-p} x_i$
Standard Deviation	$F_{\sigma}(\mathbf{x}) = \sigma(\mathbf{x}) = \left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^2\right)^{1/2}$
Skewness	$F_{\gamma}(\mathbf{x}) = \gamma(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^3}{\sigma(\mathbf{x})^3}$
Kurtosis	$F_{\kappa}(\mathbf{x}) = \kappa(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^4}{\sigma(\mathbf{x})^4}$

   The analysis should be done by feature followed by class of flower type. This analysis should provide insight into the Iris data set.

   Note: The trimmed mean is a variation of the mean which is calculated by removing values from the beginning and end of a sorted set of data. The average is then taken using the remaining values. This allows any potential outliers to be removed when calculating the statistics of the data. Assuming the data in $\mathbf{x}_s = [x_{1,s}, x_{2,s}, \cdots, x_{n,s}]$ is sorted, the resulting $\mathbf{x}_{s,p} = [x_{1+p,s}, x_{2+p,s}, \cdots, x_{n-p,s}]$. the trimmed mean allows the removal of extreme values influencing the mean of the data.

2. **Problem 2 Parts a and b**
30 Points Total 15 Points Each

In this problem we will begin to analyze Iris data based on the class of flower type using linear discriminant analysis.

(a) Implement the two class linear discriminant based on the Fisher's Linear Discriminant (FLD) two-class separability (Fisher, 1936) described below. This is also shown in the two class linear discriminant function presented in (Bishop, 2006) Section 4.1.1 Two classes. For this exercise you will want to separate your Iris data into three sets and focus on any two class combination. For example, from the iris data take the first 50 observations for class 1, the next 50 as class 2 and the final 50 as class 3. Using the two class linear discriminant function compare class 1 verses class 2, class 1 verses class 3 and finally compare class 2 versus class 3.

(b) For this problem you will want to expand the two class case from part a to a three class case as presented in (Bishop, 2006) from Section 4.1.2 Multiple classes.

Now that we have our statistic set up let look a the mean and standard deviation between the classes (Iris flower types) and within the classes let's consider the Fisher's Linear Discriminant (FLD) to quantify two-class separability of features (Fisher, 1936). FLD is a simple technique which measures the discrimination of sets of real numbers. Without going into all of the theory of the FLD lets focus on the primary components assuming we have a two class problem, equal class sample and a covariance matrix that is generated from normal distributions. The within-class scatter matrix is defined as

$$S_W = \sum_C P_C S_C \tag{1}$$

where S_C is the covariance matrix for class $\mathbf{C} \in \{-1, +1\}$

$$S_C = \sum_{\substack{i=1,\\ i \in C}}^{l_C} (\mathbf{x}_i - \mu_C)(\mathbf{x}_i - \mu_C)^T \tag{2}$$

and P_C is the *a priori* probability class $\mathbf{C}$. That is, $P_C \approx k_C/k$, where k_C is the number of samples in class $\mathbf{C}$, out of a total of k samples. The between-class smatter matrix is defined as

$$S_B = \sum_{\mathbf{C}} (\mu_{-1} - \mu_{+1})(\mu_{-1} - \mu_{+1})^T \tag{3}$$

where μ is the global mean vector

$$\mu = \frac{1}{l} \sum_{i=1}^{l} \mathbf{x}_i \tag{4}$$

and the class mean vector μ_C is defined as

$$\mu_C = \frac{1}{l_C} \sum_{\substack{i=1,\\ i \in C}}^{l_C} \mathbf{x}_i \tag{5}$$

Now lets look at the criterion function $J(\cdot)$ written as follows:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{6}$$

where $\mathbf{w}$ is calculated to optimize $J(\cdot)$ as follows:

$$\mathbf{w} = S_W^{-1}(\mu_{-1} - \mu_{+1}) \tag{7}$$

w for the Fisher Linear Discriminant has been obtained, which will allow for the linear function to yield the maximum ratio between of the between-class scatter and the within-class scatter. Now let's determine a threshold b that will allow us to determine which class a new observation will belong to. The optima decision boundary assuming each class has the same number of samples can be calculated as follows:

$$b = -0.5(\mathbf{w}\mu_{-1} + \mathbf{w}\mu_{+1}) \tag{8}$$

Now, if we have a new input observation x we can determine which class the new observation belongs to based on the following

$$y = \mathbf{w}\mathbf{x} + b \tag{9}$$

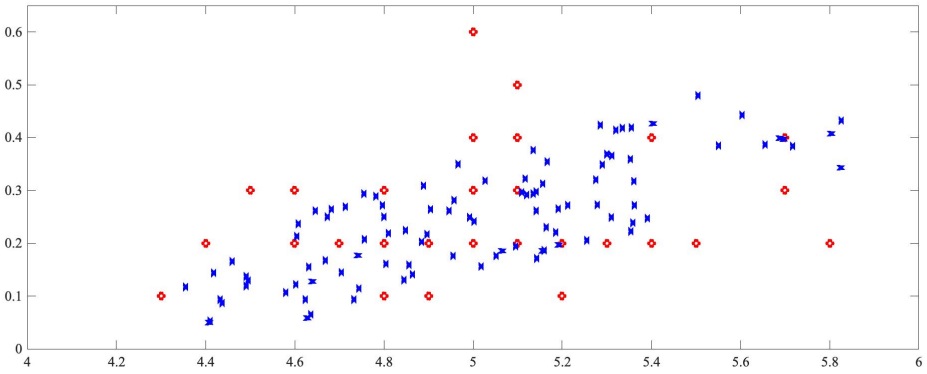
where $y < 0$ is class -1 and $y \geq 0$ is class $+1$.

The previous discussion is based on the FLD and is simplified as a two class linear discriminant function presented in (Bishop, 2006) Section 4.1.1 Two classes. Credit is given to Fisher for his work in this area of linear discrimination.

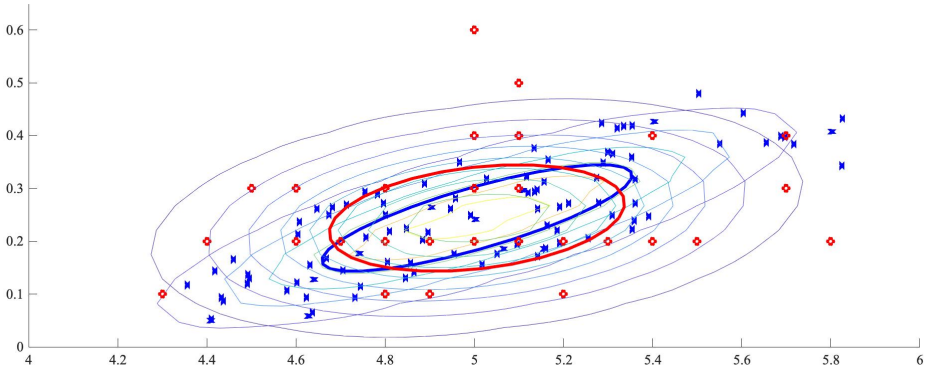
3. **Problem 3** ***Note this is a Collaborative Problem***
25 Points Total

In this problem the Iris data set is to be expanded with synthetic data so that 100 additional observations are generated for each flower class resulting in 300 additional observations. Once the data is generated make similar figure as provided in Figure 1 (a) for each set of paired features and classes.

So let's take the first 50 observations, the first feature (sepal length) and fourth feature (petal width) shown in red as observed in Figure 1. The 100 additional observations generated are show in blue. In this example the data has similar covariance matrix, mean, minimum and maximum. The synthetic data was generated using the covariance matrix, mean, minimum and maximum of the data. Random data was generated that contained 100 observations and 4 features. The random data was multiplied by the covariance matrix, normalized to fit the original Iris data in terms of minimum and maximum values then the mean of the data was set based on the Iris mean.



(a) Synthetic Data (blue) vs Iris Data (red)



(b) Distributions

Figure 1: Synthetic Data vs Iris Data (a) shows the synthetic data in blue and the original Iris in red, (b) the distributions of the data are shown for context.

4. **Problem 4** ***Note this is a Collaborative Problem***
25 Points Total

In some application areas of data science, data retrieval and data cleansing are critical to the entire analysis process. One example is portfolio analysis. Elsevier's Scopus (https://www2-scopus-com.proxy1.library.jhu.edu/search/form.uri?display=basic) is the largest abstract and citation database of peer reviewed literature: scientific journals, books and conference proceedings. It covers nearly 36,377 titles from approximately 11,678 publishers, of which 34,346 are peer-reviewed journals in top-level subject fields: life sciences, social sciences, physical sciences and health sciences.

(a) Go to the Scopus website and search for data science and machine learning related documents. Plot the distribution of the number of documents by year from at least the last 10 years. What is the story that the plot tells you?

(b) Limit the search to 2016 and 2017. List the possible data fields/columns you may need to export in order to answer the question of author and/or institution collaborations in this scientific area during this timeframe.

(c) Within the possible fields you suggest to export, which fields need data cleansing and why, in order to provide robust input for performing portfolio analysis?

References

[1] Bishop, Christopher M., *Neural Networks for pattern Recognition*, Oxford University Press, 1995

[2] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006, https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[3] Bruce, Peter and Bruce, Andrew, *Practical Statistics for Data Science*, O'Reilly, 2017

[4] Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronal L., and Stein, Clifford, *Introduction to Algorithms*, 3rd Edition, MIT Press, 2009

[5] Duin, Robert P.W., Tax, David and Pekalska, Elzbieta, *PRTools*, http://prtools.tudelft.nl/

[6] Fisher, R. A., *The use of Multiple Measurements in Taxonomic Problems*, Proceedings of Annals of Eugenics, Number 7, pp. 179-188, 1936

[7] Franc, Vojtech and Hlavac, Vaclav, *Statistical Pattern Recognition Toolbox*, https://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

[8] Fukunaga, Keinosuke, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972

[9] Machine Learning at Waikato University, *WEKA*, https://www.cs.waikato.ac.nz/ ml/index.html

[10] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Jan 31, 1986

[11] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, Cambridge University Press, September 10, 2007

[12] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, http://numerical.recipes/

[13] Press, William H., *Opinionated Lessons in Statistics*, http://www.opinionatedlessons.org/