ISTE-470 Assignment 2

##### Fall 2221

Answer the following questions as a team and submit them to the Assignment 2 dropbox by the specified due date. Keep in mind that even though this assignment will be submitted as a team, each team member is individually responsible for understanding the answers to each question and can independently write code and perform necessary analyses on an exam.

Scoring: Questions 1-5 are worth 10 points each; Question 6 is worth 50 points.

1. Why are variance and standard deviation sensitive to outliers? **Hint: think of their relationship to the mean**.

Because to calculate variance and standard deviation you must use the mean. When calculating the mean, it will always move towards the outlier. So, for this reason outliers heavily affect variance and standard deviation.

2. What is the mean, median, and IQR of the data set “11 3 1 6 7 5 4 5”? (Show your work to get full credit)

Mean =

Mean = 5.25

Median =

Median=5

Q1 = Q2 =

Q1 = 3.5 Q2 = 6.5

IQR = Q2 – Q1 = 6.5 – 3.5 = 3

3. Why is it important to provide axis labels when visualizing data?

Axis labels are important in interpreting visual data. The y and x axes show what the visual data is representing, which allows the viewer to properly understand what is being portrayed through the graph. It is important to ensure that the axis ticks are the same between two different graphs when comparing the two graphs. For example, if one graph were to have a y-axis that reaches 120 grams while another graph that is used for comparison reaches 40 grams, it can be easy to misinterpret the graphs and come to an incorrect conclusion. The two graphs can look similar as they are not to the same scale. Furthermore, providing units and axis labels are essential in allowing the interpreter correctly understand what is being portrayed.

4. We saw that no single attribute, or pair of attributes, allow us to visually separate the three classes of irises. What could we ask a domain expert to do to help us with this problem?

5. Open iris.arff in Weka. For each attribute in the data set, note if the attribute is continuous or discrete. If the attribute is continuous, list its min, max, mean, and standard deviation. If the attribute is discrete, list each attribute value, its frequency, and if there is a single mode among the attribute values.

Sepal Length (Continuous): Min: 4.3, Max: 7.9, Mean: 5.843, Standard Deviation: 0.828

Sepal Width (Continuous): Min: 2, Max: 4.4, Mean: 3.054, Standard Deviation: 0.434

Petal length (Continuous): Min: 1, Max, 6.9, Mean: 3.759, Standard Deviation: 1.764

Petal width (Continuous): Min: 0.1, Max: 2.5, Mean: 1.199, Standard Deviation: 0.763

Class (Discrete):

* Iris Setosa: Frequency: 0.333, No single mode
* Iris Versicolor: Frequency: 0.33, No single mode
* Iris Virginica: Frequency: 0.333, No single mode

6a. (10 points)Using your three selected features from Exercise 2 (**you should be selecting the three features with the largest range**), create three separate scatter plots (one for each combination of features). For example, if you have features F1, F2, and F3, you would create the following scatter plots: F1 vs. F2, F1 vs. F3, and F2 vs. F3. Insert the three plots below and provide labels for each plot.

F1 (PLA) vs F2 (PRA)

Chart, scatter chart

Description automatically generated

F1 (PLA) vs F3 (PRV)

Chart, scatter chart

Description automatically generated

F2 (PRA) vs F3 (PRV)

Chart, scatter chart

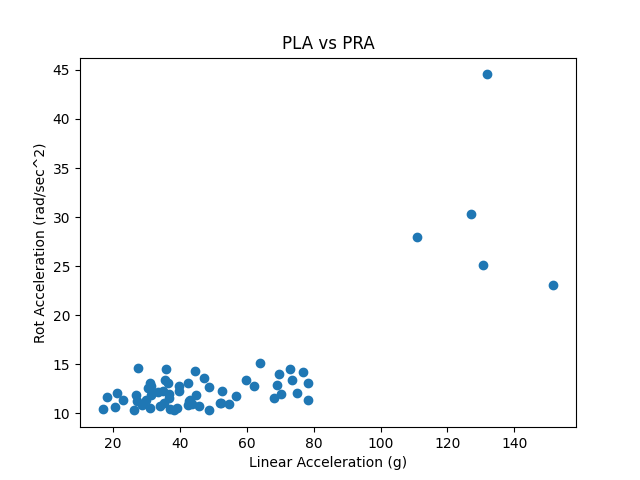
Description automatically generated

6b. (10 points) Find the “Top 5” largest feature values for each of the features and note which data instance they belong to. For example, if the largest value for feature F1 is at index 2, then that would correspond to Instance 3. Fill out the table below by listing the instance numbers in sorted ascending order.

|  |  |
| --- | --- |
| **Feature** | **Instance Numbers of Top 5 largest feature values** |
| PLA | 8, 9, 11, 19, 20 |
| PRA | 8 ,9 ,11, 19, 20 |
| PRV | 8, 9, 11, 34, 61 |

6c. (15 points) Two of the features should have the exact same instance numbers in the table above, which means that these are the five “different” data instances. Based on what you learned from our discussion of visual separability in lecture, describe how the fact that two features have the same five “different” data instances manifests itself in the scatter plot for those two features, as compared with the other two feature scatter plots. You should insert the scatter plot images in your answer to assist in your discussion.

Because two of the graphs happen in the same instance and their values are obviously higher than the rest, like what we see that in graph 1 (PLA vs PRA). We can see that in other graphs points tend to not be as isolated because their highest values don’t happen in the same instance.



6d. (15 points) Insert a waveform plot of one the “different” data instances and a waveform plot of one of the other 60 data instances. Provide a visual comparison of the two sets of waveforms. Make sure you refer to the three features you selected as part of your discussion.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

We can see how on the image of instance 19, we can see that the peak values of Linear Acceleration and Rotational Acceleration happen at the same time, and we can see that the peaks from rotational velocity happen at different times. When Rotational Acceleration and Linear Acceleration have their peaks, we can see that Rotational Velocity has a drop. In the instance 59, in contrast we can see that all the peaks are happening at different times and they are not related at all, but what stays true is that rotational acceleration and linear acceleration both go up whenever Rotational Velocity has a drop.