assignment-1

January 29, 2018

1 Homework 1

1.0.1 Problem 1 (5 points):

Answer each of the following questions with a few sentences:

a. What is the difference between classification and clustering? Classification is a type of problem in which a set of predefined classes are known and the goal is to find the class that new object belongs to. In classification the labels are known.

Clustering is a technique used to attmpt to group sets of objects and infer relationships based on patterns that an algorithm learns. In clustering the labels are not known.

b. What is the difference between data warehouse and database? Data warehouses and databases are different in their functionalities. Data warehouses are used for data modeling and are optimized for reading operations. Databases are used for relational modeling and are optimized for writing operations. They also differ in that data warehouses have de-normalized data while databases have normalized data. Databases are used for OLTP and data warehouses are used for OLAP.

c. What is the difference between data mining and OLAP? Data mining and OLAP are very similar to each other. Perhaps their biggest difference is the way they operate on data. Data mining has a lower level of detail which allows for more detailed patterns to be seen in data. OLAP aggregates data on high levels of detail allowing for more summary calculations.

d. What is the difference between data marts and data warehouse? The difference between a data mart and a data warehouse is the level of data they hold. A data mart holds a very concetrated subset of data (often data from a data warehouse) whereas a data warehouse contains multiple subsets of data in the same location. Data marts are typically more aggregated and summarized, while data warehouses are very detailed.

e. In a data table, what do the columns represent and what do the rows represent? A row in a given table represents an individual record of whatever the table describes (i.e. Baseball player first name, last name, and team). A column in a table represents an attribute of the table, and more specifically any given row (i.e. The baseball player's first name).

1.0.2 Problem 2 (5 points):

Answer each of the following questions with a few sentences:

a. (2 points) After loading new data into SPSS, describe two tasks you might do to clean your data. The first thing I would do is drop any fields that are unrelated to the problem that I am trying to solve. After that I would look for and either remove or fix all data that is incorrectly formatted, has duplications, is NaN or NA, etc.

b. Explain which type of data mining algorithm (also called data mining functionality) would you use to answer each of these questions? i. What are five groups of customers who buy similar things?
This would be descriptive, and might use a clustering algorithm to find patterns among customer purchases
ii. What are different sets of products that are often purchased together?
Again I would use a clustering algorithm to look at total customer purchases and see if there are any patterns among purchase combinations
iii. I sell milk
Not sure about this one.. haha I would say predictive if we were trying to predict our milk sales?

1.0.3 Problem 3 (5 points):

Explain whether or not each of the following activities is a data mining task.

a. Dividing the customers of a company according to their gender. This would not be data mining since it's simply just aggregating a field by an attribute.

b. Computing the total sales of a company. This would not be data mining since it is a summary and not at a detailed level.

c. Sorting a student database based on student identification numbers. This would not be data mining since it's a database query.

d. Predicting the outcomes of tossing a (fair) pair of dice. This would be data mining since it would require a predictive model.

e. Predicting the future stock price of a company using historical records. This would be data mining since is would also require predictive modeling.

1.0.4 Problem 4 (15 points):

a. Visualize the relationship between the two sepal variables, sepal length and sepal width using a scatter plot. Use different colors or symbols per class so you can see how the classes are related to this pair of variables. We talked in class about how classifiers work broadly-speaking. Do you think that a classification algorithm using these two variables will be successful in classifying data with respect to the class labels we have? Explain why or why not and include the plot image with your answer.

b. Repeat part (a) for the petal variables.

c. Create a histogram for each of the four variables. Histograms in SPSS are just a different graph type from scatterplots. Describe what you can tell about the distribution of each variable.

d. Determine if there are any outliers in the data with respect to the sepal length.

e. Repeat d. for the petal length.

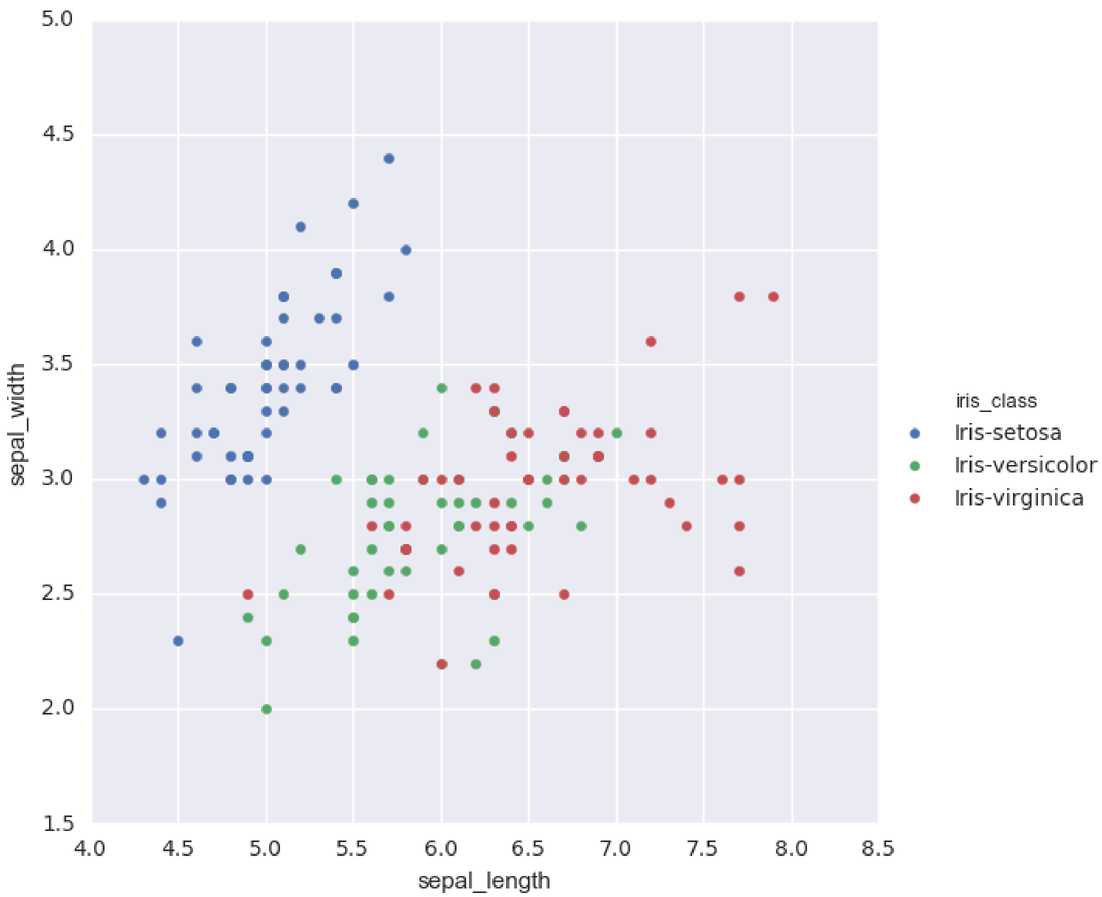
```
In [39]: #Import packages
         import pandas as pd
         from matplotlib import pyplot as plt
         import seaborn as sns
         %matplotlib inline
```

```
In [11]: #Import data
         url = "https://github.com/PixarJunkie/fundamentals-of-data-science/raw/mas
         df = pd.read_csv(url)
         df.head()
```

```
Out[11]:    sepal_length  sepal_width  petal_length  petal_width   iris_class
         0           5.1          3.5           1.4          0.2  Iris-setosa
         1           4.9          3.0           1.4          0.2  Iris-setosa
         2           4.7          3.2           1.3          0.2  Iris-setosa
         3           4.6          3.1           1.5          0.2  Iris-setosa
         4           5.0          3.6           1.4          0.2  Iris-setosa
```

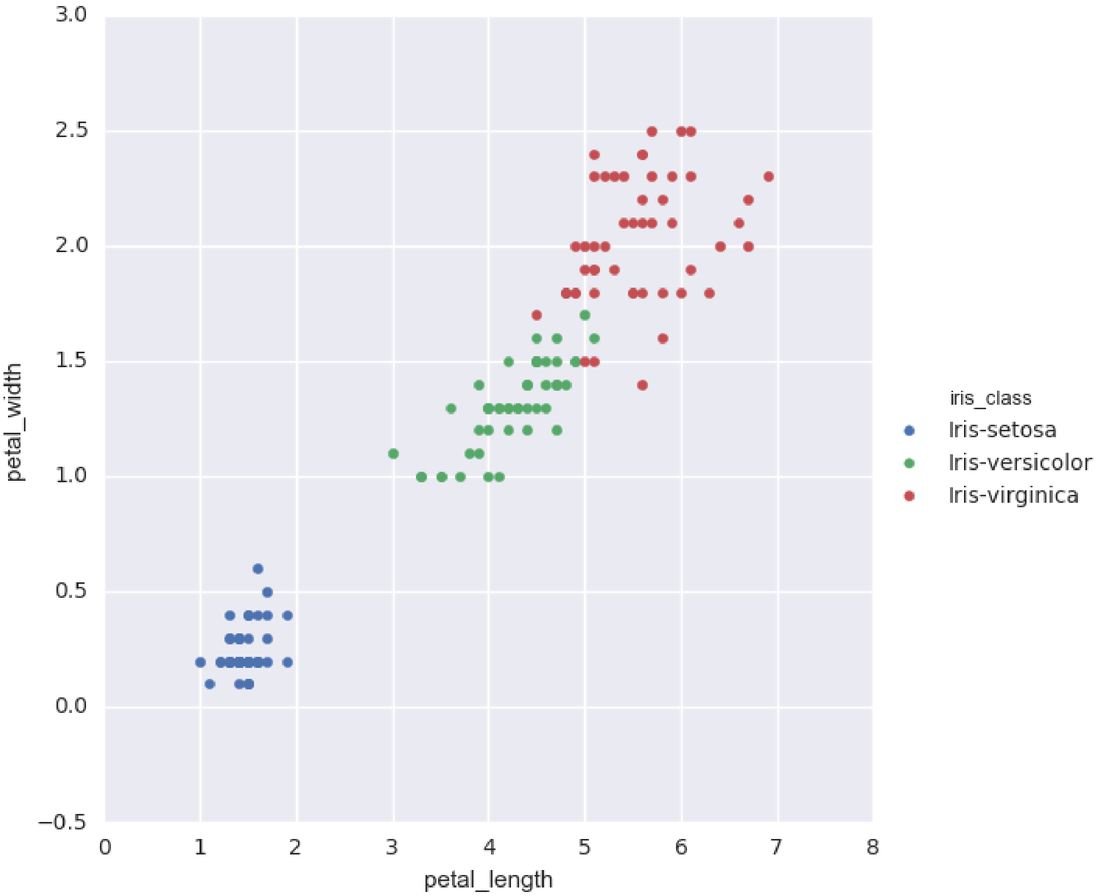
Problem 4-a I think a classification algorithm using just these two variables would do decently. It would have a hard time classifying the difference between the Iris-versicolor and the Iris-virginica since there is a lot of overlap. From the scatterplot, it looks like a classification algorithm would classify the Iris-setosa quite well.

```
In [44]: #sepal_length vs. sepal_width plot
         sns.FacetGrid(df, hue = 'iris_class', size = 6).map(plt.scatter, "sepal_le
         sns.plt.show()
```



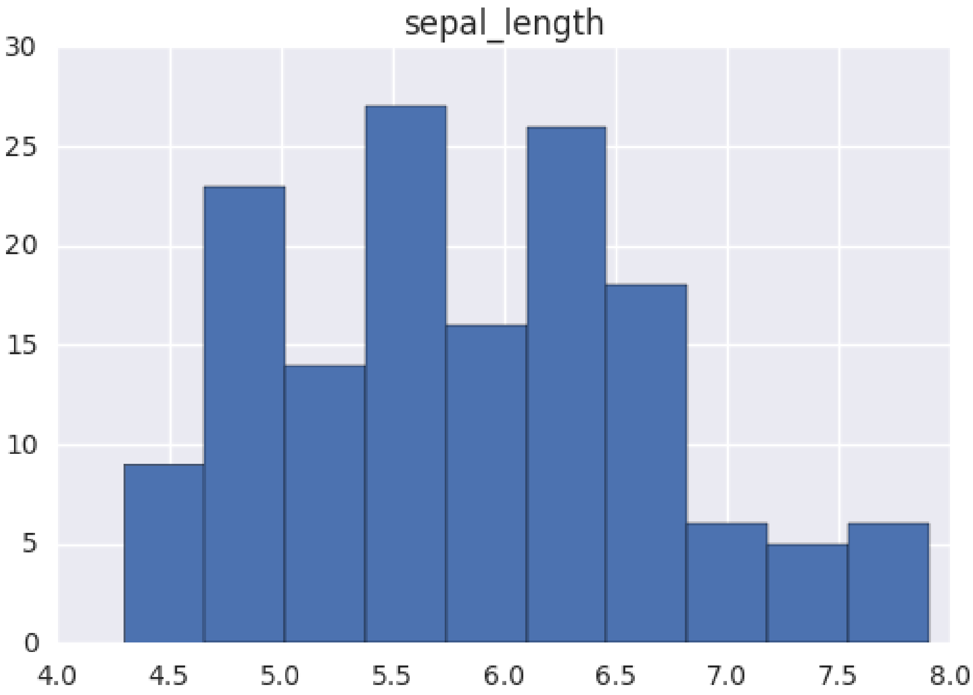
Problem 4-b A classification algorithm using just these two variables would perform much better than the sepal variables. It would have the same problem with mis-classifying between the Iris-versicolor and the Iris-virginica in some ranges. Overall though, there are quite noticable differences between the three Iris classes and their respective pedal lengths/widths.

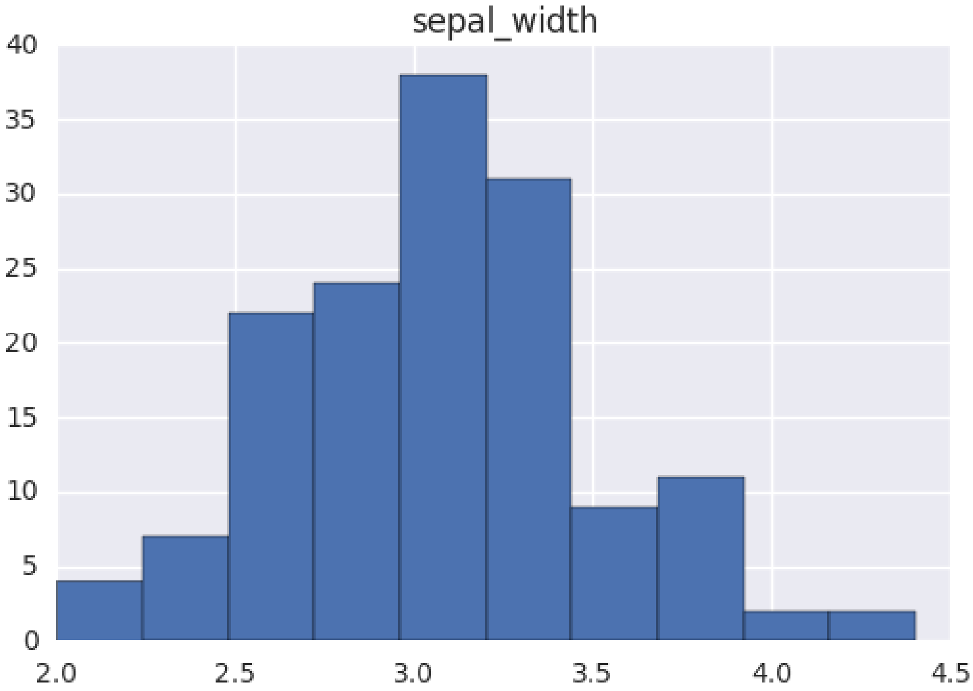
```
In [46]: #petal_length vs. petal_width plot
         sns.FacetGrid(df, hue = 'iris_class', size = 6).map(plt.scatter, "petal_le
         sns.plt.show()
```

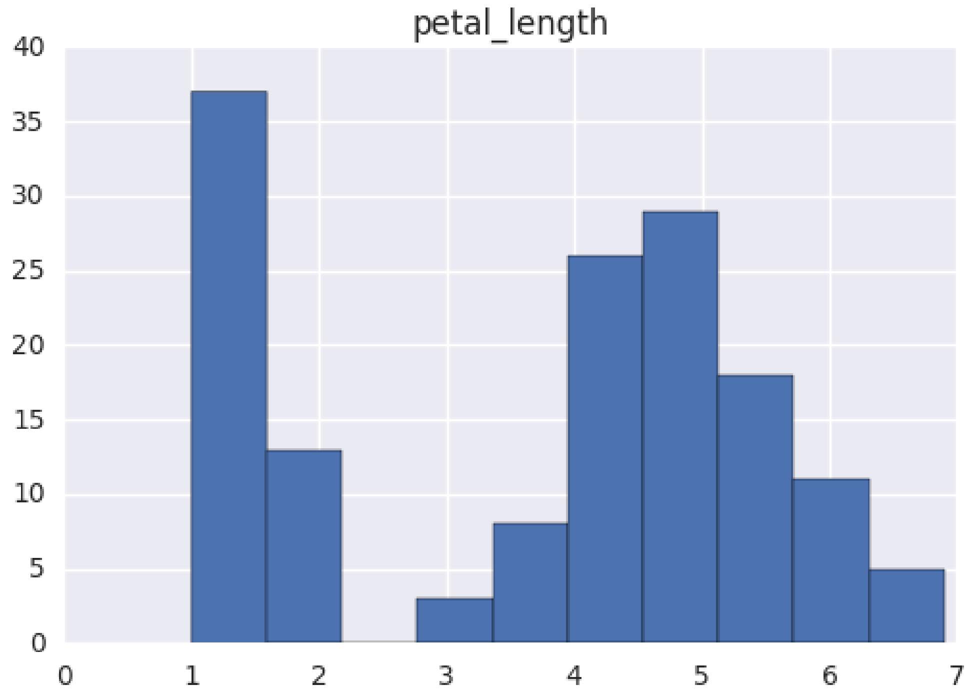


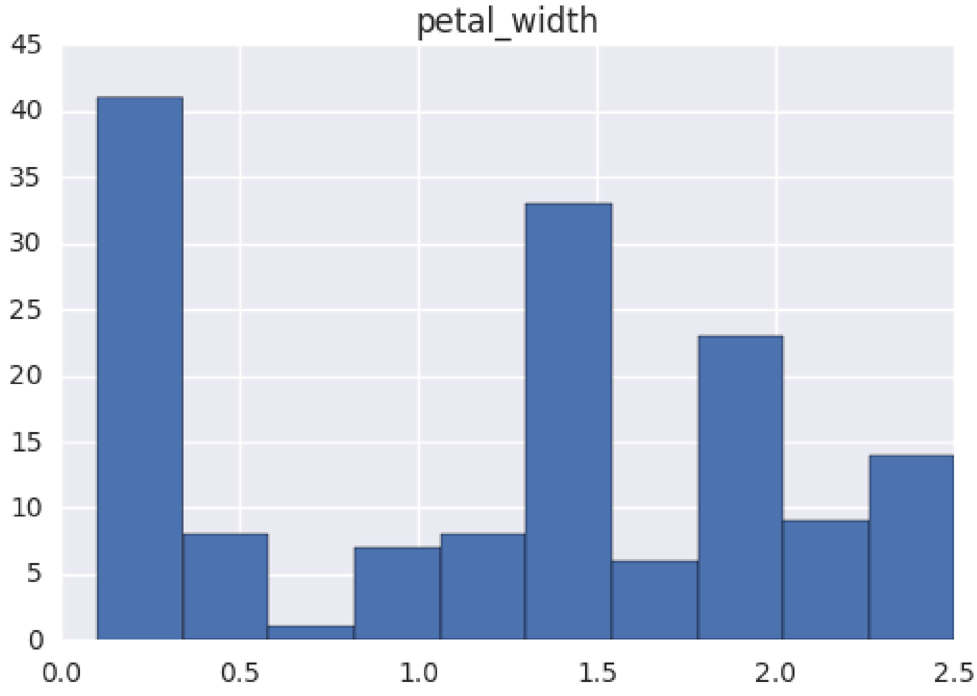
Problem 4-c The sepal length and sepal width are the most "normally" distributed. The petal length appears to by bimodal while the petal width seems to be multimodal.

```
In [61]: for col in df.iloc[:, :4]:
             df.hist(col)
             plt.show()
```


sepal_length
30
25
20
15
10
5
0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


sepal_width
40
35
30
25
20
15
10
5
0
2.0
2.5
3.0
3.5
4.0
4.5


petal_length
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7


petal_width
45
40
35
30
25
20
15
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5

Problem 4-d From looking at all sepal lengths falling outside the 2.7th standard deviation, there were no sepal lengths outside that range. This would indicate that there are no statistical outliers for the sepal length variable.

```
In [76]: #Creating boolean dataframe of sepal lengths outside the 2.7 standard devi
         outlier_std = 2.7

         outlier_trans = df[['sepal_length', 'iris_class']].groupby('iris_class').t

         outlier_df = outlier_trans.abs() > outlier_std
         outlier_df[outlier_df.sepal_length == True]
```

```
Out[76]: Empty DataFrame
         Columns: [sepal_length]
         Index: []
```

Problem 4-e Using the same method as above, the conclusion for petal length was the same. Since there were not petal lengths outside the 2.7th standard deviation, there are no statistical outliers.

```
In [77]: #Creating boolean dataframe of petal lengths outside the 2.7 standard devi
         outlier_std = 2.7

         outlier_trans = df[['petal_length', 'iris_class']].groupby('iris_class').t

         outlier_df = outlier_trans.abs() > outlier_std
         outlier_df[outlier_df.petal_length == True]
```

```
Out[77]: Empty DataFrame
         Columns: [petal_length]
         Index: []
```