CSC240 HW1 Yingcan Chen ID:29849802

1.1

**(a) Is it another hype?**

No, different from many other impractical concept, data mining becomes popular because the need to deal with huge amounts of data everywhere. In short, Data Mining is the method we use to gather information and knowledge from data by discovering the patterns inside, and it has been used to improve products in many different industries.

**(b) Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?**

No, it’s not just a simple transformation, but rather a result of evolution of information technology and development in other disciplines

**(c)We have presented a view that data mining is the result of the evolution of database technology. Do you think that data mining is also the result of the evolution of machine learning research? Can you present such views based on the historical progress of this discipline? Address the same for the ﬁelds of statistics and pattern recognition.**

After searching on the Internet, I found that the classification technique is discovered by deep learning, and pattern discovery is strongly related to statistics. SO pattern recognition is the result of these two fields combined together and they were all very important for data mining, this data mining is definitely benefit from the evolution of machine learning

**(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.**

The first step (data cleaning) is to remove the noise of the data and any inconsistent data in between

The second step (data integration) is to combine data from different sources

The third step (data selection) is to select data that’s relevant to the task from the data base

The fourth step (data transformation) is to transform data to usable form for mining

The fifth step (data mining) is to use proper techniques to find patterns in data

The sixth step (pattern evaluation) identify patterns, do analysis and find out new knowledge

The seventh step (knowledge presentation) is to present new knowledge to users using visualization technique.

1.2

**How is a data warehouse different from a database? How are they similar?**

Although both were used to store data and study pattern. A database is basically a place where interrelated data are stored, whereas data in data warehouse were processed via data cleaning, data integration, data transformation, data loading, and periodic data refreshing. In another word, the data are summarized.

1.4 **Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?**

For example, food chain or grocery stores may want to sell bundles of items and they need to find out what products will sell well together, so they can mining frequent sets to find out the answer. And these patterns cannot be generated without data mining techniques.

1.5 **Explain the difference and similarity between discrimination and classiﬁcation, between characterization and clustering, and between classiﬁcation and regression.**

Discrimination and classification: discrimination is the comparison of features of the target class against that of other classes (use similar methods to characterization), classification finds a model to describe and distinguish different classes. Both compare difference but through different ways

Characterization and clustering: characterization summarize the characteristic of target class and may rely on class-labeled data sets, clustering analyze data without class-label and can be used to generate class labels. Both are used to analyze data

Classification and regression: Classification predicts scattered labels whereas regression deal with functions that were continuous and predict numerical values. Both need relevance analysis

1.7 **Outliers are often discarded as noise. However, one person’s garbage could be another’s treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.**

The two method that are used to detect outliers are statistical tests and distance measures. There are different kinds of statistical tests, but their idea is using the distribution model to find out data that lie outside of the range. While distance measures find outlier by selecting data that is far from clusters. I think in the case of bank fraud, distance measure is more reliable, because it detect abnormally large transaction as outlier and this is exactly what distance measure does, every data far from clusters will be think as outlier

**1.9What are the major challenges of mining a huge amount of data (e.g., billions of tuples) in comparison with mining a small amount of data (e.g., data set of a few hundred tuple)?**

The challenge of mining a huge amount of data is that whether we can make the system only to find interesting patterns and ignore uninteresting ones, if so, can it figure out all the interesting patterns? Will it miss? Will it ignore all the uninteresting pattern? Or it might still generate some uninteresting pattern

In another word, the challenge is to make the system generating all interesting patterns and no uninteresting pattern without missing.

Dataset:

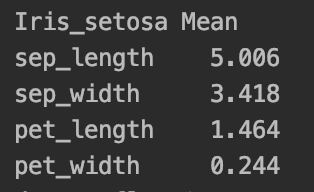
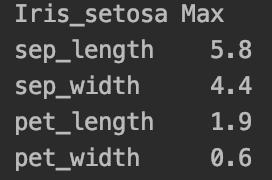
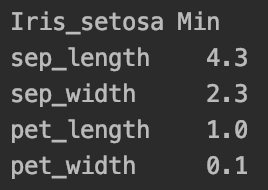
Interpret statistic

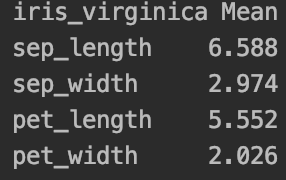
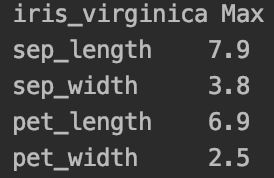
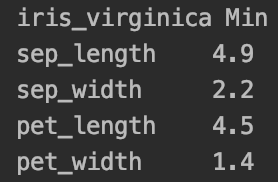
The Data set I am working with it the IRIS dataset which consist of 150 data of 3 kinds of iris plants, 50 each (<https://archive.ics.uci.edu/ml/datasets/Iris>), by looking at this data, I was thinking if I am able to find some pattern in these data so that we can categorize these plants just by their data.

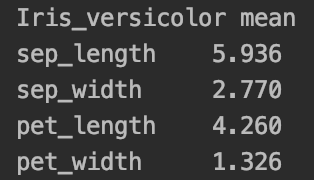
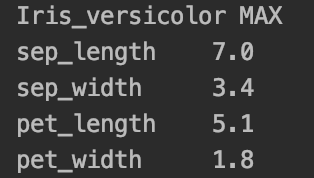
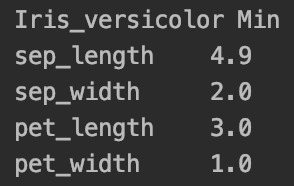
The attribute information for the data:

1. sepal length in cm   
2. sepal width in cm   
3. petal length in cm   
4. petal width in cm   
5. class:   
-- Iris Setosa   
-- Iris Versicolour   
-- Iris Virginica

The First thing I did with the data is that I calculate the Max, min and mean of every measurement for each single kind of IRIS just as shown below:



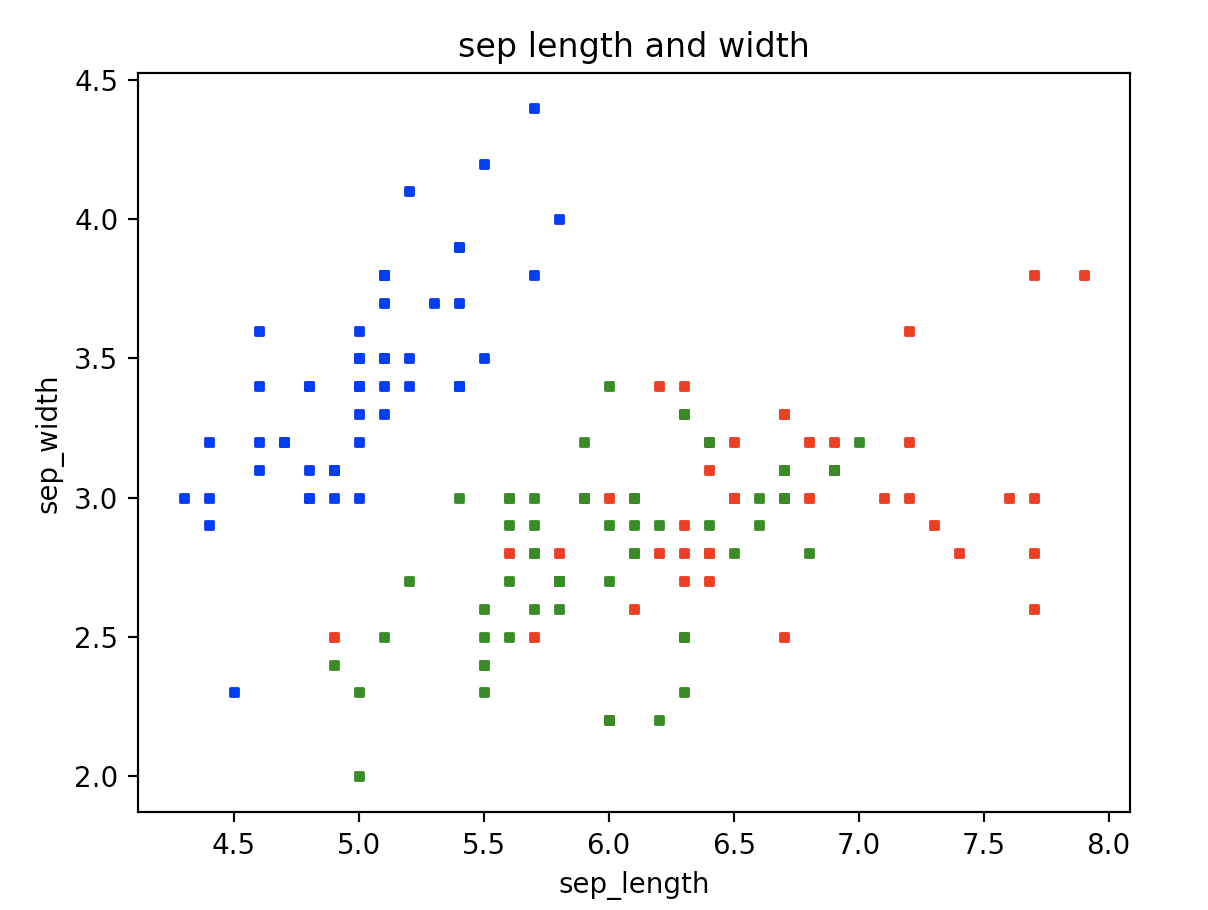




Just from looking at these data, we can see that there is a pretty big difference between Iris-Setosa and the other two, while iris\_virginica and iris\_versicolor are sort of mixing together, they are a little bit difficult to distinguish from each other.

**And the scatter\_plot confirmed my guess:**

Scatter plot using sepal length and sepal width



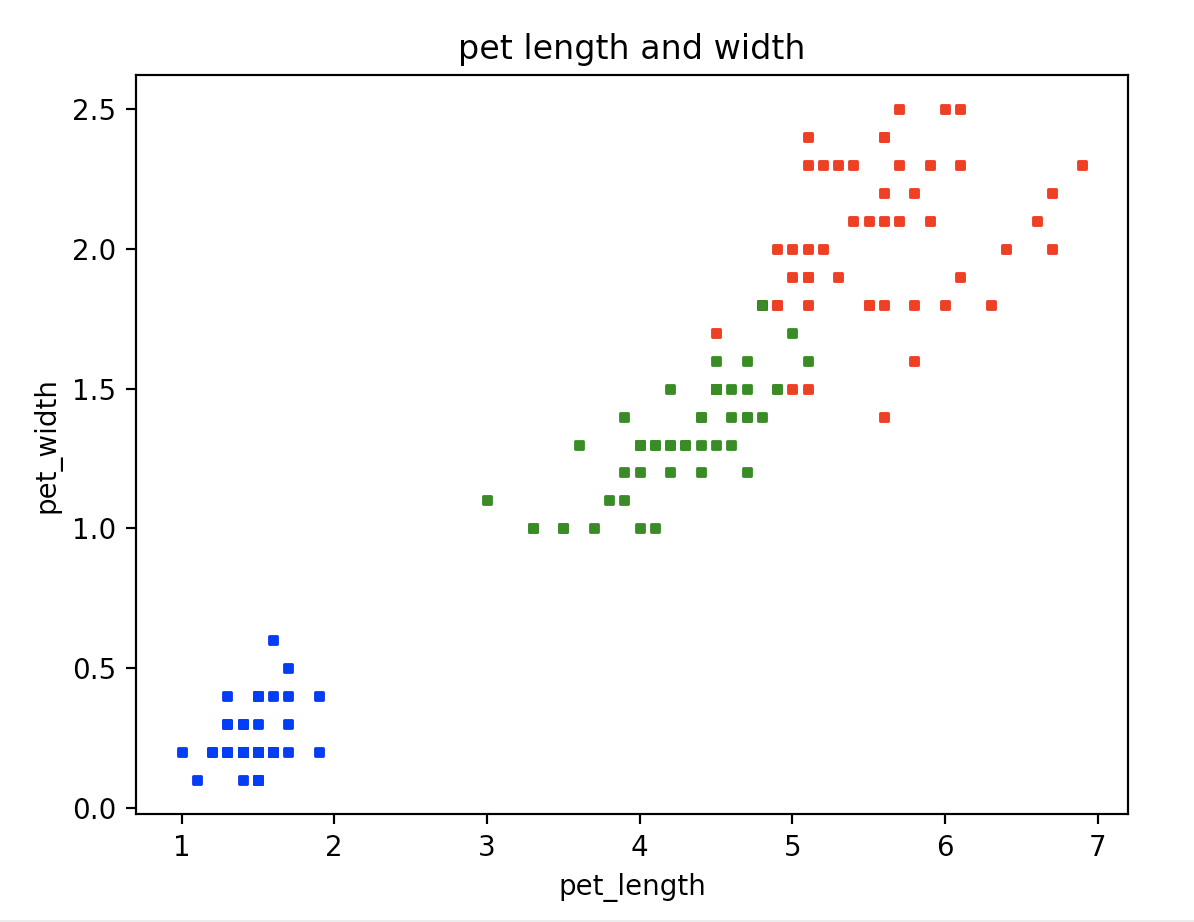
(blue representing Iris-Setosa)

(Green representing iris\_versicolor)

(red representing iris\_virginica)

In this graph we can see that all the blue dots representing Iris\_Setosa are away from red and green dots, while red dots and green dots are clustered together)

Scatter plot using petal length and sepal width



(blue representing Iris-Setosa)

(Green representing iris\_versicolor)

(red representing iris\_virginica)

Same thing happened in this graph where blue dots are far away from green and red where green one and red ones are clustered and inseparable from each other

**For Bonus Point**

So far it seems that we cannot distinguish iris\_versicolor and iris\_virginica by their data, they just seems to be very similar, In order to find a difference, I tried to scatter plot all the combination of data and had a interesting discovery.

When I draw the plot using petal width and sepal length, I found that there is a clear boundary between red dots and green dots and I think this is a clear sign that we are able to distinguish iris\_versicolor and iris\_virginica using this two parameters. I think we will be able to model it by doing a linear regression, and get two linear regression function, and we will be able to distinguish the type of IRIS based on their sepal length and petal width data.

