Liqi Zhu Assignment 1

Part I

1.1

Data mining: the computerized process of discovering valid, novel, useful and understandable patterns or models in massive databases.

(a) Instead of being a hype, I prefer to view data mining as the result of the evolution of information technology, especially when the need for turning huge amounts of data into useful information and pattern is arising rapidly. Data Ming is the result of a natural process.

(b) It's not the simple transformation of technology developed from databases, statistics, machine learning and pattern recognition. It incorporated many techniques from domains such as those and has its own methodology, extensive applications and goals.

(c) As we know, around 1990s scientists began creating programs for computers to do analysis based on large amounts of data and learn conclusions from the results using machine learning technology, one of the main part of data mining is to get the pattern and learn information from data, incorporated with machine learning, I do think that data mining has the combination of machine learning methodology, and without machine learning's development the data mining may face more difficulties than we think. As my opinion, data mining is the result of evolution of machine learning technology.

And for statistics and pattern recognition, that's definitely one of the fundamental technology that data mining incorporated, and the technology of statistics and pattern recognition supports the data mining process, without the development of technology in this area, there can be tool missing or theory missing in the data mining process. In my opinion, data mining is also the result of evolution of statistics and pattern recognition.

(d)

Data cleaning: to remove noise and inconsistent data.

Data integration: where multiple data sources may be combined.

Data selection: where data relevant to the analysis task are retrieved from the database.

Data transformation: where data are transformed and consolidated into forms appropriate for minding by performing summary or aggregation operations.

Data mining: an essential process where intelligent methods are applied to extract data patterns.

Pattern evaluation: to identify the truly interesting patterns representing knowledge based on interestingness measures.

Knowledge presentation: where visualization and knowledge representation techniques are used to present mined knowledge to users.

1.2

Database system normally consists of database files and a database management system. Database files contain original data, indices, metadata of the database logic, and information such as log data for maintaining the database. Which represents the status of the data at a time. While a data warehouse is a decision support system; a structured environment designed to store and analyze all or significant parts of a set of data. The data are logically and physically transformed from multiple source applications into business structure and are updated and maintained over a long time period.

Similarities: Both are used for persistent information storage

1.4

I used to work in data department of Coca-cola China, and for those companies in FMCG market, they have to apply the huge amounts data into finance and marketing strategies and budget based on customers' behavioral and sales volume. Coca-cola China used the database bought from Nelson, and obviously the mining of associations and outlier detection need to be

done to raw data so that detecting the data for application in various cases, outlier detection is always helpful since the number matters a lot when it's directly connected with finance. Mining the connection is the key part of business analysis, without enough useful data and analysis of connection, the finance report and budget for next year is NOT reliable, for those companies, which is unacceptable.

For customers' behavior and sales volume, maybe data query processing and statistical analysis are able to accomplish what we want, but when it came to connections between those numbers, if we want to know the reason of customers' behavior and want the data to support us, than it's not what simple statistical analysis or data query processing can accomplish.

1.5

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. Classification is the process of finding a model or function that describes and distinguishes data classes or concepts. The model is used to predict the class label of objects for which the class label is unknown.

Similarities: Both are used for analysis of the feature of class data.

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query. While Unlike classification and regression, which analyze class-labeled (training) data sets, clustering analyzes data objects without consulting class labels. In many cases, class-labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data.

Similarities: Both are aimed at analyzing grouping together data.

Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data. While classification is the process to find a model or function describes and distinguishes data classes or concepts.

Similarities: Both are tools for prediction, classification is usually used for predicting the class label and regression analysis is used for predicting missing data values.

1.7

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

Both sounds reliable and which one is better really depends. If someone travels a lot then the second method is useless for him. If the frequency is not high enough than unusually large amount also means nothing. So I think the most reliable method is to combine the category of amount, location and frequency together, if the outlier happened in two or three categories, it's more likely to be fraudulent usage.

1.9

First challenge is the efficiency and scalability of the algorithms. With large amounts of data, the algorithm should be efficiency enough to extract information from databases within predictable and acceptable running time.

Second challenge is the processing of the algorithms. Some complicated algorithms require parallel processing and then incorporate into one. Due to the high cost, compare to those small databases, algorithms for huge amounts of data must be dynamic and can be up to date without re-search the data.

Part II
Programming language: Python
Dataset: Iris from UCI

Computation and interpretation
1. Basic statistics results:

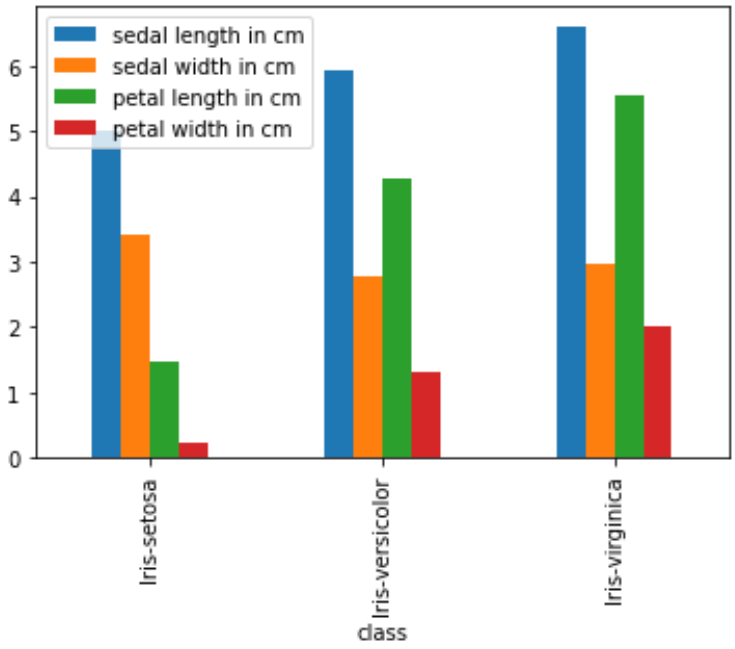
	sl	sw	pl	pw
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

class	Iris-setosa	Iris-versicolor	Iris-virginica
sedal length in cm	5.006	5.936	6.588
sedal width in cm	3.418	2.770	2.974
petal length in cm	1.464	4.260	5.552
petal width in cm	0.244	1.326	2.026

Here showed the basic statistics results with four variables and the mean average of each class of iris. What can be found is that each class has its own feature. Compared to Versicolor and Virginica , Setosa's petal size is smaller than its own sepal size and way more smaller than the other two classes. The pattern of variables are quite the same for Virginica and Versicolor but Virginica is obviously larger.

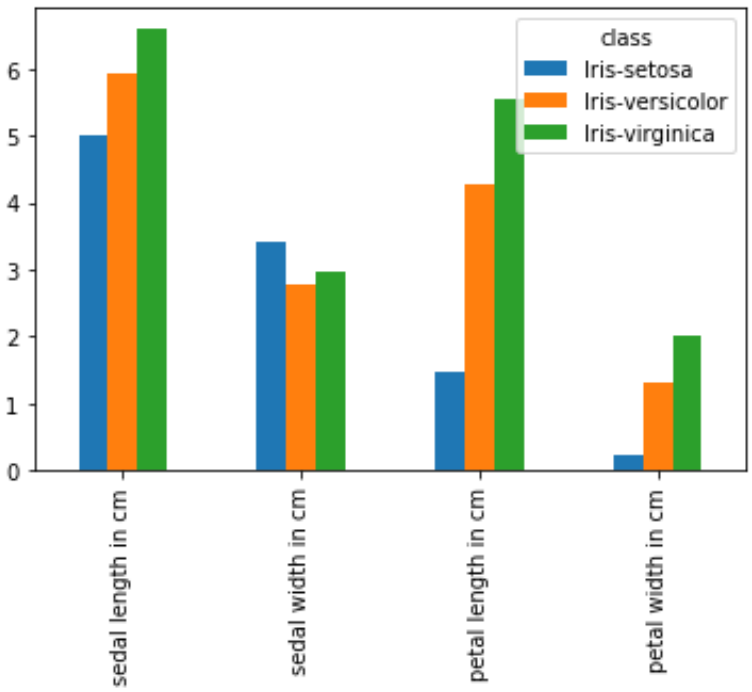
2. Plot

Then I tried to visualize the data with bar plot first, the results are showed:



What can be found in this plot is that the figure and shape of the iris has it's pattern, Setosa's shape is different from neither of the other two, while Versicolor and Virginica share the same pattern of figure.

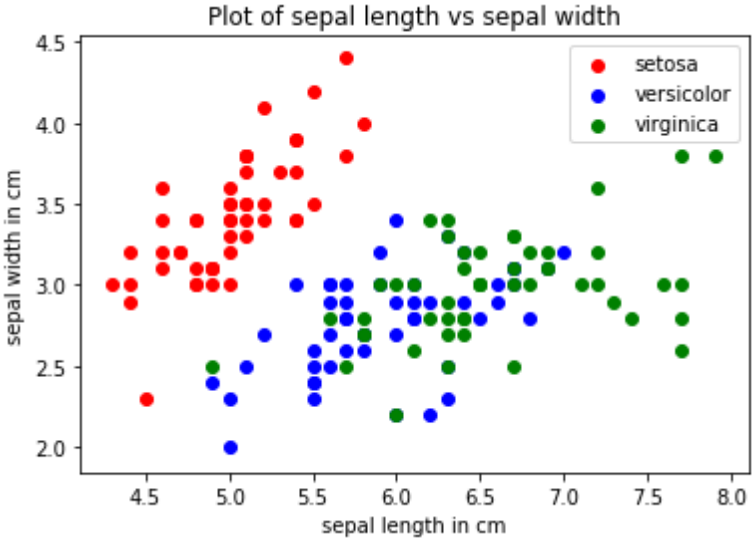
Therefore, Setosa can be the easiest to classify among there.



This basically is the same chart as above, divided by length and width.

The pattern is not as clear as the plot above, but we can find that the biggest difference among three kind of iris is its petal size, not sedal size.

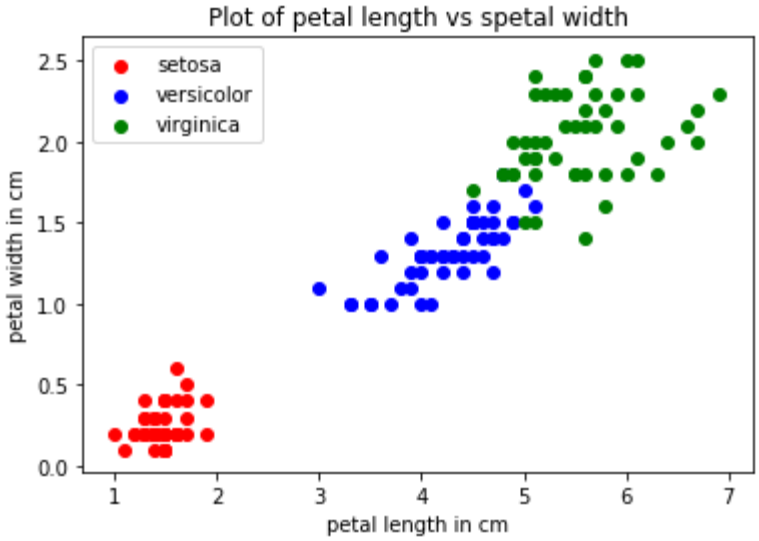
Then I tried the scatter plot to verify the conclusion above and tried to find something new.



It's obvious that Setosa's area doesn't intersect with the other two classes.

While Versicolor and Virginica seem intertwined with each other.

Setosa is the easiest to classify.



When it comes to petal size, Setosa is again the easiest to classify.

Versicolor and Virginica showed their difference, some points are intertwined but it's possible to classify those two with a high accuracy.

Comparison to former plot, it also shows the biggest difference among three is the petal size.