Assignment 1 Meiying Chen

Part 1 Textbook problems

1.1

I do not think Data Mining is another hype. Mainly because this is a highly application-oriented field, according to the growing interests and researching efforts in both academic and industrial aspects, we can infer the data mining technologies did benefit a lot of people in practice. Truly useful technology is not hype.

Data mining is not just a simple transformation or application of the mentioned technologies. There are huge challenges in applying and developing these existing methods of data mining. For instance, to apply statistical analysis on scalable datasets that data mining researchers often dealing with, you may need to design innovative and efficient algorithms to speed up the calculation. Developing an adapted algorithm is never an easy task. Also, it has different aims between doing machine learning researches and applying machine learning methods into data mining research. The former focuses on the algorithm itself and the performance evaluation index like accuracy. However, besides these aims, data mining researches also emphasize the efficiency and ability to handling large size datasets, complex datatypes and the exploration of alternative methods.

Data mining can be viewed as a process of knowledge discovery, which iteratively contains the following steps:

a. Data cleaning: to remove noise and inconsistent data
b. Data integration: combine multiple data sources
c. Data collection: retrieve data that is relevant to the analysis task from databased.
d. Data transformation: using summary or aggregation operations to transform and consolidate data into appropriate forms
e. Data mining: extract data patterns
f. Knowledge representation: use visualization and knowledge representation methods to convey mined knowledge to users

In conclusion, data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

1.2

Both the database and data warehouse can be used to manipulate different kinds of data efficiently. Basically, these all apply store and inquiry functions with efficient data structures. Except for these similarities, data warehouse technology also includes OLAP, which is data cleaning, data integration, and online analytical processing. These functions allow data warehouse to be a tool to summarize data, consolidate conclusions.

1.4

For example, companies using market basket analysis to boost their selling. Market basket analysis is a modeling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This data mining technique may allow the retailer to understand the purchase behavior of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly.

The purchase behavior is the key pattern that needs to be mined. We achieve this using differential analysis comparison of results between different stores, between customers in different demographic groups. This is too integrated that data query processing or simple statistical analysis cannot do.

1.5

	difference	similarities
discrimination VS classification	comparing general features for a population or sample VS contrasting classes of the same population or sample	measure nominal data type and analyze object
characterization VS clustering	models or functions to describe or distinguish data classes to model and predict VS summary of general characteristics or features of the target population or sample	grouping of objects or related data to compare against data set values
classification VS regression	the process of finding a set of models at of the population or sample VS predicts data that it isn't available at the time of the analysis this data is often numerical data values	predict possible trends

1.7

Fraud Detection

Method 1: Rule-based detection system. We may have a bunch of human fraud detection experts concludes how they recognize frauds in the form of rules. The computers detect frauds according to those very rules.

Method 2: Clustering-based program. A clustering algorithm is applied to records that are classified as fraudulent or non-fraudulent. Use this model to identify whether a new record is fraudulent or not.

The second is more reliable, because, it is based on a large amount of data, which made the model less sensitive to specific situations. And the latter one has the ability to learn, which means if the distribution of the dataset changes, say you apply the system to another city, the algorithm will be flexible and adjust itself accordingly.

1.9

As the dataset becoming larger, the efficiency and scalability of the data mining algorithm are challenged. That is, the running time of a data mining algorithm must be predictable, short, and acceptable by applications. Efficiency, scalability, performance, optimization, and the ability to execute in real-time are key criteria that drive the development of many new data mining algorithms.

Possible solutions could be developing a parallel, distributed, and incremental mining algorithm; using cloud computing and cluster computing, etc.

(part 2 is on next page)

Part 2 Playing with data

Dataset: Breast Cancer Wisconsin (Diagnostic) Data Set

url: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Data Set Characteristics:	Multivariate	Number of Instances:	569	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	32	Date Donated	1995-11-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	990016

Statistical descriptions of python terminal outputs:

1. Data overview

Data size is: (569, 31)

Features: ['radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean', 'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se', 'fractal_dimension_se', 'radius_worst', 'texture_worst', 'perimeter_worst', 'area_worst', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'concave points_worst', 'symmetry_worst', 'fractal_dimension_worst']

Classification aim is: 'diagnosis'

2. Measuring the central tendency

Take the feature **texture_mean** as an example:

mean: 19.289648506151142

median: 18.84

mode: 14.93

Because mean > median > mode, we can infer it is a positive skewed distribution. Now draw a figure to validate this:

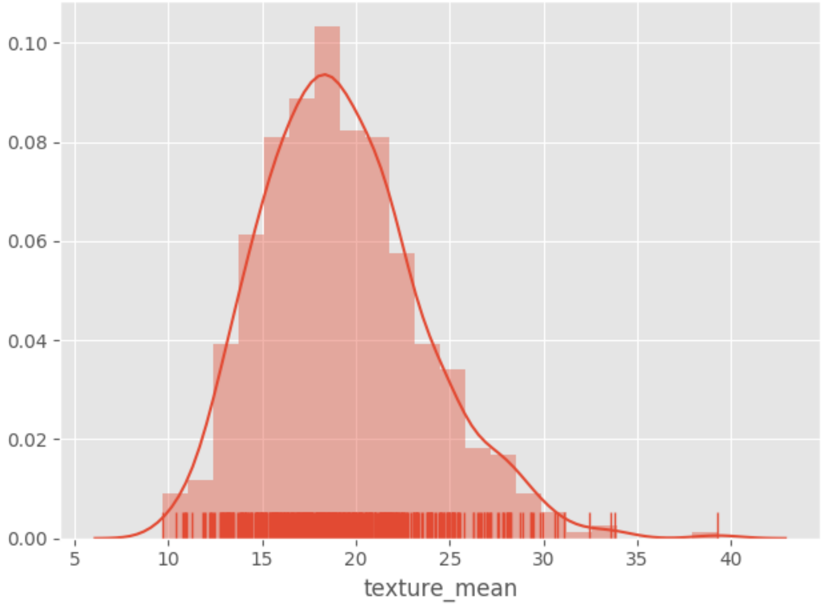

Figure 1 Positive Skewness Distribution

3. Measuring the Dispersion of Data

Name: texture_mean, dtype: float64

describe method	value
count	569
mean	19.289649
standard deviation (std)	4.301036
min	9.710000
25%	16.170000
50%	18.840000
75%	21.800000

max	39.280000

25%, 50%, 75% number are similar and close to the min value, which also indicates negative skewness.

4. Histogram

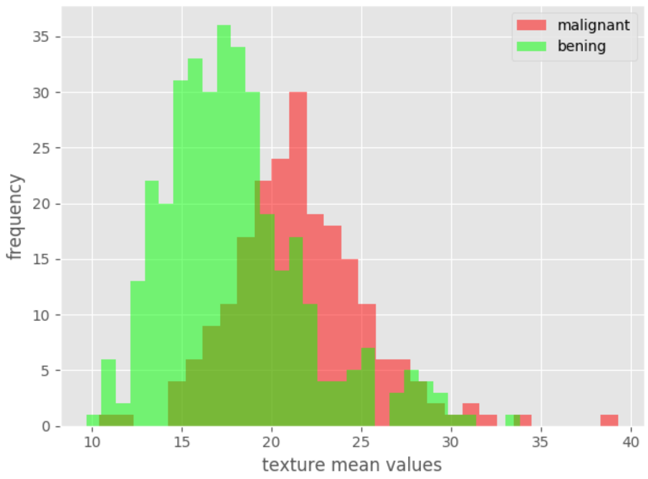

Figure 2 Histogram of Texture Mean for Tumors

As we can see from the picture, the most frequent malignant texture mean is about 21. And the distributions of tow aim values, that is malignant and bening, are different.

5. Box lot

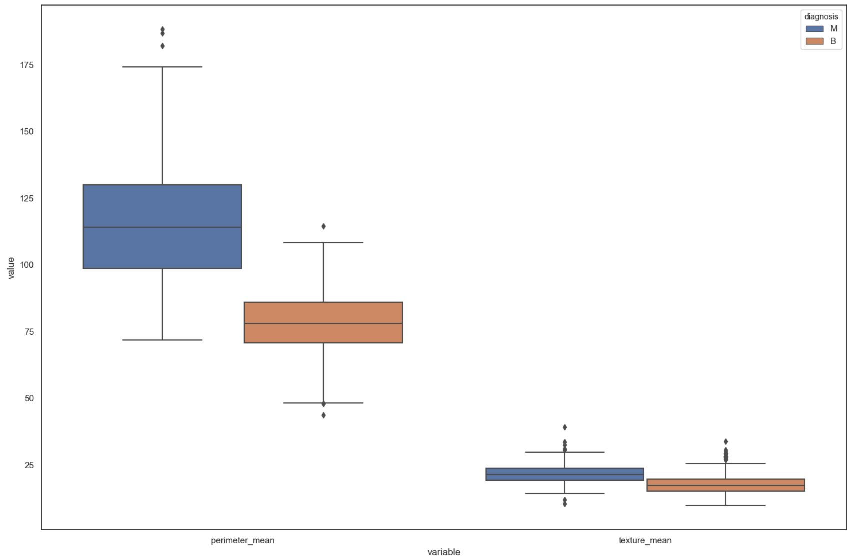

Figure 3 Box Plot

We can use box plot to see outliers. Take parameter_mean and texture_mean as an example, the dots outside the boxes are outliners.

6. Correlation between features

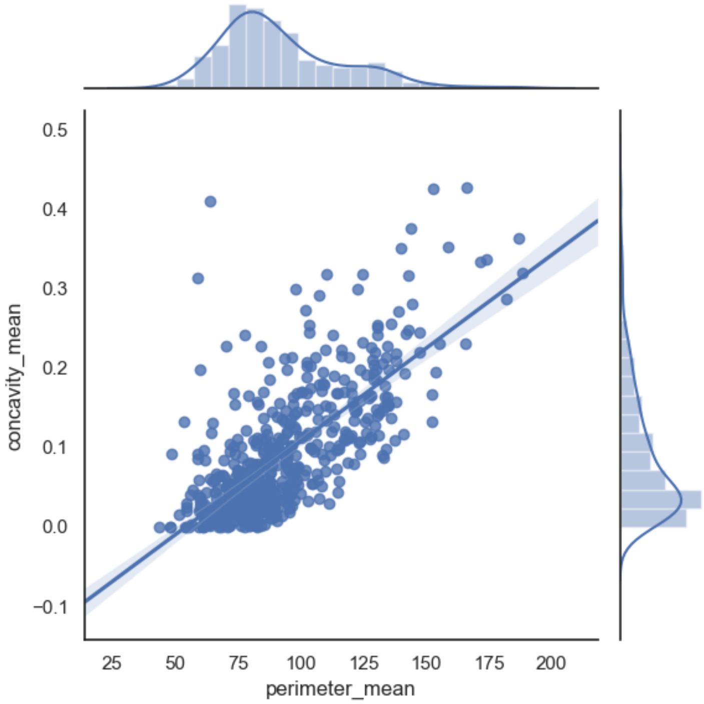

Figure 4 Scatter Plot

Scatter Plot is the simplest way to check relationship between two variables. In scatter plot above you can see that when parameter mean increases, concavity mean also increases. Therefore, they are positively correlated with each other.

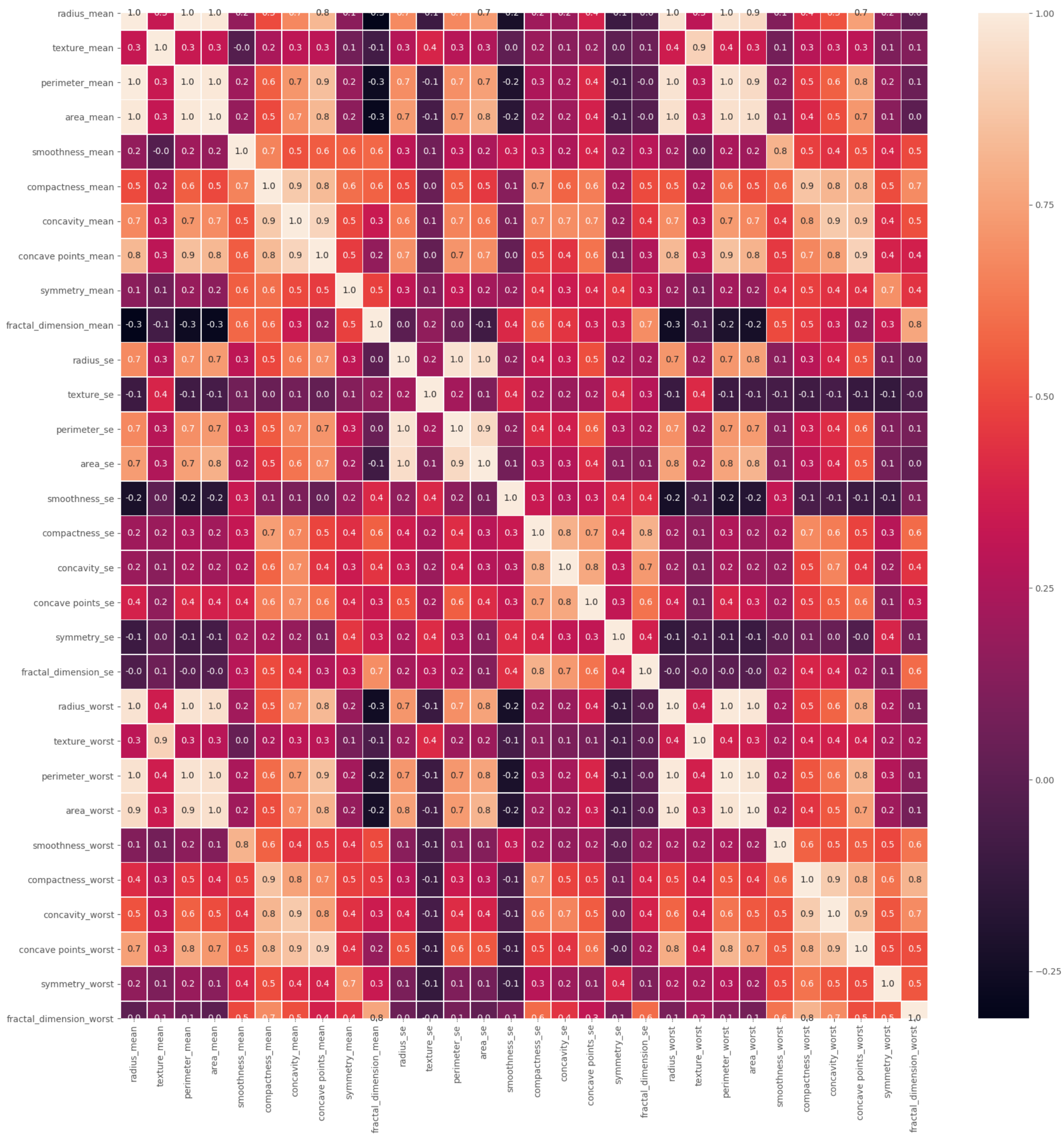

Figure 5 Heat Map

Heat map shows the strength of the relationship between two variables among all the features. For example, concavity mean and parameter mean. According to the yellow little box indicating correlation strength, we can say that concavity mean and parameter mean are positively related with each other.

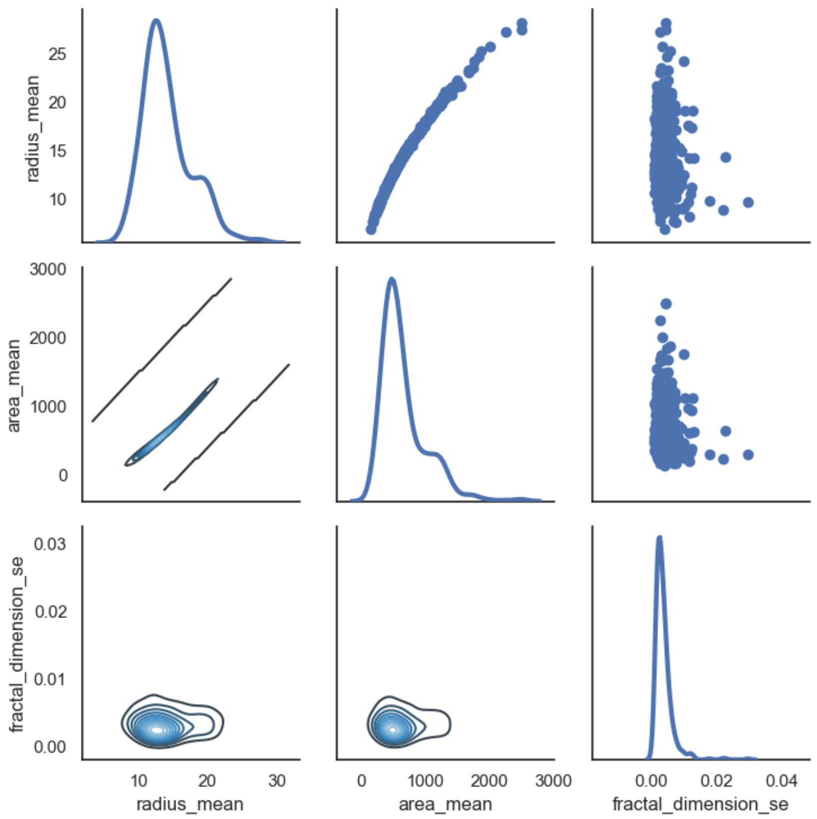

Figure 6 Relationship between 3 Features

We can also plot relationship between more than two features.