A：what is data science?

1: Describe in your own terms what Data Science is.

Data Science is a new area which possess more knowledge and methodology. It is basic on statistics and computer science and we may analyze noisy or huge data and get a prediction or conclusion according to the data.

2: Describe in your own terms the technological advances and other phenomena that have caused the Big Data phenomenon.

The technological advances and it can store more data and save money and we can use data more efficiently. We can analyze the data in machine learning, statistics, data mining etc. so that the model is more powerful.

3: Describe in your own terms what you think are the important skill sets needed to be a data scientist. Elucidate this by drawing on your own experience in carrying out the assignments in this course or other experience you have had outside this course.

Computer science, statistics, machine learning, math, data visualization.

B: Exploratory Data Analysis and the Data Science Process

• Describe in plain terms what Exploratory Data Analysis is. List the kind of things that can be achieved by carrying out EDA.

EDA is an approach to analyze data and show the main character of a set of data. We can use some tools show the model after we analyze the set of data, such as graph, plots. With the help of EDA, we can understand a set of data clearly, we can get the main idea(mean, minimum, maximum) of the data and we may use it to prove our hypothesis or use it find some things, like tendency, we use it to predict what may happen in the future.

• Contrast EDA against Confirmatory Data Analysis (e.g. in terms of what the focus is, what the techniques are, whether or not an assumption is made.)

CDA concerns itself with model and hypothesis. EDA has no hypothesis or model. CDA pays more attention to check the hypothesis which has existed and find if it is right. EDA pays more attention to find the correlation or distribution of a set of data and we may use it to build a model and get some conclusion.

• Discuss some of the basic tools that can be employed to carry out EDA.

Plots, graph, summary statistics.

• Give a schematic description of the Data Science Process and discuss how its components interact. Explain the significance of EDA in the data science process.

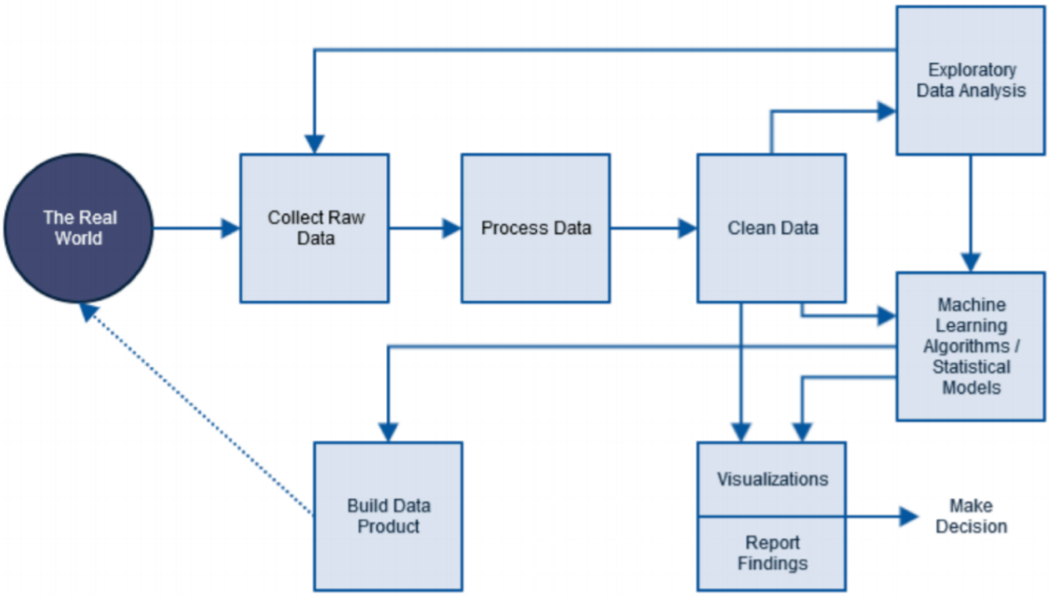


Fig 1. The data science process.

Firstly, in real world, there are lots of activities and we may get some data from them. Then, we can get the raw data based on our activities. However, these data can not be used effectively because they are not clean or may possess some problems. Hence, we may use python or R to clean the data. Then, after we get the clean data, we can do EDA and then, we can build a model(machine learning, regression). Next, we can build a data product and check if it is useful in real life. Then, repeat the loop.

EDA is important because it can help us check if a set of data is actually clean(missing value, absurd outliers). If it is not clean, we may collect more data and clean it again. It can also help us build a model, such like machine learning algorithm or statistics model.

• Describe in what ways EDA helped you to either (i) define a problem or (ii) develop a model in the project or the assignments you carried out in this course.

For instance, in assignment 2, I used EDA to find the correlation between Mpg and some other factors about a car….%$%^$%^$瞎编

C. Data Wrangling

• Use and explain the following functionalities of the data transformation tool dplyr:

– the five “verbs” – filer, select, arrange, mutate, summarise

– Group-by and ranking functionalities

Filter: pick a specified row or column

Select: only keep the columns which we want to use.

Arrange: can sort the value of data (like, from high to low)

Mutate: use the data we already have and build a new column (multiply) based on the data.

Summarise: summary the result of a set of data.(min, max,average)

Group-by: group data of the same name in a column

- Explain what the notion of “tidy” data means

Tidy data means that we use some method, like combine, delate, group or pick parts of column, so that we can use and analyze data more efficiently.

- Use and explain the following functionalities of the tool tidyr – gather, separate and spread

Gather: to combine more groups into one group(combine man and woman into gender)

Separate: to separate one group to more group(U30.F, separate age,gender)

Spread: divide different type(factors) in one column into different type.

- Explain and demonstrate how to work with multiple tables (joins) in dplyr

Find a key which possess in different table and it can connect two or more tables and then, join them in one table.

D. Machine Learning Overview

- Characterize the two major categories of machine learning and clearly state the difference: supervised learning and unsupervised learning.

Supervised learning: we use the data which has been labeled. We can use a map function and the input to find the output.

Unsupervised learning: we are not told the desired output and we may use it to find something we don’t know.

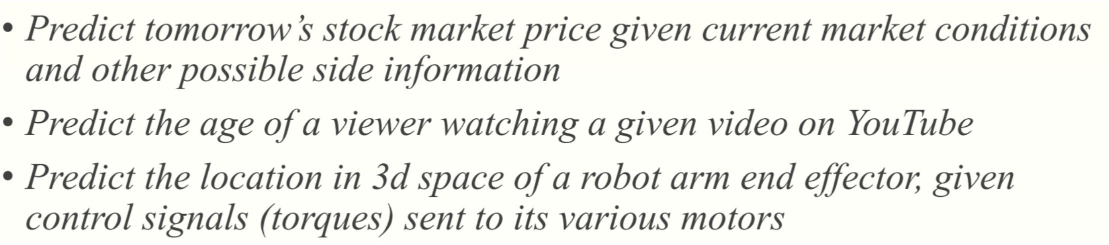
- Describe/characterize when a supervised learning problem is called a classification problem and when it is a regression problem.

A classification problem is to classify a set of data by the features into different types which has been labeled before. A regression problem is similar to a classification problem, however, the response variables are continuous.

- Give an example of a classification problem. Give an example of a regression problem.

classification problem: email spam.

regression problem: predict a age of viewers of a youtube video.



- *Predict tomorrow's stock market price given current market conditions and other possible side information*
- *Predict the age of a viewer watching a given video on YouTube*
- *Predict the location in 3d space of a robot arm end effector, given control signals (torques) sent to its various motors*

E. Linear Regression

• Describe what linear regression is. Discuss how it works.

Linear regression is a linear approach to build a correlation between the response variable and some explanatory variables. We can build a model between the response variable and explanatory variables and find the parameter for each predictors so that we can get a function which can show the relationship.

• Give an intuitive definition of p-value? What does it measure?

The smaller p-value can show that there is a stronger correlation between the two features.

• Give an intuitive definition of R2 ? What does it measure?

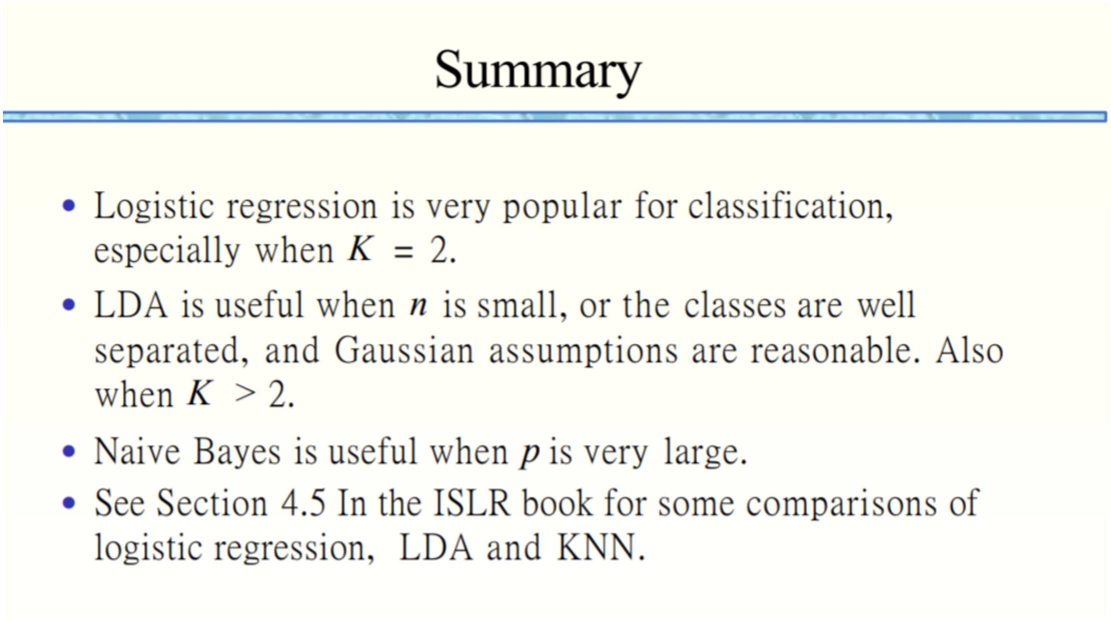
R2 is the coefficient of determination, the domain of R2 is from 0 to 1 and it can show the proportion of predictor variables which can explain or make a connection to the response variable.

• Suppose you are given a linear model with five predictors X1, . . ., X5 in which two of the predictors represent interaction between other predictors. Suppose further you are given the corresponding fitted coefficients β0, . . ., β5. Given these, (i) Predict (calculate) the response for a given combination of predictors. (ii) Answer questions that require interpretation of the regression coefficients

I: $Y= \beta0+ \beta1X1+ \beta2X2+ \beta3X3+ \beta4X4+ \beta5X5$

Ii:

F. Classification



Summary

- Logistic regression is very popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naive Bayes is useful when p is very large.
- See Section 4.5 In the ISLR book for some comparisons of logistic regression, LDA and KNN.

• Explain in basic terms what logistic regression is

Logistic regression is a statistical model which can use a logistic function to classify a set of data.

• Explain in basic terms what linear discriminate analysis is

linear discriminate analysis is a method to find a linear combination of features that characterizes or separates two or more classes of objects or events.

• Explain in basic terms what confounding is

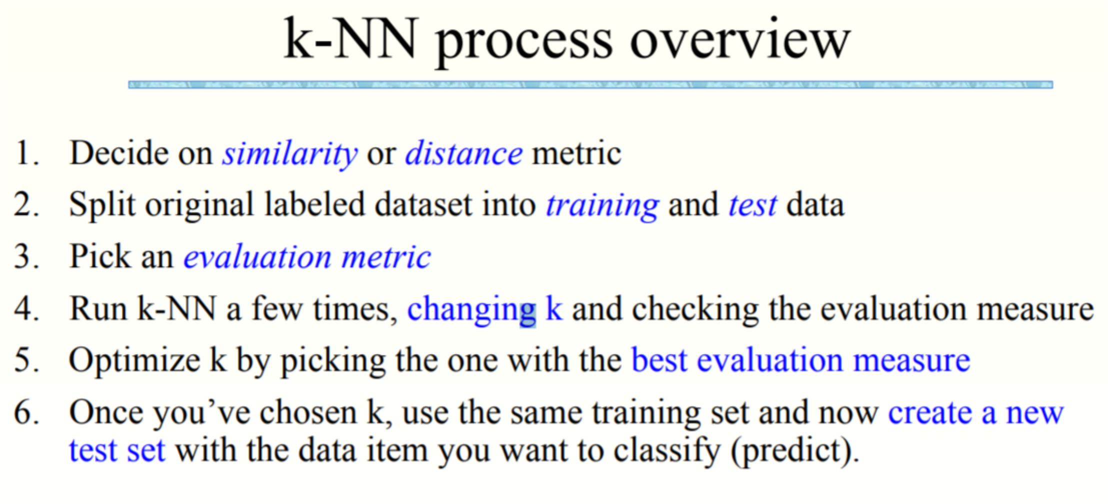
Confounding is a variable which may influence the dependent variable and independent variable.

• Describe in basic terms what Naive-Bayes classifier does

Naive-Bayes classifier can use the naïve-bayes function and get the probability for a event or a response variable. According to the probability, we can classify the training data in different set.

• Describe how the k-Nearest Neighbors classification algorithm works. Give an overview of the k-NN process.

k-Nearest Neighbors classification algorithm can consider the most similar items and put the new item into their group. We should firstly decide a distance metric. Then, split the original data into training set and test set. Next, we pick an evaluation metric. Next, we run K-NN and change k so that we can find the best k for the set of data. After we choose k value, we can use it for prediction.



k-NN process overview

1. Decide on *similarity* or *distance* metric
2. Split original labeled dataset into *training* and *test* data
3. Pick an *evaluation metric*
4. Run k-NN a few times, changing k and checking the evaluation measure
5. Optimize k by picking the one with the best evaluation measure
6. Once you've chosen k, use the same training set and now create a new test set with the data item you want to classify (predict).

• Give an intuitive definition of the following measures of similarity (distance) metrics: Euclidean distance, Cosine similarity, Jaccard distance, Manhatan distance.

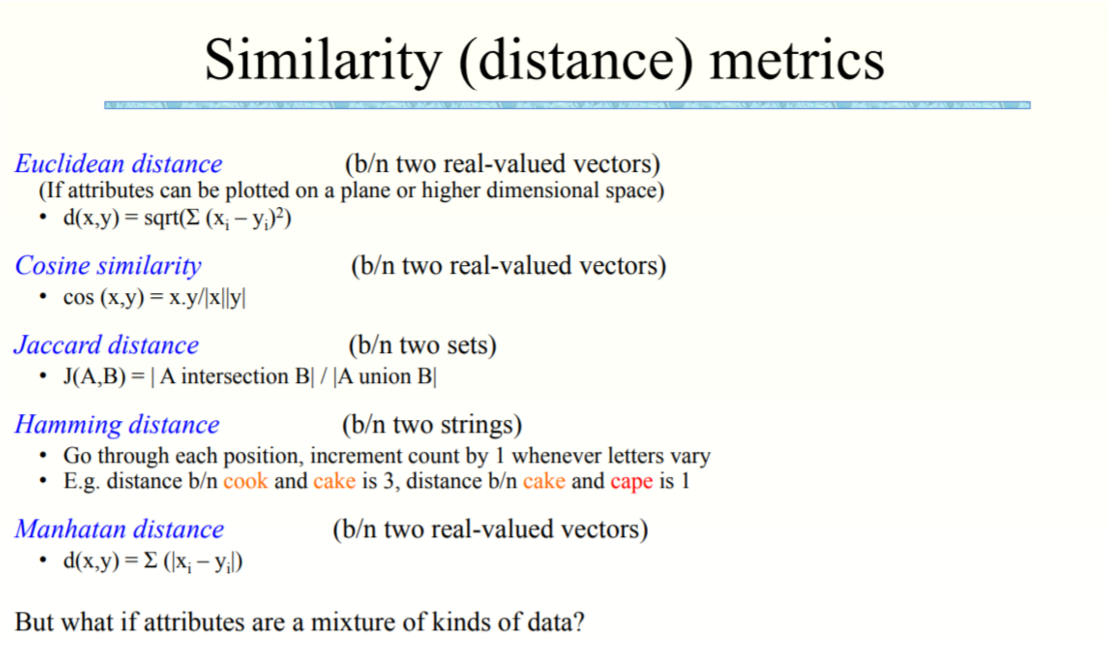
similarity (distance) metrics: find the distance between elements and measure how close two elements are.

Euclidean distance: the distance between 2 points in n-dimension space.

Cosine similarity: measure the similarity(distance) of two vectors' angle.

Jaccard distance: measure the similarity of two sets.

Manhatan distance: the distance between two points in n-dimension in coordinate axis.



Similarity (distance) metrics

Euclidean distance (b/n two real-valued vectors)
(If attributes can be plotted on a plane or higher dimensional space)

- $d(x,y) = sqrt(\Sigma (x_i – y_i)^2)$

Cosine similarity (b/n two real-valued vectors)

- cos (x,y) = x.y/|x||y|

Jaccard distance (b/n two sets)

- J(A,B) = | A intersection B| / |A union B|

Hamming distance (b/n two strings)

- Go through each position, increment count by 1 whenever letters vary
- E.g. distance b/n cook and cake is 3, distance b/n cake and cape is 1

Manhatan distance (b/n two real-valued vectors)

- $d(x,y) = \Sigma (|x_i – y_i|)$

But what if attributes are a mixture of kinds of data?

• Define the evaluation metrics precision. Do the same for recall.

Precision= the number of true positive/ (the number of true positives+ the number of false positives)

Recall= the number of true positive/(the number of true positives+ the number of true negatives)

G. Resampling: Cross-validation and Bootstrap

• Explain what cross-validation is

Cross-validation is a method that divides a data set to training data and validation data so that it can test the model ability to predict new data. It can show the problem of a data set, like overfitting and selection bias.

• Explain how k-fold cross-validation works and why it is attractive

Firstly, we consider about the training data and divide the data into k subsets. Then, we pick each subset as a validation and the other (k-1) subsets as training set. Then, we can get the value of MSE for tach test and we can calculate the average of the MSE so that we can evaluate the data set effectively.

• Discuss how one might do cross-validation the wrong way

We do cross-validation works before we get the outcome. That is wrong.

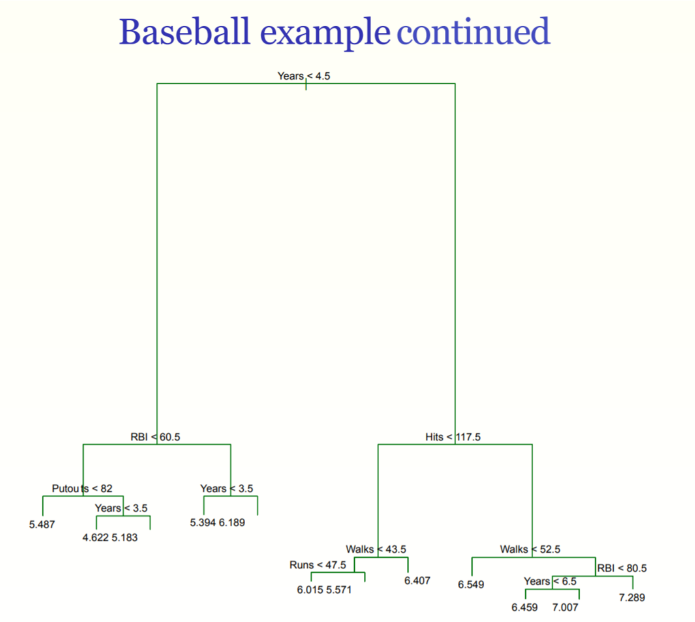
- Explain in plain terms what bootstrap is and what it can be used for

Bootstrap is a statistical tool to quantify the uncertainty associated with a given estimator or statistical learning method.

H. Tree-based methods

- Explain how a tree-based method is used to solve a regression problem

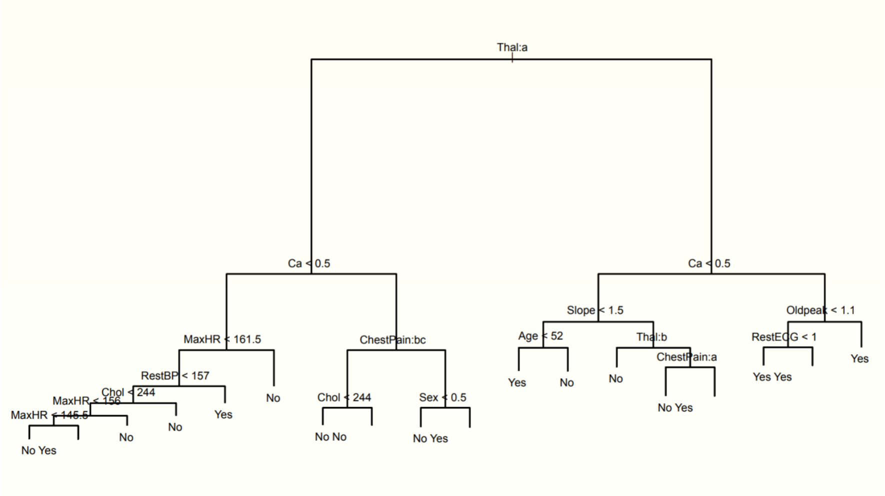
At each internal node, we have a function which can judge it belong to lefthand branch or righthand branch. Since this is a regression problem, it is continuous and the terminal node will be end with a number.



Baseball example continued

- Explain how a tree-based method is used to solve a classification problem

At each internal node, we have a function which can judge it belong to lefthand branch or righthand branch. Since this is a classification problem, the terminal node will be end with a label.



• Give an overview of a tree-based algorithm

Use recursive binary split to grow a large tree and just end at the terminal node which the terminal node has fewer than a value. Then, use pruning tree instead of the large tree. Next, use K-fold cross-validation and minimize the average error as α. Finally, use α and repeat the steps.

• Describe what bagging, random forest and boosting are

Bagging: it is a general-purpose procedure for reducing the variance of a statistical learning method.

random forest: it is an ensemble learning method which can reduce variance when we average the trees.

boosting: it is similar to bagging. It is a general approach that can be applied to many statistical learning methods for regression or classification. just for decision tree.

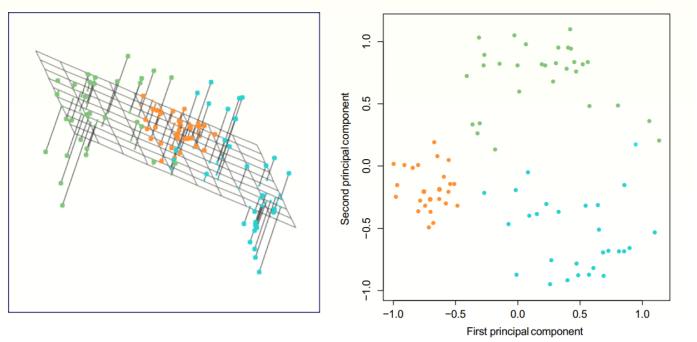
I. Unsupervised Learning: Principal Components Analysis and Clustering Readings:

• Explain the idea behind Principal Components Analysis. What does it find?

PCA can provide a low dimension space of a dataset. It can find a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated.

• Be able to give geometric interpretation of PCA.

To map a data set which in high dimension to a lower dimension space. We may map a set of data into a 2-dimension plot and see it clearly.



• Explain the importance of scaling in PCA.

We can use PCA for data visualization and pre-processing before supervised techniques. It can get some unlabeled data easily.

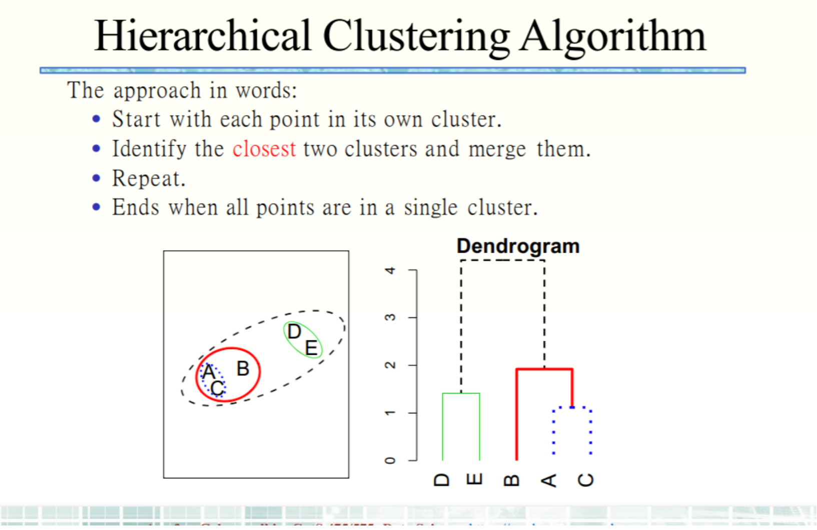
• Explain the notions of "Proportion of variance explained" and "scree plots"

scree plots:

• Describe how the k-means clustering algorithm works.

Firstly, choosing a number k so that we can divide the data in k groups. Then, pick the cluster centroid or one point. Then, we calculate the distance between each point and the centroid(or one point). Next, put the points whose distance is closer to the centroid or the point into one group. Next, repeat the steps until k centroid's distance lower than a value.

• Explain how hierarchical clustering works. Given a dendrogram, be able to interpret what it represents.



Hierarchical Clustering Algorithm

The approach in words:

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.