CZ1016 Notes

Syllabus

1. Basic Data Science Introduction
2. Basic Statistics and Visualization
3. Linear Regression (single, multiple)
4. Classification (tree, random forest)
5. Clustering (k-Means, alternatives)
6. Anomaly Detection (LOF algorithm)
7. Recommendations (similarity etc.)

M1 Topic1 Data Analytic Thinking

Data science Pipeline

CZ1016 Notes

M1 Topic2

Five Primary Questions

How much? How many?	Prediction: Numeric	Regression
Is it type A or type B?	Prediction: Classes	Classification
How is this organized?	Detection: Structure	Clustering
Is it a weird behaviour?	Detection: Anomaly	Anomaly Detection
What should be done next?	Decision: Action	Adaptive Learning

M1 Topic3 Structured Data in Practice

Numeric data	Highly Organized Data Clearly Defined Easy to Mine and Analyze	Numeric Continuous Variables
Categorical data		Factor/Level/Class Variables
Mixed data		Numeric and Categorical
Time Series data		Numeric with Timestamps
Network data		Nodes and Connections

M1 Topic5 Unstructured Data in Practice

Text Data	Highly unorganized data Non-obvious variables Highly context-sensitive	Words, Phrases, Emoticons
Image Data		Pixels and Objects
Voice Data		Voice signals and waves
Video Data		Images, Frames, Objects

M2 Topic 1 Uni-Variate Statistics

<u>Basic Summary of the Data</u>

- Mean: Sum of Data / Count of Data
- Standard Deviation (average deviation from the mean): Sum of Deviation / Count of Data(sum squared of deviation)

CZ1016 Notes

(The standard deviation is expressed in the same units as the mean is, whereas the variance is expressed in squared units)

SD tells the measure of dispersion of the data.

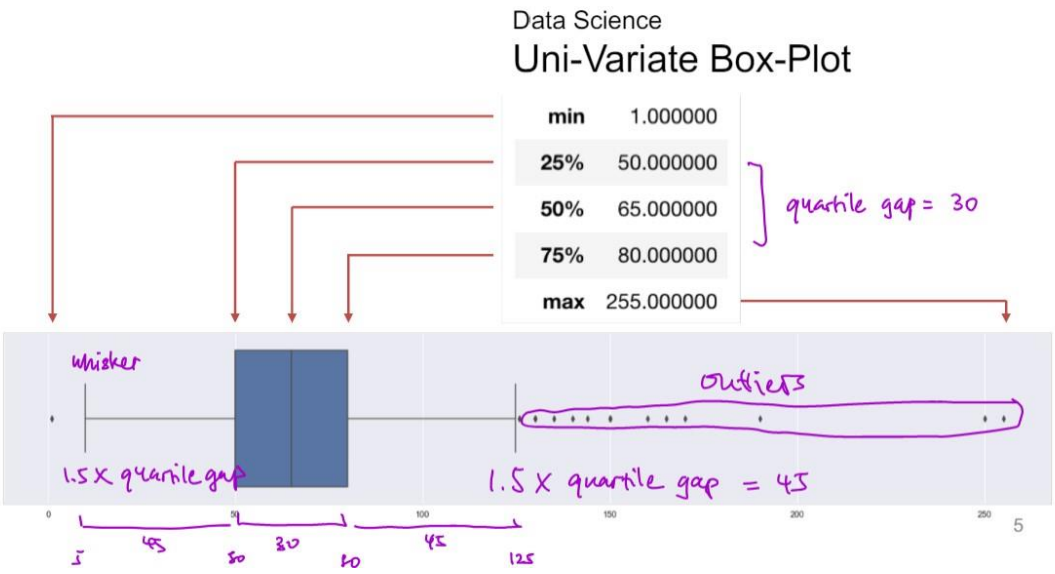
- Median: Marker to Divide the Data 50:50
- Quantiles: Markers to Divide the Data 25:50:25

Statistical Question

- Central Tendency: expected value, average
- Spread of the data: deviation from the average

M2 Topic 2 Uni-Variate Visualization

Box-plot

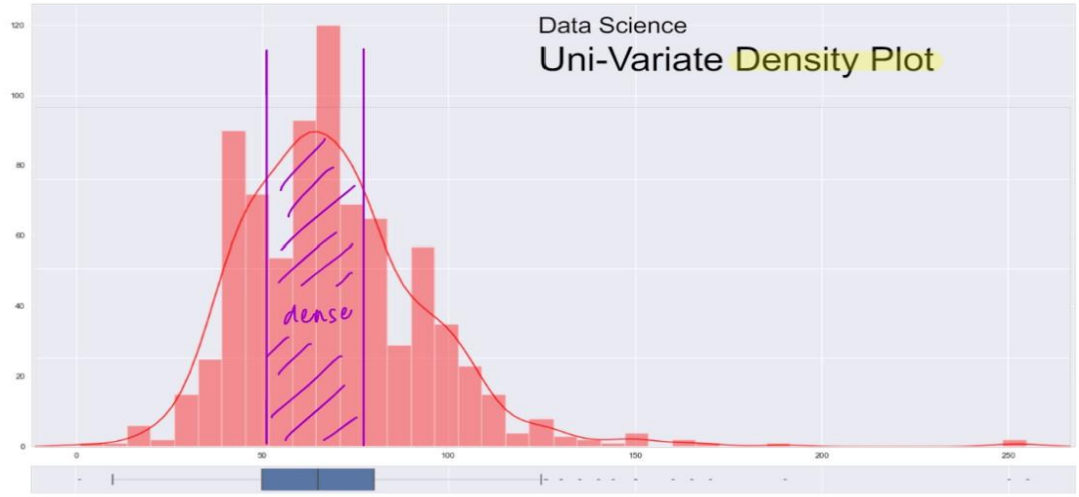


- Most of the datapoints are fall in Box and whiskers
- Outliers: points do not follow the norm
- Good summary of the data

Drawbacks of box-plot: does not show the frequency of individual points

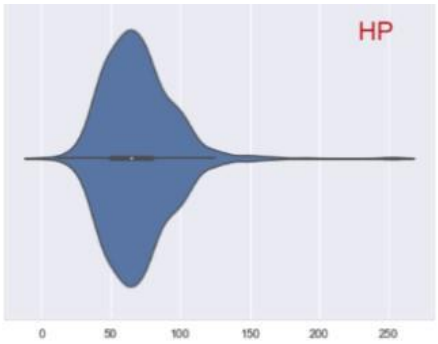
CZ1016 Notes

Density plot: smoother version of histogram

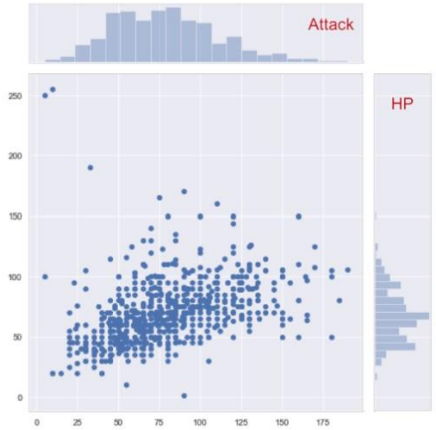


(high frequency=dense)

Violin plot: describe the data in the most comprehensive, and compact manner



M2 Topic 3 Bi-Variate Exploration



Bi-variate joint plot (pair plot)

Pearson's Correlation Coefficient:

Statistical Formula

Co-Variance / St. Dev Product

$$\rho_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$ — covar / — st. dev

11

r = correlation coefficient

x_{i} = values of the x-variable in a sample

\bar{x} = mean of the values of the x-variable

y_{i} = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

CZ1016 Notes

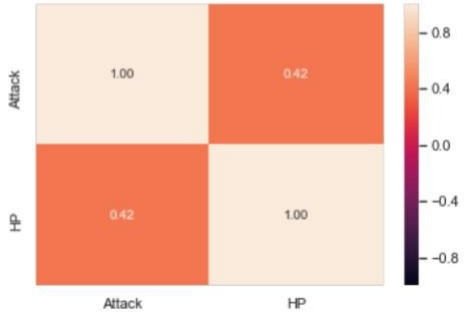
statistical intuition***[important]***

- No Dependence Corr= 0
- Perfect Positive Correlation = + 1 (if hp increase, attack must increase)
- Perfect Negative Correlation = -1 (if hp increase, attack must decrease)
  (+-1 are both good in terms of dependence)

Correlation Matrix and Plot

(heatmap)

	Attack	HP
Attack	1.000000	0.422386
HP	0.422386	1.000000

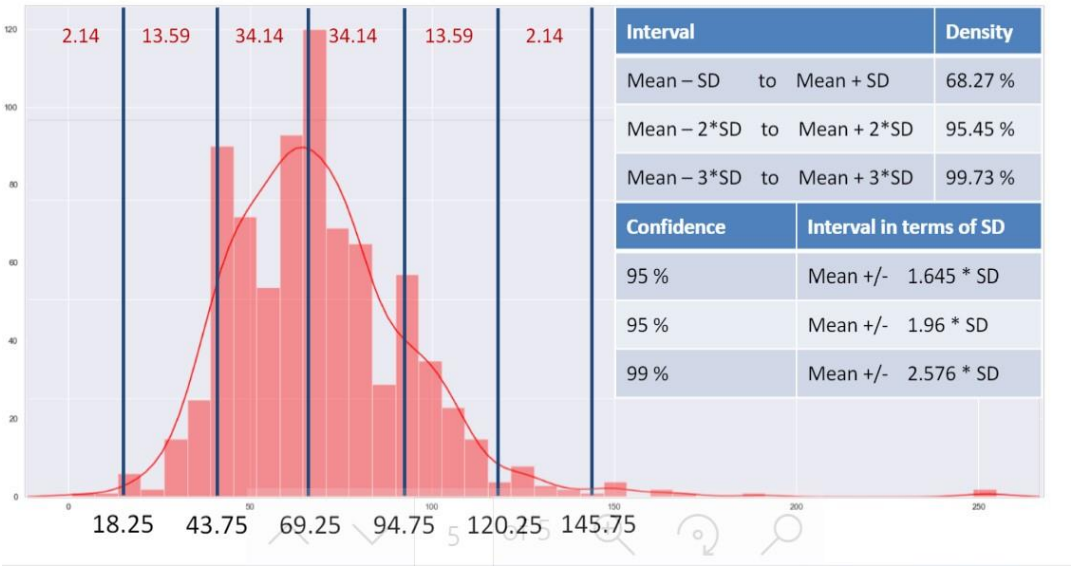


M2 Topic4 Multi-variate exploration

Pattern recognition (plots of multi-variate data)

- distributions of variables
- inter-variable dependence

M2 extra Normal Distribution



Interval			Density
Mean – SD	to	Mean + SD	68.27 %
Mean – 2*SD	to	Mean + 2*SD	95.45 %
Mean – 3*SD	to	Mean + 3*SD	99.73 %

Confidence	Interval in terms of SD
95 %	Mean +/- 1.645 * SD
95 %	Mean +/- 1.96 * SD
99 %	Mean +/- 2.576 * SD

CZ1016 Notes

Observations:

- tails of the distribution is not even
- standard Normal Distribution: MEAN=MEDIAN
- z-value for 90% confidence= 1.645
- z-value for 95% confidence= 1.96
- z-value for 99% confidence= 2.576

M3 Topic1 Basics of Machine Learning

Supervised learning

- *Regression:*

  Model: Total = f(variables)

  Three steps:

  1. Given some data to train (variables and responses(labels))
  2. Learn the model
  3. Predict: Estimate for others
- *Classification:*

  Model: P(legend) = f(variables)

  1. Given some data to train (variables and responses(labels))
  2. Learn the model
  3. Predict: Determine classes for others

Unsupervised learning

- *Clustering:* Grouping depends on "distance"
  1. Given distance on all datapoints
  2. Find optimal groups in the data
  3. Justify interpretation of the groups
- Anomaly detection (identify the weird behaviours/difference)

M3 Topic2 Uni-Variate Linear Regression

Preparation: Split the dataset to train and test

Objectives:

- Learn the relationship from Train (learn the parameter of the model)

  Parametric modelling/parametric learning
- Try to use the model to predict the Total on Test

CZ1016 Notes

Steps in Linear Regression

1. Guess the initial values of the "Parameters" for the hypothesized Linear Model.

2. Predict the values of the Response Variable for all observations in Train data.

Hypothesize a Linear Model

$$\boldsymbol{Total = a \times HP + b + \epsilon}$$

Cost Function to Minimize

$$J(a,b) = \sum (Total - a \times HP - b)^2$$

3. Compute the Errors in Train data, compared to actual values of the Response.

4. Choose a specific Cost Function (like Sum Square of Errors) for Optimization.

5. Reassign or Tune the "Parameters" of the model to Optimize the Cost Function.

*To determine which line is more representative – which line minimize J(a,b)

Ɛ/error/residual: distance from the datapoint to the linear regression line *(not parameter dependent)* It is to be minimised.

Residual sum of squares (RSS): Sum up the square of the errors (difference between actual and predicted values) of all the points in the train set

*once the training set is changed(which 75% of datapoints are used), parameters will also be changed

Goodness of fit (2 methods):

Explained Variance (R^2)

$$R^2 = 1 - \frac{\sum (Total - a \times HP - b)^2}{\sum (Total - \overline{Total})^2}$$

The higher the R^2 the better the Model

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum (Total - a \times HP - b)^2$$

The lower the MSE the better the Model

Explained Variance: 1-RSS/TSS(total sum of squared/variance) = 1- MSE/VAR

- If no error, R^2 is 1 (maximum). Worst case: R^2 is 0.
- 0<=RSS <= TSS: Maximum possible value of RSS is TSS

M3 Topic3 Multi-variate linear regression

Model: Hyperplane

(Eg. 1 parameter-straight line in 2D plane

2 parameters-plane in 3D space)

*Minimize error of J((a1,a2,a3),b)

Hypothesize a Linear Model

$$\boldsymbol{Total = a_1 \times HP + a_2 \times Attack + a_3 \times Defense + b + \epsilon}$$

Cost Function to Minimize

$$J(a,b) = \sum (Total - Total_{pred})^2$$

CZ1016 Notes

$$Response = Coeffs \times Predictors + Intercept + \epsilon$$

Explained Variance (R^2) – higher is better

$$R^2 = 1 - \frac{\sum(Response - Response_{pred})^2}{\sum(Response - Response_{mean})^2}$$

Mean Squared Error (MSE) – lower is better

$$MSE = \frac{1}{n}\sum(Response - Response_{pred})^2$$

Goodness of Fit

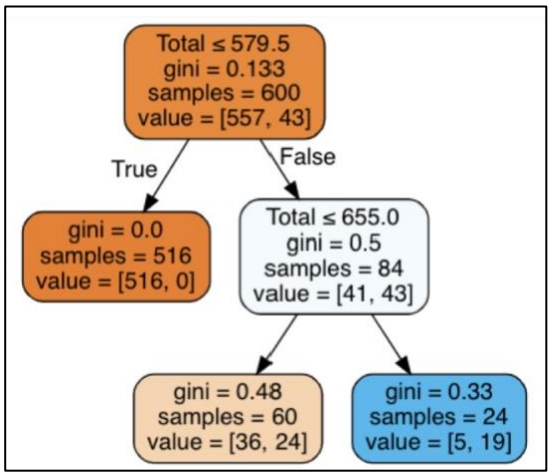
- If there is not any other information, guessing value is usually *mean*.
- Then MSE will be same as variance and R^2 will be 0. (Worst case scenario)
- "explained variance": explain how many % of variance in total

*Goal of linear regression: to explain the response variable in as detailed as possible

M4 Topic1 Binary Classification (2 classes)

Decision Tree:

Partitions made in the Data Space methodically represented using consecutive Binary Decisions



Gini index (metric of misclassification):

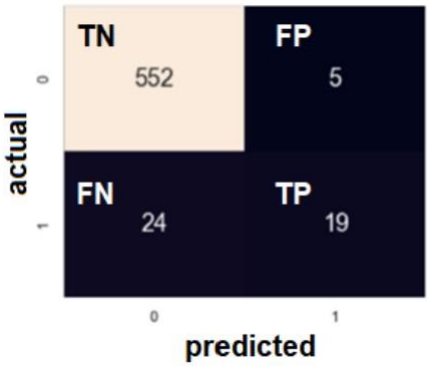
Decision of Partition depends on the Gini Index (metric of misclassification)

$$gini = \frac{x}{n}\left(1 - \frac{x}{n}\right) + \frac{y}{n}\left(1 - \frac{y}{n}\right)$$

- Tells about the *chance of misclassification* in a particular node of the tree (a specific partition of the data)
- How many elements belong to one class VS how many elements belong to the other class
- Gini index: the smaller, the better

CZ1016 Notes

Confusion Matrix



Classification Accuracy

Fraction of Correct Predictions

TP : True predicted as True
TN : False predicted as False

Accuracy in Train Data	0.952
Accuracy in Test Data	0.935

Classification Errors

FN : True predicted as False
FP : False predicted as True

Train	$fpr = \frac{5}{557}$,	$fnr = \frac{24}{43}$
Test	$fpr = \frac{0}{178}$,	$fnr = \frac{13}{22}$

- Based on the question to decide which error to reduce
- classification accuracy = $\frac{TN + TP}{TN + TP + FN + FP}$
- FP rate = $\frac{FP}{TN + FP}$
- FN rate = $\frac{FN}{TP + FN}$

Multi-Variate Decision Tree

When passing 4 variables into a 2-lvl decision tree, only “total” is seen, as total is good enough, and much better than other variables in predicting the legendary.

As the depth of the tree increases, other variables are also included into the prediction.

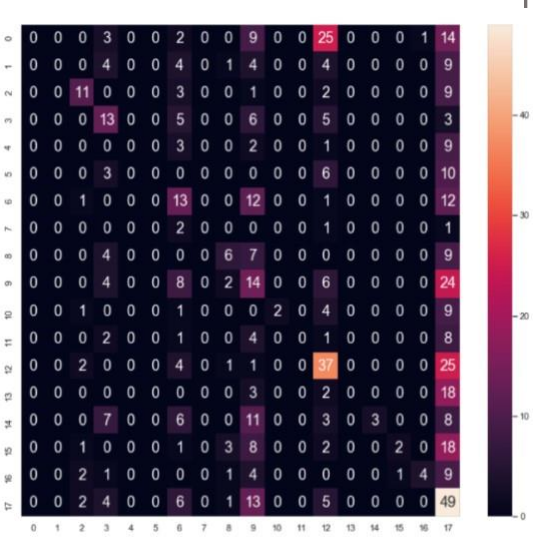
Increasing the number of levels in a tree generally gives a better fit on training data.

M4 Topic 2 Multi-Class Classification

Every binary decision will partition the space into two pieces.

For multiple dimensional spaces, it can choose which axis to partition on.

N X N confusion matrix: main diagonal elements are correct predictions (off-diagonal elements are wrong predictions)



Train Dataset 75% of the Data

CZ1016 Notes

Goodness of Fit of the Model

To improve,

- increase the depth of the tree (*might be prone to overfitting) or
- Merge the types to allow the prediction to work better (too many types are complicating the problem)

Decision of Partition depends on the Gini Index (metric of misclassification)

$$gini = \sum \frac{x_i}{n}\left(1 - \frac{x_i}{n}\right)$$

M4 Topic 3 Random Forest

Which variables for the Decisions?

Repeat for a number of trees:

- Step 1. Choose Variables
- Step 2. Choose Data Points
- Step 3. Create one Decision Tree

Multiple trees with different confusion matrix (different accuracy) – scope for collaboration

(analogy: each team member read 200/500 pages of the book and discuss/correct each other -> trees in a random forest "talk to" each other and correct their mistakes-> take the results out from the trees and take a vote, eg. Peek the answer that the majority predict as correct)

- Scenario A

All Variables and All Data Points

Step 1. Choose all the Variables
Step 2. Choose all the Data Points
Step 3. Create one Decision Tree

Repeat for a number of Trees

Identical Trees – No Collaboration

All trees will behave identically

(one tree doesn't learn anything new from "talking to" another)

- Scenario B

Each person will pay attention to the numerical instead of theoretical portion of the book, even though they are reading different pages, they have the similar training -> not complementing each other

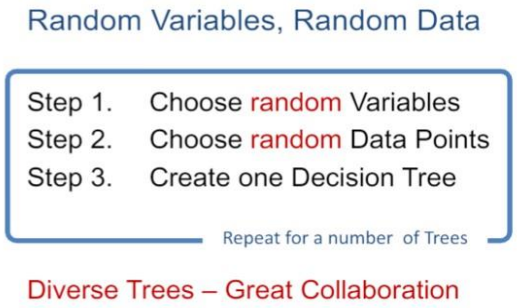
All Variables and Random Data

Step 1. Choose all the Variables
Step 2. Choose random Data Points
Step 3. Create one Decision Tree

Repeat for a number of Trees

Similar Trees – Some Collaboration

CZ1016 Notes



- Scenario C

Complete random forest, most of the trees are different from each other-> different accuracy-> strong and weak points will complement each other

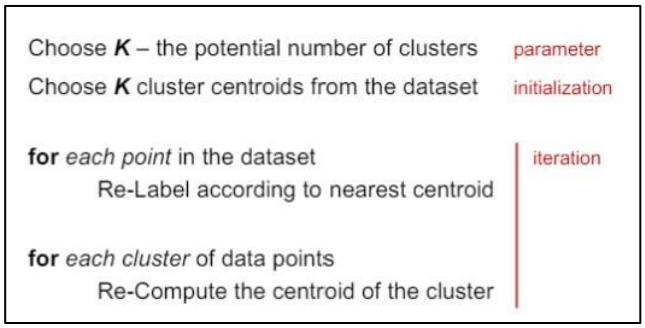
M5 Topic 1 Clustering Patterns

Clusters: points that are close enough

Questions to ask:

- How many clusters are visible?
- Can we identify the clusters?
- What do the clusters signify?

K-means clustering steps:



1. Assume there are n clusters, choose n(arbitrary) *centroids. [*centroid: mean of all the points]

2. For each point, Calculate the distance between it and the centroids to determine which cluster it belongs to, and label the data points

3. For each cluster, recompute the centroid for each cluster, now the partition change again.

4. Relabel the points and recompute the centroids.

5. Repeat steps 3 and 4 until the clusters do not change. (Convergence/Finalization of the algorithm)

How to identify the number of clusters?

- Criteria: Within Cluster Sum of Squares (WSS)
- The fewer number of clusters, the larger the WSS.
- The more the number of clusters, the smaller the WSS (min 0).

CZ1016 Notes

Elbow plot(angle plot): whenever an elbow band, there is a good number of clusters.

A rough way to choose the number of clusters is by inspection of the graph, but this does not ensure that our choice is optimal. To determine optimal number of clusters, we can plot the total within cluster sum of squares (WSS) of all clusters against k. This should give an elbow curve. Here is an example:

Let us say the elbow occurs at k = kelbow.

• Where k < kelbow, WSS is decreasingly quickly and thus increasing k would significantly decrease WSS. Thus, k is not optimal and should be increased.

• Where k < kelbow, WSS does not decrease much as we increase k and thus has reached its saturation. We should consider decreasing k.

• Thus, the optimal k is k = kelbow where WSS has decreased sufficiently before it reaches its saturation.

Drawbacks and solutions of K-means:

• Due to the use of distance, k-means can only be used for **spherical clusters**. For data with non-spherical clusters, we can consider using density-based clustering methods such as **DBSCAN** which includes points which are distance ε apart from the points in the cluster.

• k-means is also sensitive to the **initial choice of points** – we can first try using points farthest from each other and after that try out other random starting points

• k-means is also sensitive to the **number of clusters** – we can try using hierarchical clustering and draw a dendrogram where we do not need to specify number of clusters

• We can also consider using **expectation-maximization (EM)** clustering which maximizes probabilities of within clusters using a Gaussian (normal) distribution, rather than maximising distance. This means our decision boundary is no longer as 'hard' compared to k-means.

M5 Topic 3 Anomaly Detection

Nearest neighbours

Anomaly/outliers: do not have a lot of neighbours, far apart from other points.

Local Outlier Factor (LOF) Algorithm:

1. Choose K–the total number of neighbours

2. Choose d–fraction of anomalies in data (d should be small)

3. For each point in the dataset

- Find the K nearest neighbours in data
- Compute if the density is high enough

When the number of dimensions increase, the anomaly detection becomes more and more accurate (Eg. Even if some points that are not anomalies in 2 dimensions, they might end up being anomalies in more dimensions.)

If project to lower dimensions, some information will be lost.

M5 Topic4 Recommendation Systems

- Tracking customer ratings: Rating of the product (central tendency)
- Tracking purchase pattern(amazon): "customers who bought this item also bought"
- Tracking browsing pattern(amazon): "customers who viewed this item also viewed"
  [Which product they like? Which product they are comparing with? How do they browse overall? What are the alternatives?]
- Learning "similar" products: sponsored product related to this item
- Promoting popular products: "Trending now"/ "popular on Netflix"
- Promoting similar products: "Because you watched___"
- Promoting the choices of "similar" customers: based on the user profile/demographics
- "top picks": what a person watched in the past and what he might be watching in the future

Collaborative filtering:

- A method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Item-similarity & User-similarity

User-Item Matrix

Euclidean distance between users: **small** distance means **high** similarity

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

- Problem: if both users do not buy many items, their Euclidean distance will still be small

CZ1016 Notes

Cosine Similarity: **small** angle (higher value) means **high** similarity

$$\cos(x,y) = \frac{\bar{x}\cdot\bar{y}}{\|\bar{x}\|\|\bar{y}\|} = \frac{x_1y_1+x_2y_2+\cdots+x_ny_n}{\sqrt{x_1^2+x_2^2+\cdots+x_n^2}\sqrt{y_1^2+y_2^2+\cdots+y_n^2}}$$

- Better measure of the similarity

Jaccard similarity: **large** intersection means **high** similarity

$$J(x,y) = \frac{Set(x)\cap Set(y)}{Set(x)\cup Set(y)} = \frac{\text{Common items between } x \text{ and } y}{\text{All items in } x \text{ and } y \text{ put together}}$$

- Large intersection= users who have bought many items that are the same, but does not tell whether the user are frequent in the market place or not
- New user: do not have too many intersections with others, normalise it by union-> better measure of similarity (Eg. 2/2 VS 2/200)

Revision Lecture:

1. First Question(objective type) 40 marks + 4 questions*15 marks
2. Q1 (syllabus 1&2, some calculations may be needed)
3. Q1: some questions are related to visualization:
    a. Box-plot: breaks open the data into equal proportion
    b. Histogram: shows frequency in the “bins”, counts data in specific intervals
    c. Keep in mind the definition of outliers (outliers in box-plot vs outliers in histogram, eg. Mean+-2 SDs for normal distribution)
4. Lin Reg algorithm(4 steps in LAMS), cost function(constants, variables), Multi-variate regression
5. Computing R^2, definition, concepts about overfitting and underfitting (definition, solutions)
6. Prediction: Confidence and Errors **[Extrapolation]**
7. Confusion matrix: computation(no precision issue)
8. Random forest: would you prefer it? What does RF do? What is the randomization point of it? Input data to each of the trees + the splits for each tree at each level is also randomized + controlled by a random subset of the variables
9. Bootstrap (sample) + different variables give rise to different structures of the trees
10. How to get the number of clusters (visualisation + elbow method) expect short essay for >6 marks qn
11. WSS & BSS (Justify the reason why it is considered)

12. Advantage of K-means: very fast, use as an initialization method for other clustering methods (find optimal number of clusters using K-means as benchmark)
13. LOF algorithm and how does density comparison happen
14. Outliers on individual axes vs overall (eg. Points fall inside the Smiling face)
15. Recommendation: Pattern Matching
16. Which notion of “user-similarity” to use (eg. Jaccard or cosine)
17. How to treat sparse data?
18. What is Nearest neighbour?
19. What happens to the confusion matrix when there is a set of imbalanced data? Which rate will be affected?