Selected topics in Advanced Machine Learning

Lecture 02 – Preprocessing

January 24, 2022

Objectives

- Machine Learning: Review
- Missing Values Treatment
- Outlier Detection

+ Review: (What is Machine Learning?)

- "A Computer program is said to ***Learn from Experience*** with respect to some ***class of Task T*** and Performance measure ***P***, if its performance at task in ***T***, as measured by ***P,*** improves with experience ***E".***

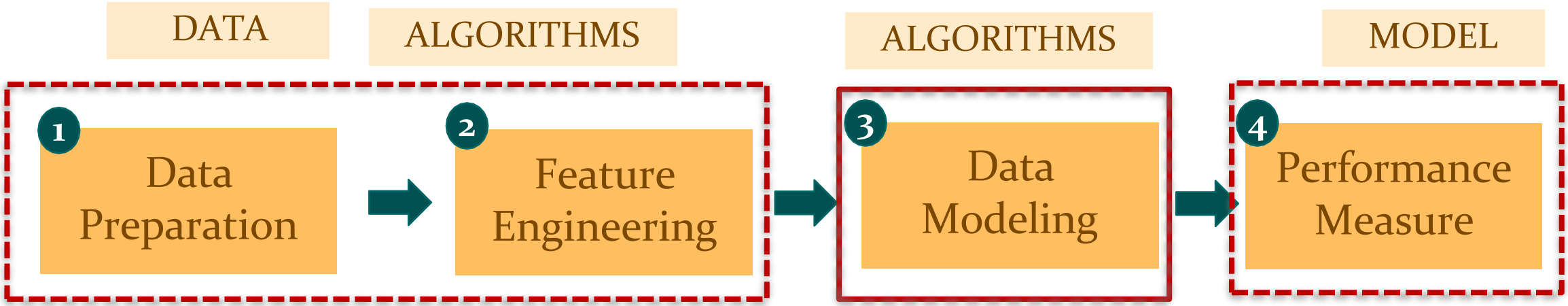
Tom M. Mitchel, Computer Scientist, 1997

What is Machine Learning Model?

- *"A **Machine Learning model** intends to determine the **optimal structure** in a dataset to **achieve** an **assigned task**.*
- *It results from **Learning algorithms** applied on a **training dataset**.*

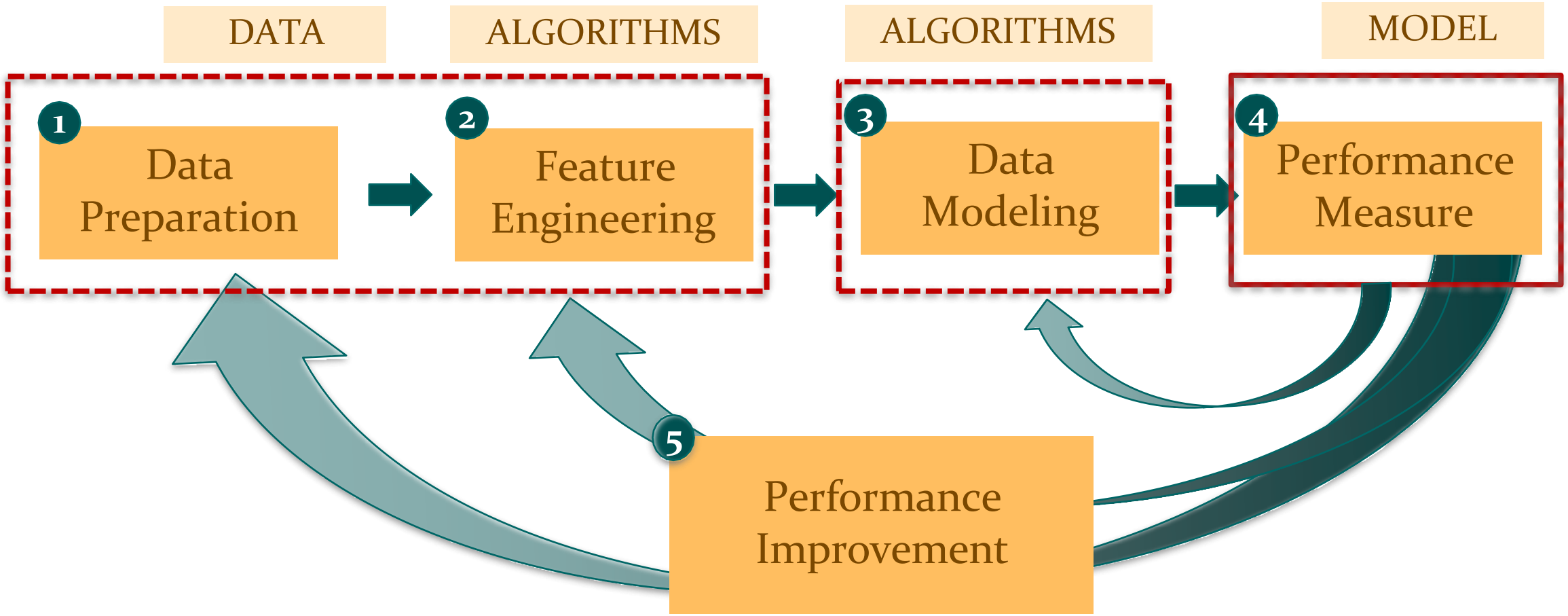
Machine Learning Model

- Machine Learning Model contains 4 basic steps.



Machine Learning Model

- Machine Learning Model is an iterative process.



- Until the model reaches a satisfying performance!

+ Machine Learning Model

1. **Data Preparation**
   - How can you ***import*** your ***raw*** data?
   - What are the ***most common*** data ***cleaning methods***?

2. **Feature Engineering**
   - How do you ***turn raw data*** into ***relevant data***?
   - Turing data to ***meaningful*** for a ***learning algorithm***?
   - How can you make the ***difference*** between ***useful*** and ***useless*** data in a huge dataset?

3. **Data Modeling**
   - What are the different types of ***machine learning algorithms***?
   - Which one should you ***choose*** to build your model?

Machine Learning Model

4 Performance Measure

- What is the ***right method*** to ***access the performance*** of your ML algorithm?
- Which ***indicator*** should you ***use***?

5 Performance Improvement

- What are the ***reasons why your ML model is not performing well***?
- What are the ***most common techniques to improve the performance***?

Three Things about ML

- ***Feature :*** Representation of raw data
- ***Model:*** Mathematical summary of features
- ***Making Something that work:*** Choosing the right model and features, given data and Task

What is Features?

- The ***initial pick*** of feature is always an ***expression*** of ***prior knowledge.***
- ***Images*** → pixels, contours, textures, etc.
- ***Signal*** → samples, spectrograms, etc.
- ***Time series*** → ticks, trends, reversals, etc.
- ***Biological data*** → DNA, marker sequences, genes, etc.
- ***Text data*** → words, grammatical classes and relations, etc.

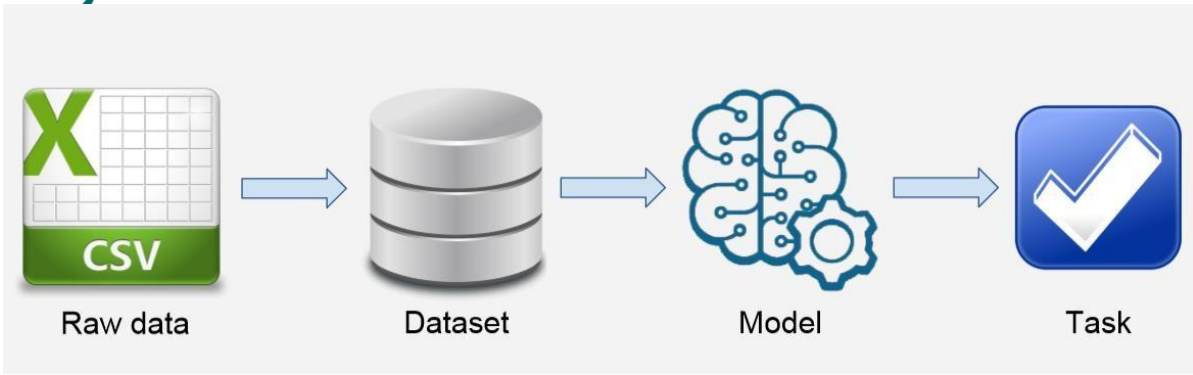
Problem: Where to focus attention ?

- *Garbage In Garbage Out (GIGO)*
- *"Sometimes, less is better!".*

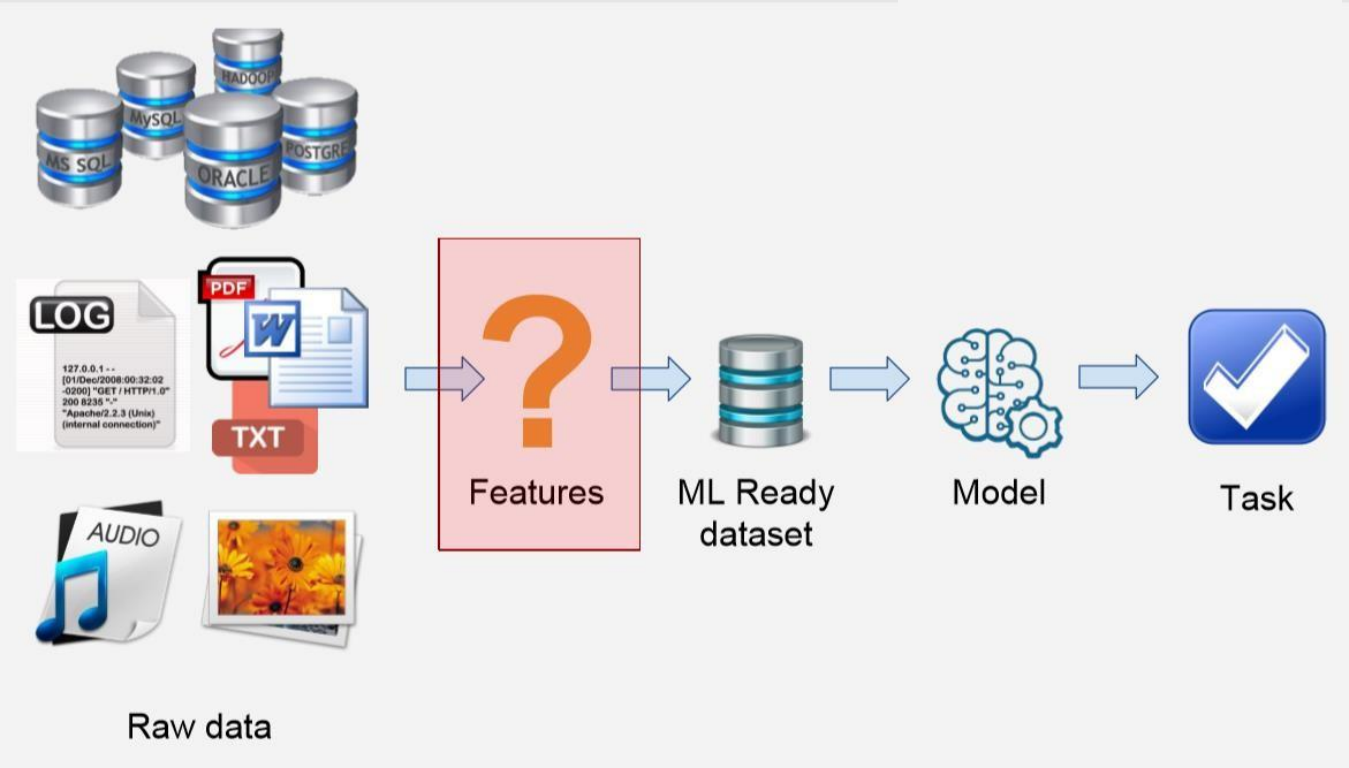
- A universal problem of intelligent (learning) agents is *where to focus their attention*.
- What *aspects* of the *problem* at hand are *important*/necessary to *solve* it?
- *Discriminate* between the *relevant* and *irrelevant* parts of *experience*.

+ Dream Vs. Reality

DREAM



REALITY



+ Missing Values Treatment

- Why missing value treatment is required ?
- Why data has missing values?
- Which are the methods to treat missing value ?

+ Missing Values Treatment

- Missing values are representative of the ***messiness*** of real-world data.
- There can be a ***multitude of reasons*** why they occur—ranging from
  - ***human errors*** during data entry,
  - ***incorrect sensor*** readings,
  - to ***software bugs*** in the data processing pipeline.
- Treating missing data is the ***fundamental*** and ***core element*** for the ***data analysis*** and / or ***machine learning***

Why missing values treatment is required?

- *Missing data* in the training data set can *reduce the power* / *fit* of a *model* or can *lead to a biased model* because we have *not analysed* the *behaviour and relationship* with other variables *correctly*.
- It can lead to *wrong prediction or classification*.
- Example:

Results with not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females.

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Result with treated missing values, we can see that females have higher chances of playing cricket compared to males.

Dealing With Missing Values

- Some of your columns will ***certainly*** contains ***missing values***, often represented as '***NaN', empty column, zeros***.

$$\textbf{\textit{Compute Ratio}}(\textbf{\textit{Missing Values}})R_m = \frac{\textbf{\textit{Number of Missing Values}}}{\textbf{\textit{Total Number of Values}}}$$

- If R_m ***is high***, You might need to ***remove*** the whole ***Column***.
- If R_m is ***reasonable low***, to ***avoid losing data***, you can ***impute*** the ***mean***, the ***median*** or the ***most frequent*** value in place of the missing value

Methods to Treat Missing Values ?

- The best is to get the *actual value that was missing* by going *back* to the *Data Extraction & Collection* stage and *correcting possible errors during these stages*.
  - ***Which is not possible in most of the cases***

- There are *two main* techniques to treat missing data.
  - Deletion
  - Imputation

1. Deletion

- Unless the nature of missing data is '**Missing completely at random**', the best avoidable method in many cases is deletion.

- ***Listwise:*** In this case, rows containing missing variables are deleted. It suffers the ***maximum information loss***.

User	Device	OS	Transactions
~~A~~	~~Mobile~~	~~NA~~	~~5~~
B	Mobile	Android	3
~~C~~	~~NA~~	~~iOS~~	~~2~~
D	Tablet	Android	1
E	Mobile	iOS	4

- ***Pairwise:*** In this case, only the ***missing observations*** are ***ignored***, and analysis is done on ***variables present***. The problem is that even though it takes the available cases, ***one can't compare analyses because the sample is different every time***.

User	Device	OS	Transactions
A	Mobile	~~NA~~	5
B	Mobile	Android	3
C	~~NA~~	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4

- ***Deleting Columns:*** In most cases if the missing data constitutes more than 90% of the data then the column is dropped as it would not contribute to the mode.

2. Imputation

- Replacing With Mean/Median/Mode
- Assigning A Unique Category
- Assigning A Most frequent Value
- Using Algorithms Which Support Missing Values

1. Replacing With Mean/Median/Mode

- This strategy can be applied on a ***feature*** which has ***numeric data*** like the age of a person or the ticket fare.
- We can calculate the ***mean, median or mode of the feature*** and ***replace*** it with the missing values.
- This is an ***approximation*** which can add ***variance*** to the data set.
- It is a ***statistical approach of handling the missing values***

OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

OS	Global Mean	Group Mean
Android	1,804	1,804
iOS	3,027	3,027
iOS	8,788	8,788
Android	4,145	2,696
Android	3,735	3,735
Android	1,056	1,056
iOS	9,319	9,319
Android	6,199	6,199
Android	2,235	2,235
iOS	4,145	7,045
Android	1,146	1,146

2. Assigning a Unique Category

- A ***categorical feature*** will have a ***definite number of possibilities***, such as gender, for example.
- Since they have a definite number of classes, we can ***assign another class*** for the missing values.
- Missing values can be treated as a ***separate category*** by itself.
- The missing values which can be replaced with a new category, say, ***U for 'unknown'***.
- This strategy will ***add more information*** into the dataset which will ***result in the change of variance***.

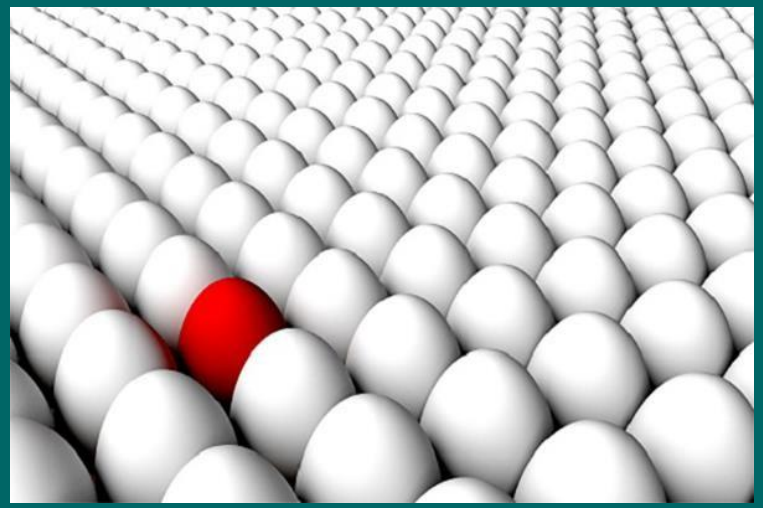
3. Assigning a Most frequent Value

- **Frequent Value:** The standard thing to do is to replace the missing entry with the most frequent one

4. Using Prediction Algorithm

- **Prediction models:** We can create a predictive model to estimate values that will substitute the missing data.
- **KNN** is a machine learning algorithm which works on the principle of ***distance measure***.
- This algorithm can be used when there are ***nulls*** present in the dataset.
- While the algorithm is applied, KNN considers the missing values by taking ***the majority of the K nearest values***.
- **RandomForest:** This model produces a ***robust result*** because it works well on ***non-linear and the categorical data***.
- It adapts to the data structure taking into consideration of the ***high variance or the bias, producing better results on large datasets.***

Outlier Detection



- Outliers and Outlier Analysis
- What Are Outliers?
- Types of Outliers
- Challenges of Outlier Detection
- Outlier Detection Methods
- Application of Outlier Detection
- Evaluation

+ Anomalies/Outliers

- ***We are drowning in the deluge of data that are being collected world-wide, while starving for knowledge at the same time****
- Anomalous events occur ***relatively infrequently***
- However, when they do occur, ***their consequences can be quite dramatic and quite often in a negative sense***

**"Mining needle in a haystack.
So much hay and so little time"**

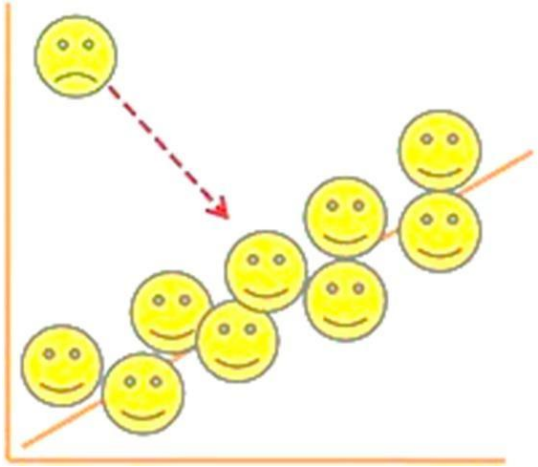
- ***Anomaly*** is a pattern in the ***data that does not conform*** to the ***expected behavior*** also referred to as outliers

What is an Outlier?

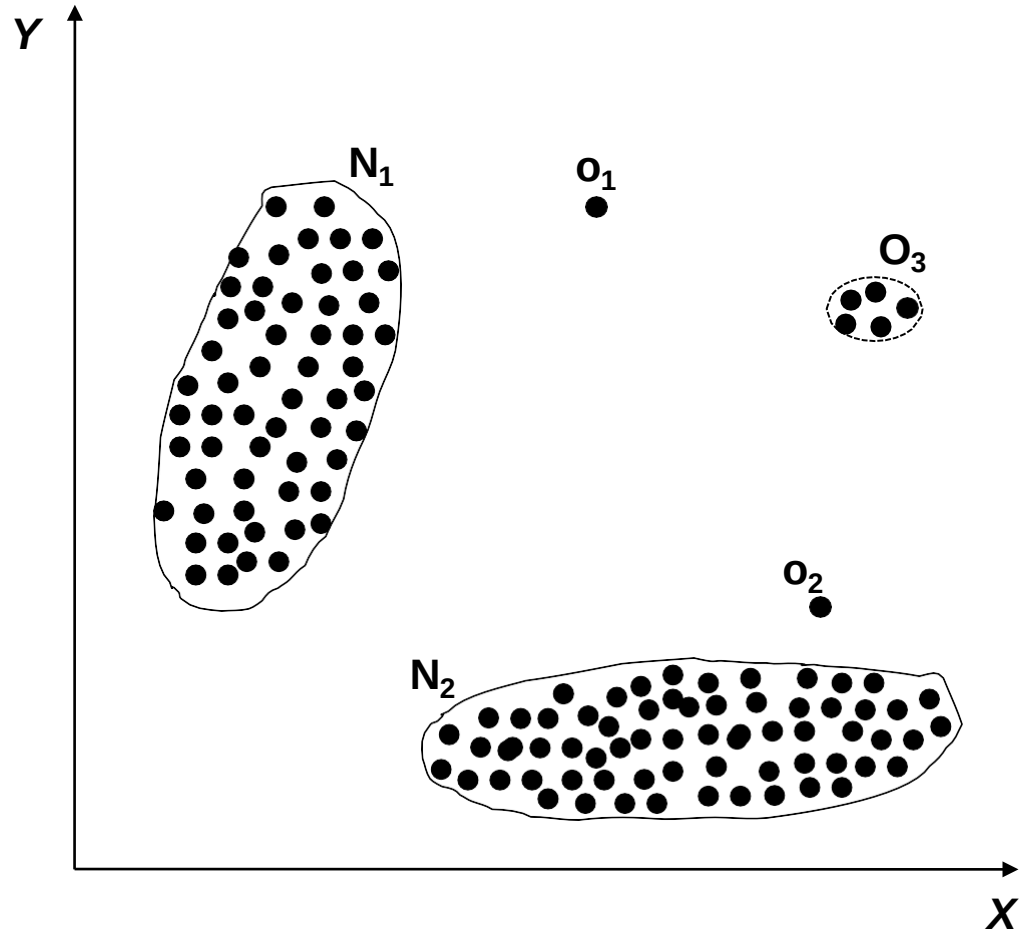
- **Outlier** is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations.

- **Simply speaking**, Outlier is an observation that appears far ***away*** and ***diverges*** from an ***overall pattern*** in a sample.

How do you even detect the presence of outliers and how extreme they are?



+ Example



- N_1 and N_2 are regions of normal behavior
- Points o_1 and o_2 are anomalies
- Points in region O_3 are anomalies

- Example: Age of a person

Anomalies/Outliers

- **What are outliers?**
  - An outlier is a ***data object*** that ***deviates significantly*** from the ***rest of the objects,*** as if it were generated by a ***different mechanism***.
  - We may refer to data objects that are not outliers as "***normal***" or ***expected data***. Similarly, we may refer to outliers as "***abnormal***" data.
- Also referred to as outliers, ***exceptions***, ***peculiarities***, ***surprise***, etc.
- **Outliers are different from noisy data**
- ***"What is noise?"***
  - Noise is a ***random error*** or ***variance*** in a measured variable.
- ***Outliers are interesting:*** an outlier violates the mechanism that generates the normal data.
- Noise is ***not interesting*** in data ***analysis***.

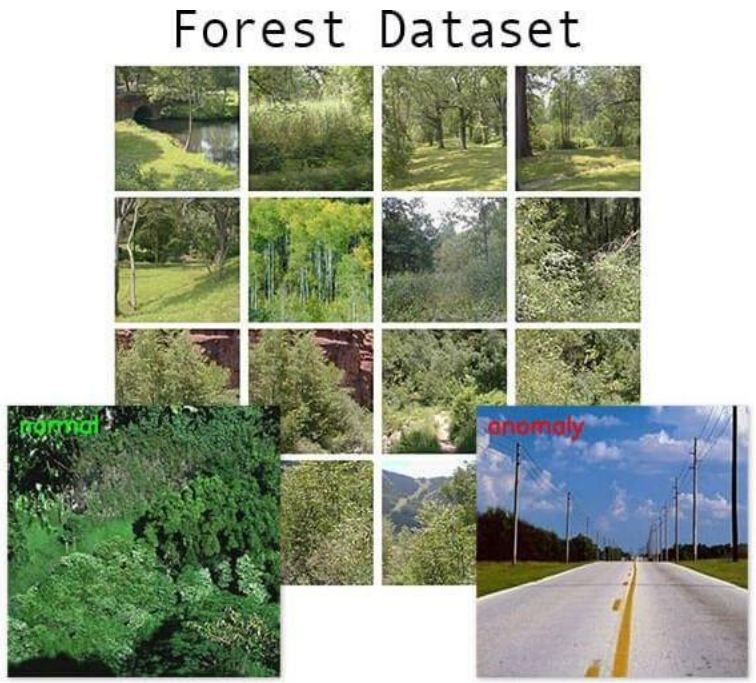
+ Anomalies/Outliers

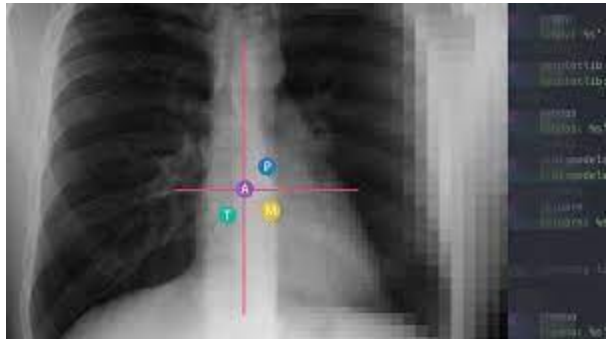
- Outliers are *interesting* because they are *suspected* of *not* being *generated* by the *same* mechanisms as the rest of the data.
- Therefore, in outlier detection, it is important to *justify why the outliers detected are generated by some other mechanisms.*
- This is often achieved by making various assumptions on the rest of the data and showing that the outliers detected violate those assumptions significantly.

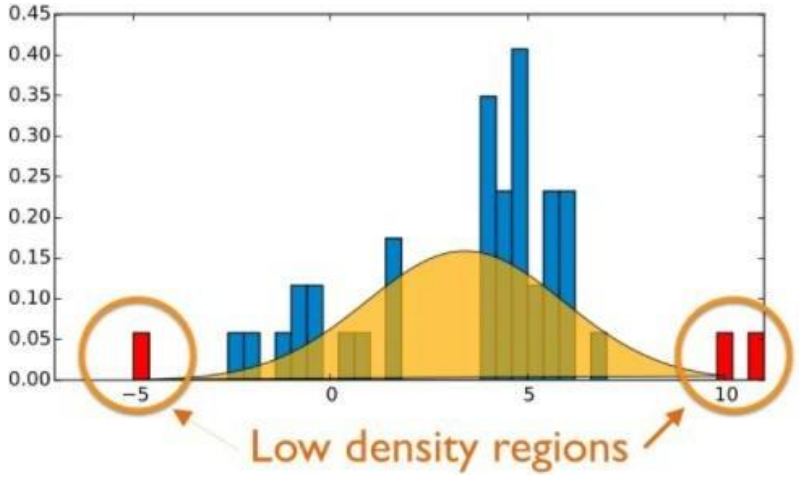
+ Novelty Detection

- Outlier detection is also related to ***novelty detection*** in ***evolving*** data sets.
- For example, by monitoring a ***social media web site*** where ***new*** content is ***incoming***, novelty detection may identify ***new topics*** and ***trends*** in a ***timely*** manner.
- Novel topics may initially appear as ***outliers***.
- To this extent, ***outlier detection and novelty detection*** share some similarity in ***modeling*** and ***detection*** methods.
- However, a ***critical difference between the two is that in novelty detection, once new topics are confirmed, they are usually incorporated into the model of normal behavior so that follow-up instances are not treated as outliers anymore***

+ Outliers/Anomalies







+ Related problems

- Rare Class Mining
- Chance discovery
- Novelty Detection
- Exception Mining

+ Key Challenges

- Defining a ***representative normal region*** is ***challenging***.
- The ***boundary*** between ***normal*** and ***outlying*** behavior is often ***not precise***.
- The ***exact notion*** of an outlier is ***different*** for ***different*** application ***domains***.
- Availability of ***labeled*** data for training/validation
- ***Malicious adversaries***.
- Data might contain ***noise***.
- Normal behavior keeps ***evolving***.

End of Lecture – 02