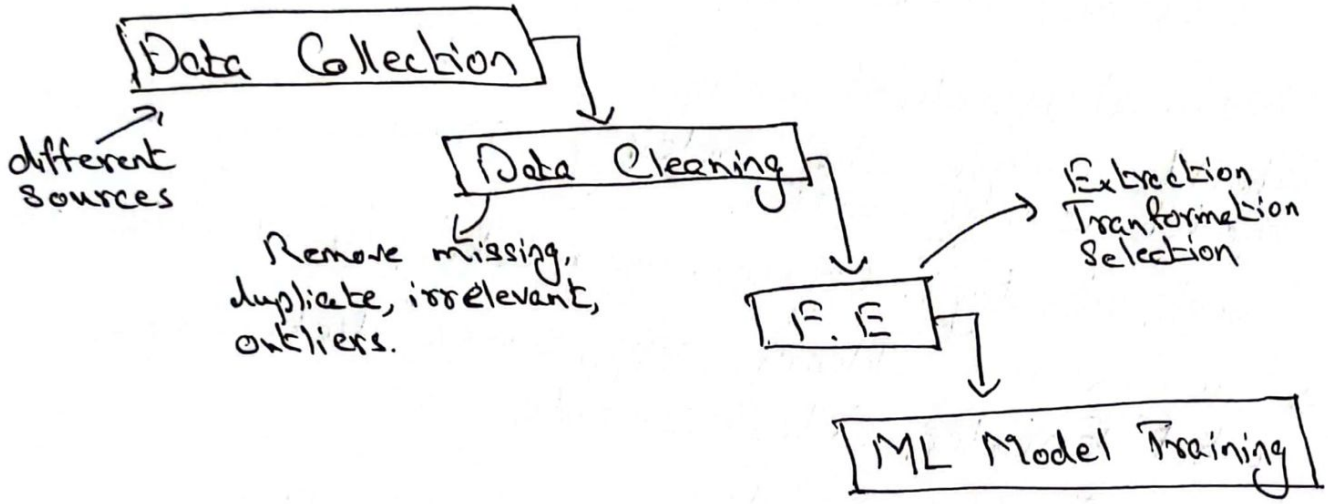
Feature Engineering

Lecture 10

→ Feature Engineering (FE) is transforming clean raw data into features/attributes for suitable ML modelling.

→ FE can be divided into
1) Feature Extraction
2) Feature Transformation
3) Feature Selection



Feature Extraction:-

→ When original Data is very different & can't be used for ML modelling.

→ Method for creating a new & smaller set of features that capture most of the useful raw data.

→ Some popular raw data types for feature extraction are Texts, images, Geospatial data, Web data and multiple sensors data.

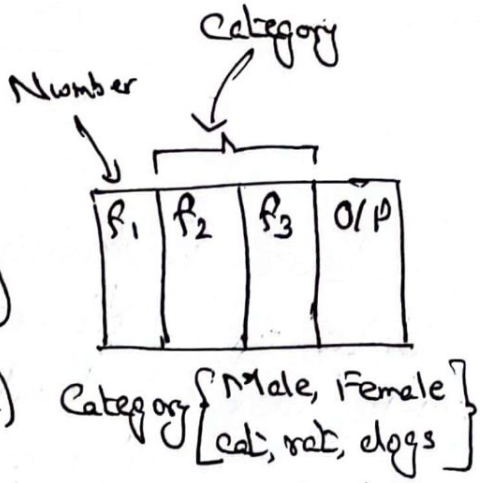
Feature Transformation:-

→ Change the feature from one form to another for improving ML model accuracy.

Feature Encoding:-

→ There are two types

1) Ordinal (Specific Ordered Group)
Order should be maintained.
Education (High School, BS, MS, Phd)
1 2 3 4
Income Level (>50k, 50k~100k, <100k)
1 2 3



2). Binary
(Yes, No), (Male, Female), (True, False).

3) Nominal (Unordered Group)
→ Should arrange the ordering alphabetically so ML model is not biased.
(cat, rat, dog); (pizza, burger, coke); (Univ/Professor Names).

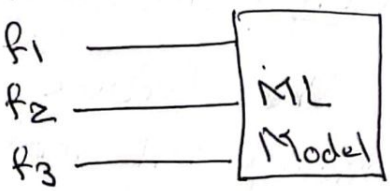
→ You can use Sklearn apply, remove, dictionary, Label Encoder, Ordinal Encoder.

→ Use One Hot Encoding
↳ Every unique value/category is added as a feature. The newly binary variables are called Dummy variables.

City	Peshawar	Lahore	Karachi
Peshawar	1	0	0
Lahore	0	1	0
Karachi	0	0	1

Feature Scaling:-

Suppose



→ So when training the Model will give more weight to f_3 the f_2 & f_1.

→ Have to bring all three features in same range to reduce biasness.

f_1	f_2	f_3
1	100	1000
2	120	2000
3	130	3000
4	140	4000

Fruit	Weight(gm)	Price(Rs)
Orange	100	1
Apple	150	2
Banana	170	4
Mango	200	5

Weight > Price

If Weight in kg

Price > Weight

→ Though all features have same weight.

→ Two Methods

1) Normalization

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

to bring it in range (0-1)

2). Standarization

mean μ

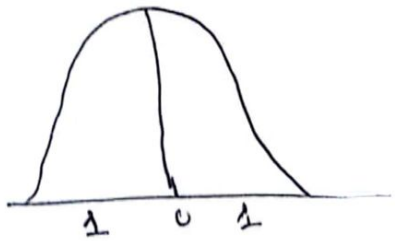
Standard Deviation α

$$X_{stand} = \frac{X - \mu}{\alpha}$$

Values are not restricted to a particular range.

3). Which one to use

→ If your data is normally distributed use standarization otherwise normalization.



→ Good fit is to fit your Model to raw, normalized & standarized data.

→ Proper is to

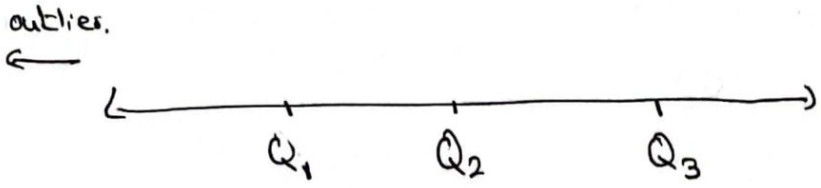
Data → Train split → FE

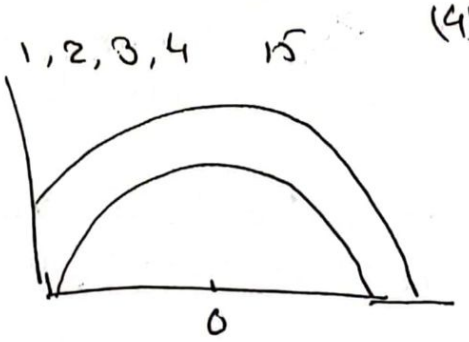
→ You can use Min-Max Scaler, Standard scaler or Robust scaler libraries.

→ Robust Scaler is used to take into account the outliers.

Robust Scaler

$$X_{(scale)} = \frac{X - median}{P_{75} - P_{25}}$$





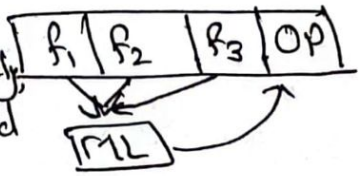
Feature Selection:-

There are three different techniques for FS.

A). Filter Base FS:

→ We use statistical method such as mean, median, variance, skewed dataset and testing (Anova, hypothesis) by choosing each individual feature.

→ Beez they perform analysis individually, they are computationaly light and good for eliminating irrelevant, redundant, constant, duplicated & correlated features.



The following are Filter Based Method types

1) Basic Statistical Filter Method.
   - Variance Threshold (Remove constant & Quasi constant feature)
   - Remove Duplicate feature
2) Correlation & Ranking based statistical Filter Method
   - Pearson's Correlation coefficient (linear data)
   - Spearman's Rank // (linear/nonlinear data)
   - Kendall's Rank // (linear/nonlinear Data)
3) Statistical Test Based Methods
   - Anova or F-test
   - Mutual Information
   - Chi Square

A.1 Variance Threshold

$f_1 = Var[f_1] = 0$ constant (no variance)

$= Var[f_2] = < 0.01 = 1\%$ Quasi constant

$= Var[f_3] = > 1\%$

f_1	f_2	f_3	O/P
0	6	2	A
0	0	4	A
0	0	5	B
0	0	9	C
0	1	10	D

→ If variance is 0 or 1, they are constant and will be removed.

→ Will calculate each feature variance, and comes under unsupervised learning.

→ Disadvantage is that this method doesn't consider correlations and dependent (output) variables.

iRemove Duplicate Feature

f_1	f_2	f_3	O/P
1	3	1	
1	10	1	
1	12	1	
1	15	1	

$f_1 == f_2$; Select 1

A.2 Correlation & Ranking based Filters

→ You should be familiar with Pearson's, Spearman's and Kendall's.

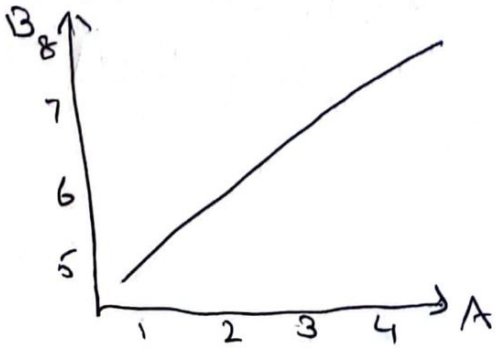
→ Covariance:

+	−	
A	B	C
1	5	10
2	6	9
3	7	8
4	8	7

. Measure the individual relationship between two or more ~~variables~~ features.

. The sign represents if two features change in same (+ve) or opposite (-ve) direction. A '0' represents two features are completely independent.

. It doesn't tell us about by how much (strength or magnitude)



→ Correlation:-

- We can use Pearson's, Spearman's & Kendell.
- For Pearson's data should be normally distributed.

$-1 \sim 1$

Highly uncorrelated — Highly Correlated.

- Remove highly correlated features.

A.3 Statistical Test Based Methods:

1) Anova or F-Test:-

→ A univariate test for testing the individual feature effect on Target.

→ It assumes linear relationship betw feature and target, and feature is normally distributed.

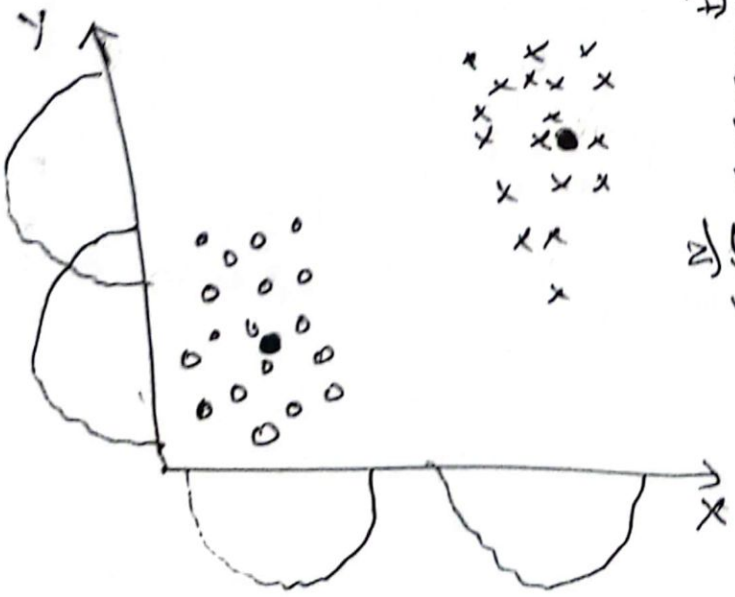
→ For Regression — features (Categorical/Numerical) — Target(N)
Classification — features (C/N) — Target(C)

→ Take one feature & target(output), do Anova test & it will provide F-Score.

$f_{s1} \rightarrow 2$
$f_{s2} \rightarrow 3$
$f_{s3} \rightarrow 30$
$f_{s4} \rightarrow 20$

Choose top 2 or 3 features.

→ F-Score represents which variable discriminates two features better.



1) If we find mean(•) of both with respect to X, the distance is greater then the distance between means found with respect to Y.

2) If we calculate var[x] & var[y] with respect to X & Y, then for ∗

$X \Rightarrow Var[x] < Var[y]$

2). Mutual Information (MI)

→ MI is measure of mutual dependence betw two random features (X & Y).

→ Measures the amount of information obtained about one feature by understanding the other.

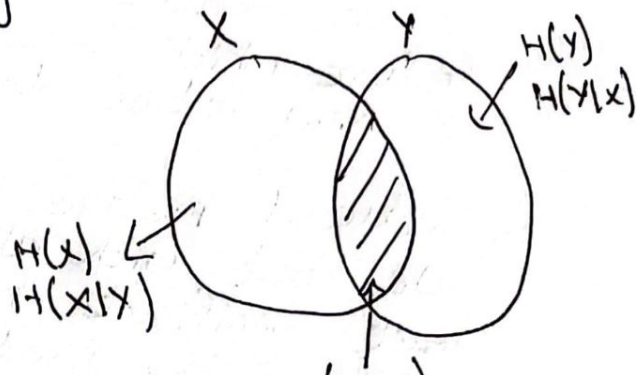
→ In ML, the MI measures how much information the presence/absence of a feature contributes to making the correct prediction on Y.

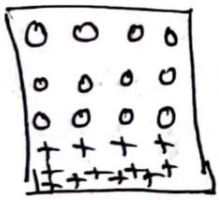
→ It is equal to '0' if and only if two random features are independent, & higher values mean higher dependency.

$$I(X:Y) = H(X) - H(X|Y)$$

$H(X)$ is Entropy of X

$H(X|Y)$ is conditional Entropy for X given Y.



→ 

Entropy → Measure of impurity

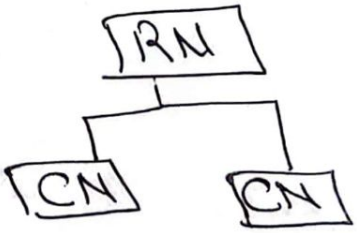
→ " " homogeneity

$$E = -p(\log_2 p) - q(\log_2 q)$$

if $p=0.5$ & $q=0.5$

$E = 1$ else $q=0$; $E = 0$

→ Plz recall DT



- IG tells us how much Entropy was reduced from going from 'RN' to 'CN'.
- In DT, IG calculates the impact of transform to a dataset.
- In MI, it calculates the dependence betw features.

$$IG(S,a) = H(S) - H(S|a)$$

- IG should be as high as possible
- In DT IG(between features), In MI (between feature & O/P)

$f_1 = 0.25$
$f_2 = 0.2$
$f_3 = 0.3$
$f_4 = 0.19$

Wrapper Method :-

→ The main idea is to select which set of features work best for ML model.

→ It follows a greedy search approach by evaluating all possible combination of features.

→ Evaluation matric for Regression (MSE, MAE, MAEP, R^2) and for classification (confusion matrix).

1). Forward Feature Selection
is an iterative method, we keep adding the feature which best imparts our model till addition of new feature doesn't improve the performance.

2). Backward Feature Selection
We start with all the features & remove the least significant feature at each iteration which improves the model performance based on performance parameters.

3). Exhaustive FS
is brute force method. Evaluation of each ~~point~~ possible combination of the variables & returns the best performing subset.

For example

f_1 f_2 f_3 f_4 Y

$[f_1, f_2]$ — [Y] → ML → 20%

$[f_1, f_2, f_4]$ — [Y] → ML → 30%

FFS
iter ① f_1 f_2 f_3 f_4

iter ② f_2, f_1
f_2, f_3
f_2, f_4

iter ③ f_1, f_2, f_3
f_2, f_3, f_4

BFS
iter ① f_1, f_2, f_3 f_4
f_1, f_2, f_3
f_1, f_2, f_4

Embedded Method

→ Wrapper method is computationaly very expensive. Used for datasets with less features. Dataset with many features, we use either filter or embedded method.

→ EM are faster than Wrapper & more accurate than Filter Methods. Less prone to overfitting.

→ Two mechanism for feature selection.

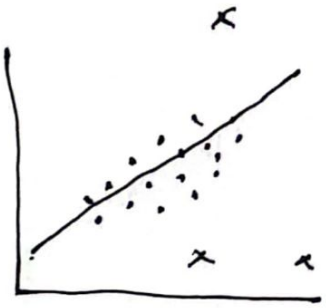
1) Regularization

→ It overcomes the overfitting problem by using penalty.

→ Lasso regression or L1 Regularization

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda * [\text{Slope}]$$

[Penalty]



$$\text{Slope} \sum_{i=1}^{n} (|x_1| + |x_2| + \cdots |x_n|)$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

λ value changes according to β of less significant x values.

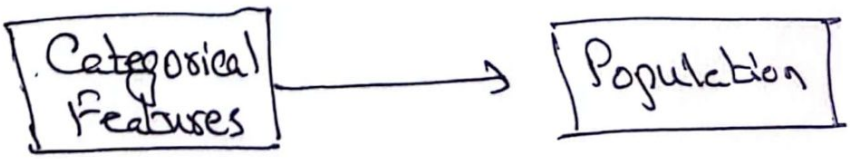
2). Tree-based FS

→ Tree based Algorithm (DT, Random Forest) concept is used.

$[f_1, f_2, f_3]$ $[f_4\ f_5]$

More significant — Less significant

Categorical FS using Chi-Squared



1) Goodness of fit (How are the feature & O/P distributed)

2) Test of Independence
(Relationship between two categorical features)