(DEEMED TO BE UNIVERSITY)

ARTIFICIAL INTELLIGENCE
&
MACHINE LEARNING

SESSION NO : 20

Feature selection methods (mutual information, information gain, Fisher score).

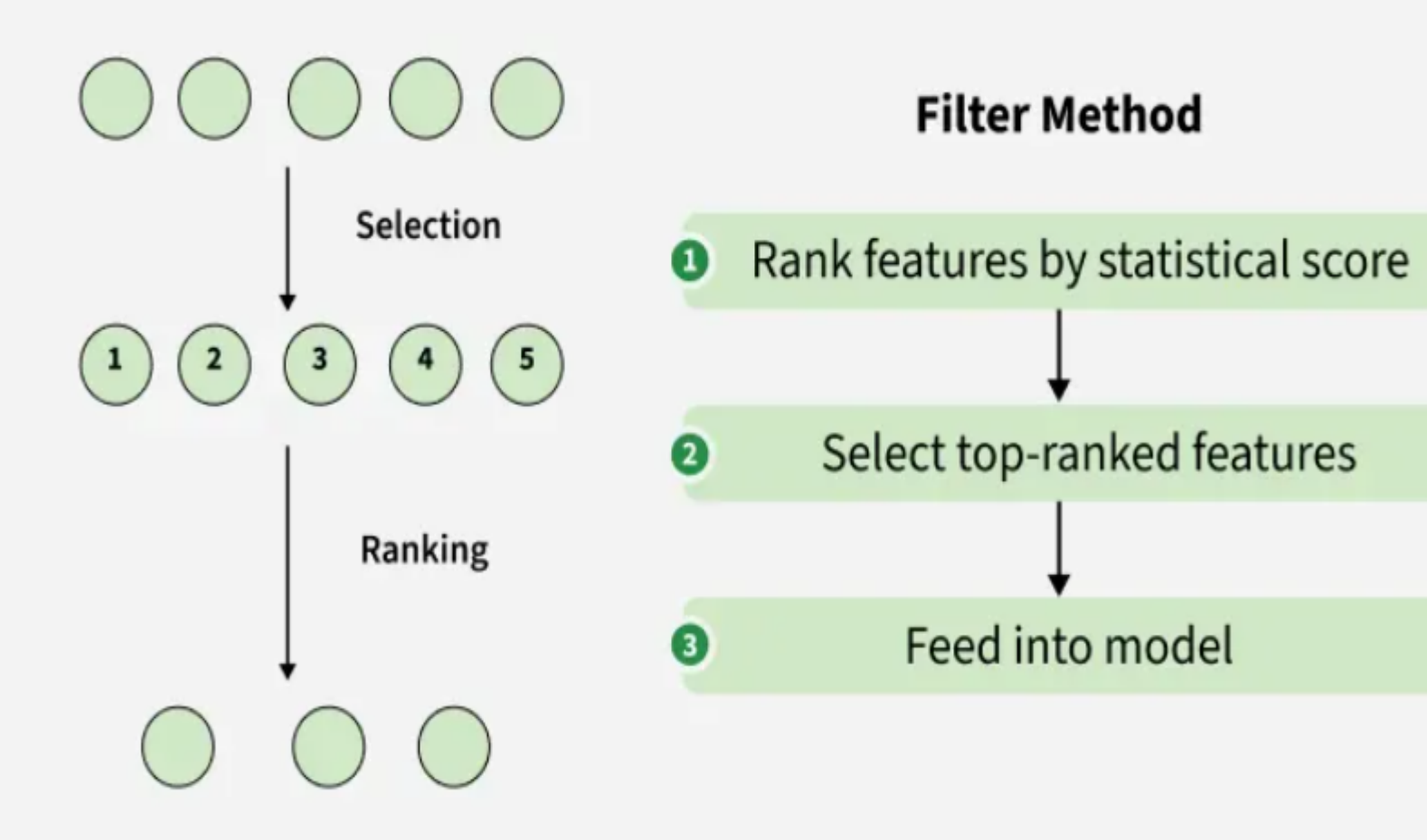
Feature Selection Methods

- In data science many times we encounter vast of features present in a dataset.
- But it is not necessary all features contribute equally in prediction that's where feature selection comes.
- It involves selecting a subset of relevant features from the original feature set to reduce the feature space while improving the model’s performance by reducing computational power.

Feature Selection Methods

- **1. Filter Methods**
- Filter methods evaluate each feature independently with target variable.
- Feature with high correlation with target variable are selected as it means this feature has some relation and can help us in making predictions.
- These methods are used in the preprocessing phase to remove irrelevant or redundant features based on statistical tests (correlation) or other criteria.

Filter Methods



Feature Selection Methods

Some techniques used are:

- **Information Gain**
- **Chi-square test**
- **Fisher's Score**
- **Pearson's Correlation Coefficient**
- **Variance Threshold**
- **Mean Absolute Difference:**
- **Dispersion ratio**

Attribute Selection Measure: Information Gain

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

Fisher Score

- The Fisher Score is a simple used to select important features for classification tasks. It works by comparing how much a feature varies between different classes versus how much it varies within the same class. Features that show big differences between classes, but are consistent within each class, are considered useful for classification.

Binary-class formula (using population variance):

$$F = \frac{(\mu_0 - \mu_1)^2}{\sigma_0^2 + \sigma_1^2}$$

Fisher Score

Sample	Feature_1	Feature_2	Class
S1	1	4	0
S2	2	5	0
S3	2	5	0
S4	3	6	0
S5	7	5	1
S6	8	5	1
S7	8	6	1
S8	9	6	1

Feature 1

Class 0: values = 1, 2, 2, 3

- $\mu_0 = \frac{1+2+2+3}{4} = 2$
- Deviations $(-1, 0, 0, 1)$ → squares $(1, 0, 0, 1)$, sum $= 2$
- $\sigma_0^2 = \frac{2}{4} = 0.5$

Class 1: values = 7, 8, 8, 9

- $\mu_1 = \frac{7+8+8+9}{4} = 8$
- Deviations $(-1, 0, 0, 1)$ → squares $(1, 0, 0, 1)$, sum $= 2$
- $\sigma_1^2 = \frac{2}{4} = 0.5$

Fisher Score

$$F_{\text{Feature\_1}} = \frac{(2-8)^2}{0.5+0.5} = \frac{36}{1} = 36$$

Fisher Score

Sample	Feature_1	Feature_2	Class
S1	1	4	0
S2	2	5	0
S3	2	5	0
S4	3	6	0
S5	7	5	1
S6	8	5	1
S7	8	6	1
S8	9	6	1

Feature 2

Class 0: values = 4, 5, 5, 6

- $\mu_0 = \frac{4+5+5+6}{4} = 5$
- Deviations $(-1, 0, 0, 1)$ → squares sum $= 2$
- $\sigma_0^2 = \frac{2}{4} = 0.5$

Class 1: values = 5, 5, 6, 6

- $\mu_1 = \frac{5+5+6+6}{4} = 5.5$
- Deviations $(-0.5, -0.5, 0.5, 0.5)$ → squares sum $= 1$
- $\sigma_1^2 = \frac{1}{4} = 0.25$

Fisher Score

$$F_{\text{Feature\_2}} = \frac{(5 - 5.5)^2}{0.5 + 0.25} = \frac{0.25}{0.75} = \frac{1}{3} \approx 0.33$$

Fisher Score

- $F_{\text{Feature\_1}} = 36.00$
- $F_{\text{Feature\_2}} \approx 0.33$

  Ranking: Feature_1 » Feature_2 → **Select Feature_1** for modeling.

Fisher Score

For multi classes

$$\text{FisherScore}(x_j) = \frac{\sum_{c=1}^{C} n_c \left(\mu_j^{(c)} - \mu_j\right)^2}{\sum_{c=1}^{C} n_c \left(\sigma_j^{(c)}\right)^2}$$

Where:

- n_c: Number of samples in class c
- μ_j: Mean of feature j over the entire dataset
- $\mu_j^{(c)}$: Mean of feature j in class c
- $\sigma_j^{(c)}$: Standard deviation of feature j in class c

Mutual Information

- Mutual information(MI)between two random variables is a non-negative value,which measures the dependency between the variables .It is equal to zero if and only if two random variables are independent ,and higher values mean higher dependency
- In short, it is the amount of information one variable gives about the other.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Mutual Information

ID	F1	F2	YY
1	High	A	Pass
2	High	A	Pass
3	High	B	Pass
4	High	B	Fail
5	Low	A	Fail
6	Low	A	Fail
7	Low	B	Fail
8	Low	B	Pass

Totals: $n = 8$.

Marginals: $P(X_1 = \text{High}) = 0.5, P(X_1 = \text{Low}) = 0.5, P(Y = \text{Pass}) = 0.5,$

$P(Y = \text{Fail}) = 0.5.$

Mutual Information

ID	F1	F2	YY
1	High	A	Pass
2	High	A	Pass
3	High	B	Pass
4	High	B	Fail
5	Low	A	Fail
6	Low	A	Fail
7	Low	B	Fail
8	Low	B	Pass

MI of X1 with YY

Joint probabilities:

- $P(\text{High}, \text{Pass}) = 3/8 = 0.375$
- $P(\text{High}, \text{Fail}) = 1/8 = 0.125$
- $P(\text{Low}, \text{Pass}) = 1/8 = 0.125$
- $P(\text{Low}, \text{Fail}) = 3/8 = 0.375$

Use

$$I(X_1; Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \quad \text{with } P(x)P(y) = 0.25.$$

Totals: $n = 8$.

Marginals: $P(X_1 = \text{High}) = 0.5$, $P(X_1 = \text{Low}) = 0.5$, $P(Y = \text{Pass}) = 0.5$, $P(Y = \text{Fail}) = 0.5$.

Mutual Information

ID	F1	F2	YY
1	High	A	Pass
2	High	A	Pass
3	High	B	Pass
4	High	B	Fail
5	Low	A	Fail
6	Low	A	Fail
7	Low	B	Fail
8	Low	B	Pass

MI of X1 with YY

Ratios and terms:

- $0.375/0.25 = 1.5, \log_2(1.5) \approx 0.585 \rightarrow 0.375 \times 0.585 \approx 0.219$
- $0.125/0.25 = 0.5, \log_2(0.5) = -1 \rightarrow 0.125 \times (-1) = -0.125$
- same two values repeat → sum $= 0.219 - 0.125 - 0.125 + 0.219 = 0.18875$

Result: $I(X_1; Y) \approx \mathbf{0.189}$ bits.

Totals: $n = 8$.

Marginals: $P(X_1 = \text{High}) = 0.5$, $P(X_1 = \text{Low}) = 0.5$, $P(Y = \text{Pass}) = 0.5$, $P(Y = \text{Fail}) = 0.5$.

Mutual Information

ID	F1	F2	YY
1	High	A	Pass
2	High	A	Pass
3	High	B	Pass
4	High	B	Fail
5	Low	A	Fail
6	Low	A	Fail
7	Low	B	Fail
8	Low	B	Pass

MI of X2 with YY

By construction, X_2 is independent of Y:

- $P(A) = 0.5, P(B) = 0.5$
- Joint counts: A with Pass $= 2$, A with Fail $= 2$; B with Pass $= 2$, B with F

  → Every $P(x_2, y) = 0.25 = P(x_2)P(y)$.

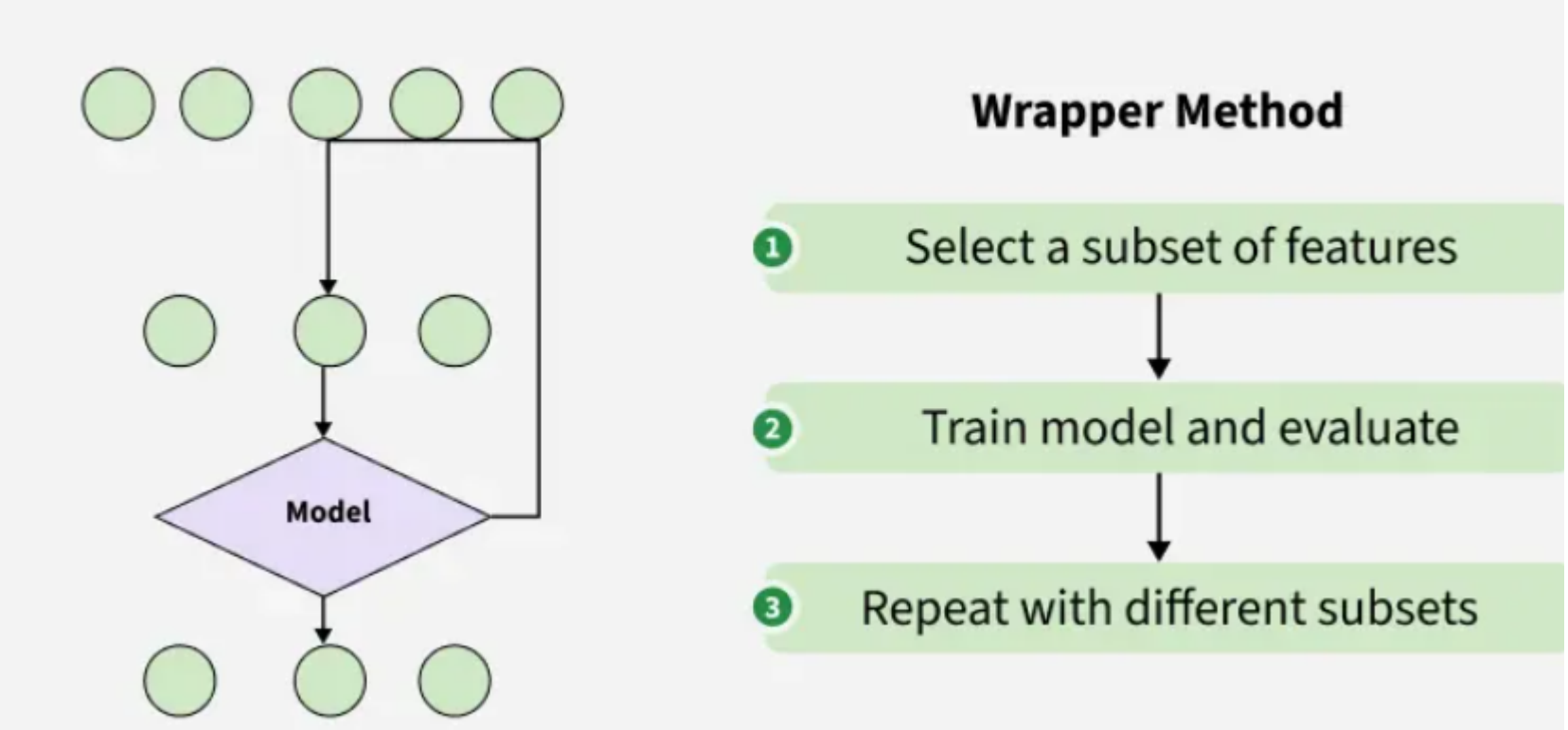
Thus each term has $\log_2 \frac{0.25}{0.25} = 0 \rightarrow I(X_2; Y) = \mathbf{0}$.

Feature Selection Decision (by MI ranking)

- $I(X_1; Y) \approx 0.189$ bits
- $I(X_2; Y) = 0$ bits

Wrapper methods

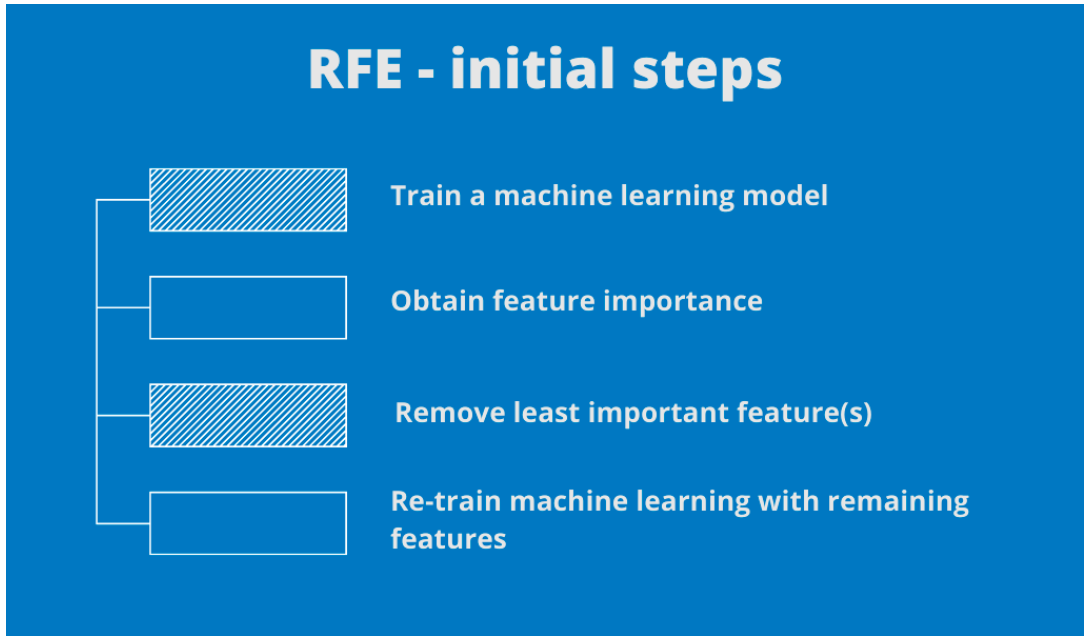
- Wrapper methods are also referred as greedy algorithms that train algorithm. They use different combination of features and compute relation between these subset features and target variable and based on conclusion addition and removal of features are done.



Wrapper methods

- **Forward selection**
- **Backward elimination**
- **Recursive elimination**

Forward selection	Backward elimination	
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $=> \{A_1\}$ $=> \{A_1, A_4\}$ $=>$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $=> \{A_1, A_3, A_4, A_5, A_6\}$ $=> \{A_1, A_4, A_5, A_6\}$ $=>$ Reduced attribute set: $\{A_1, A_4, A_6\}$	I {



(DEEMED TO BE UNIVERSITY)

THANK YOU