Key drivers of feature Selection → Feature Relevance
→ Feature redundancy

Feature Relevance:

In Supervised learning Based on training data which is already labelled a model is inducted, now this model is capable of assigning class labels to new unlabelled data based on some predictor variables.

predictor variable
- $\xrightarrow{\text{not contributing any info}}$ Irrelevant variables
- $\xrightarrow{\text{little info}}$ Weakly relevant variables
- $\xrightarrow{\text{Significant contri-bution in info}}$ Strongly relevant variables

In unsupervised learning $\boxed{\text{training data labelled dat}}$ $\xrightarrow{(\times)}$ Not there

So grouping of Similar data instances are done based on differen variables.

- Some variables don't contribute any info for similarity or dissimilarity in data instances (or) in grouping process these are irrelevant variables in the context of unsupervised learning.

eg:- in the context of the supervised task of predicting Student weight or Unsupervised task of grouping students with similar academic merit, the Variable Rollno is quite irrelevant

- So while selecting a subset of features irrelevant features are rejected. rejection of weakly is based on case to case.

Feature Redundancy:

- A feature may contribute info which is similar to the info contributed by one or more features

eg:- In weight prediction problem features age, height Contribute Similar info

- so if either age or height is not part of feature subset results are almost same

These are said to be potentially redundant

How to find out irrelevant, potential redundant features?

Measures of feature relevancy & redundancy

Measures of feature relevance:

Mutual information is considered as good measure to determine the info contribution of a feature to decide the class label.

Mutal Info ↑ relevancy ↑

for supervised learning:

k = no. of classes
C = class variable
f = feature set

$$MI(C,f) = \underbrace{H(C)}_{\text{marginal entropy of a class}} + \underbrace{H(f)}_{\text{marginal entropy of feature 'x'}} - H(C,f)$$

$$-\sum_{i=1}^{k} P(c_i)\log_2 P(c_i)$$

$$-\sum_{c} P(f=x)\log_2 P(f=x)$$

for unsupervised learning:

Since there is no class variable

Entropy of the set of features without one feature at a time is calculated for all features using shannon's formula

$$H(f) = -\sum_{x} p(f=x)\log_2 P(f=x)$$

and are maintained in descending order

Measures of feature redundancy: (to measure the similarity in info contribution)

1) Correlation based measures
2) Distance based measures
3) Other coefficient based measures

① correlation is the measure of linear dependency b/w two random variables /features

Pearson's correlation Coefficient: $\rho = \dfrac{cov(F_1, F_2)}{\sqrt{Var(F_1) \cdot Var(F_2)}}$

$$cov(F_1, F_2) = \sum (F_{1i} - \bar{F_1}) \cdot (F_{2i} - \bar{F_2})$$

$$Var(F_1) = \sum (F_{1i} - \bar{F_1})^2 \text{, where } \bar{F_1} = \frac{1}{n}\sum F_{1i}$$

$$Var(F_2) = \sum (F_{2i} - \bar{F_2})^2 \text{, where } \bar{F_2} = \frac{1}{n}\sum F_{2i}$$

→ $-1 < \rho < 1$

→ 1 indicates perfect correlation

→ 0 indicates no relationship

→ Usually, A Threshold value is adopted to decide similarity

2

Euclidian Distance:

$$d(F_1, F_2) = \sqrt{\sum_{i=1}^{n} (F_{1i} - F_{2i})^2}$$

Apti (F_1)	communication (F_2)
2	6
3	5.5
6	4
7	2.5
8	3
6	5.5
6	7
7	6
8	6
9	7

Sample table

Minkowski Distance:

$$d(F_1, F_2) = \sqrt[r]{\sum_{i=1}^{n} (F_{1i} - F_{2i})^r}$$

with $r = 2$ (Euclidian Dis)

$\longrightarrow$ L_2 norm

Manhattan Distance

$$d(F_1, F_2) = \sum_{i=1}^{n} |F_{1i} - F_{2i}|$$

with $r = 1$ (L_1 norm)

3 Jaccard Index / coefficient:

$$j = \frac{n_{11}}{n_{01} + n_{10} + n_{11}}$$

$\Rightarrow$ Measures similarity b/w two features

Jaccard distance:

$$d_j = 1 - j$$

$\Rightarrow$ Measures dissimilarity b/w two features. it's Complement of Jaccard ind

eg:-

F1	0	1	1	0	1	0	1	0
F2	1	1	0	0	1	0	0	0

$$J = \frac{2}{1+2+2} = \frac{2}{5} = 0.4$$

$$d_J = 1 - 0.4 = 0.6$$

i.e F_1, F_2 are 40% similar 60% dissimilar

Simple matching coefficient (SMC):

$$SMC = \frac{n_{11} + n_{00}}{\underbrace{n_{00} + n_{01} + n_{10} + n_{11}}_{\text{all combinations}}} \Rightarrow$$

Measures similarity b/w 2 features, includes cases where both features having value 0.

$$SMC = \frac{2+3}{3+1+2+2} = \frac{5}{8} = 0.625$$

$$d_{SMC} = 1 - 0.625 = 0.375$$

Cosine similarity:

Most poupular measure in text classification

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

$x \cdot y$ → dot prod

(x, y are vectors)

$\|x\| \cdot \|y\|$ → magnitudes

i.e $x \cdot y = \sum_{i=1}^{n} x_i \cdot y_i$

$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$ and $\|y\| = \sqrt{\sum_{i=1}^{n} y_i^2}$

eg:-

$x = (2,4,0,0,2,1,3,0,0)$ and $y = (2,1,0,0,3,2,1,0,1)$

$x \cdot y = 2*2 + 4*1 + 0*0 + 0*0 + 2*3 + 1*2 + 3*1 +$
$0*0 + 0*1 = 19$

$\|x\| = \sqrt{2^2+4^2+0^2+0^2+2^2+1^2+3^2+0^2+0^2} = \sqrt{34} = 5.83$

$\|y\| = \sqrt{2^2+1^2+0^2+0^2+3^2+2^2+1^2+0^2+1^2} = \sqrt{20} = 4.47$

$$\cos(x,y) = \frac{19}{5.83 \times 4.47} = 0.729$$

- Cosine Similarity measures angle b/w x, y vectors

$\cos(x,y) = 1$ indicates angle $= 0°$ i.e x,y are Same except magnitude

$\cos(x,y) = 0$ indicates angle $= 90°$ i.e x,y don't Share similarities

in prev eg:- $\cos^{-1} 0.729 = 43.2°$