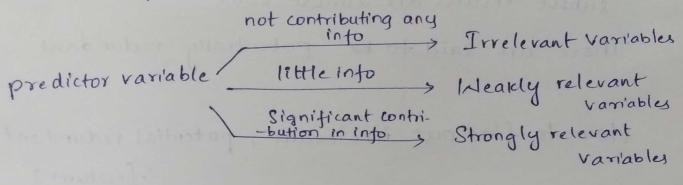
Key drivers of feature Selection Feature Relevance
Feature redundancy

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.



In unsupervised learning tra

training data (X)
labelled dat Not there

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

- Some variables don't contribute any into 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

"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 into which is similar to the into contributed by one or more features

eg: In weight prediction problem features agesheight 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 relevance:

Measures of feature relevance:

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

for supervised learning:

MI(C,f) = H(c) + H(f) - H(c,f)

marginal entropy of

a class

feature 'x'

- Sp(ci) log p(ci)

i=1

Mutal Info 1 relevancy 1

k = no of classe

C = class variable

f = feature set

marginal entropy of

feature 'x'

- Sp(f=x) log 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

$$H(f) = - \sum_{x} p(f=x) \log_{x}^{p} (f=x)$$

and are maintained in descending order

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

- i) Correlation based measures
- e) Distance based measures
- 3) Other coefficient based measures
- Ocorrelation is the measure of linear dependency blue two mandom variables I features

Pearsons correlation Coefficient:
$$\rho = \frac{\text{cov}(F_1, F_2)}{\sqrt{\text{Var}(F_1) \cdot \text{Var}(F_2)}}$$

var (Fi) =
$$\sum (F_{ii} - F_{i})^{2}$$
, where $F_{i} = \frac{1}{n} \sum F_{ii}$

$$Var(F_2) = \sum (F_{2i} - \overline{F_2})^2$$
, when $\overline{F_2} = \frac{1}{h} \sum F_{2i}$

- -> 1 indicates perfect correlation
- -> o indicates no relationship
- -> Usually, A Threshold Value is adopted to decide Similarity

(2)

Euclidian Distance:

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

Minkowski Distance:

Manhattan Distance

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

with r=1 (LI norm)

Jaccard distance:

F1 0 1 1 0 1 0 1 0 0 0 0

Apti	Communi
(FI)	Communication (F2)
2 3	6
6	5-5
8	2.5
6	3 5.5 7
7 8	6
9	7
Samuel Lall	

Sample table

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

i.e FIIF2 are 40%. similar 60%. dissimilar

Simple matching coefficient (SMC):

$$SMC = \frac{n_{11} + n_{00}}{n_{00} + n_{01} + n_{10} + n_{11}}$$
all combinations

Measures Similarity blw
2 features, includes cases
where both features
having value 0.

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

Cosine Similarity:

Most poupular measure in text classification

n=(2,4,0,0,2,11,3,0,0) and y=(2,1,0,0,3,2,1,0,

91. y = 2 d 2 + 4 d 1 + 0 = 0 + 0 d 0 + 2 - 3 + 1 = 2 + 3 = 1 + 0 = 0 + 0 d 1 = 19

 $||x|| = \sqrt{2^2 + 4^2 + 0^2 + 0^2 + 2^2 + 1^2 + 3^2 + 0^2 + 0^2} = \sqrt{34} = 5.83$ $||y|| = \sqrt{2^2 + 1^2 + 0^2 + 0^2 + 3^2 + 2^2 + 1^2 + 0^2 + 1^2} = \sqrt{20} = 4.42$

cos(2,4) = 19 = 0.729

· Cosine Similarity measures angle blux, y vectors

cos(x1y)=1 indicates angle=0° i-e x1y are

Same except magnitude

Cos(niy)=0 indicates angle=90 i-e ny done Share similarities

in pru eg: cost 0.729=43.20

