Feature Selection

Assume that we want to distinguish between Indians and Americans, based on their features like height, weight, skin color, eye color, hair color, education, etc. We can say that skin-color is an important feature that can distinguish between the two communities.

Our aim here is to develop an algorithm can identify the above mentioned task and also identify other features that might also be important.

The feature selection method should result in automatically must detect important features.

Feature selection problem formulation

Feature $X_1, X_2, X_3 \ldots X_D$

b no of features to be selected $b < D$
(value of b varies from problem to problem)

b — known or unknown

From the point of view of computational complexity / Constraints (like the amount of computations to be done) or Co-variance matrix cannot be more than a specified by.

Sometime the user may give a range of no of features for eg give me features between (15 - 20)

Uses of FS.

1. Redundant features act as noise. Noise removal

$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ Assume that $X_2 = 2X_1 - 1$ i.e

X_1	X_2
1	1
3	5
-10	-21

There is a relationship between X_1 & X_2 then
2 do not require both features. Hence X_1 & X_2 are called redundant features. X_1 & X_2 are two way relationship hence I can say X_1 or X_2 are redundant

Non linear relationship

$$X_2 = 2X_1^3 - 10X_1^2 + 5X_1 - 7$$

exponential or any other relationship.

→ we cannot Establish the relationship b/n X_2 & X_1

$$X_2 = 10^{2^{-X_1}} - 3e^{X_1} + 4\log_2 X_1$$

2. Insight into classification problem.

Steps for feature Selection.

1. Objective function J which attaches a value to Every subset of features is to be defined, i.e we need an objective function that (provides/measures) the importance of the collection of features.
2. optimize / maximize depending on the function.

Here let's assume that the objective function is defined, & b is known.

D = 100 features

b = 10

How many ~~such~~ possible such sets containing 10 elements can be found from D?

$\binom{100}{10}$ This many subsets of size 10 we can have from 100 features

$\binom{100}{10} > 10^{12}$

Can we evaluate all these subsets to find the optimal subset? No Since 10^{12} is a huge no. Suppose I do not search the whole space, then Can we gaurrente that I will get the optimal Solution.

There is no feature selection algorithm that gives optimal feature subset for any criterion function without doing an exhaustive search.

Any criterion function means that satisfy some properties we can exploit these properties to obtain a feature selection algorithm which will give optimal feature subset without doing an exhaustive search.

TM COVER & CAMPENHOUT - IEEE Tr. Information theory 1973

"The two best features are not necessarily the best two"

Assume criterion fn J to maximize and 4 variables x_1, x_2, x_3, x_4

$$J(\{x_1\}) \geq J(\{x_2\}) \geq J(\{x_3\}) \geq J(\{x_4\})$$

If we want to select 2 features x, will the best one to be picked up

$^4C_2 = 6$ pairs

$J(\{x_3, x_4\}) \geq$ the other 5 pairs

$\Downarrow$

This pair is better than any other pair

This happens because at there exist a two way relationship between x_1 & x_2 then if we put them together there is no extra information we can derive.

$S = \{x_1, x_2 \ldots x_D\}$ $b < D$

We need to define a function J

$P(S)$ is a power set of S (set of all subsets)

$P(S) = 2^D$ element will be there

$= \{B : B \subseteq S\}$

The objective function J must be defined from $P(S)$ to $(-\infty, \infty)$ i.e the domain of J is $P(S)$ a range is $(-\infty, \infty)$

$J : P(S) \rightarrow (-\infty, \infty)$ general form may be $(0, \infty)$

and J needs to be optimized!

$A_b = \{B \; B_s \subseteq S, B$ contains b elements$\}$

i.e we are going to look at all subsets of S containing b elements