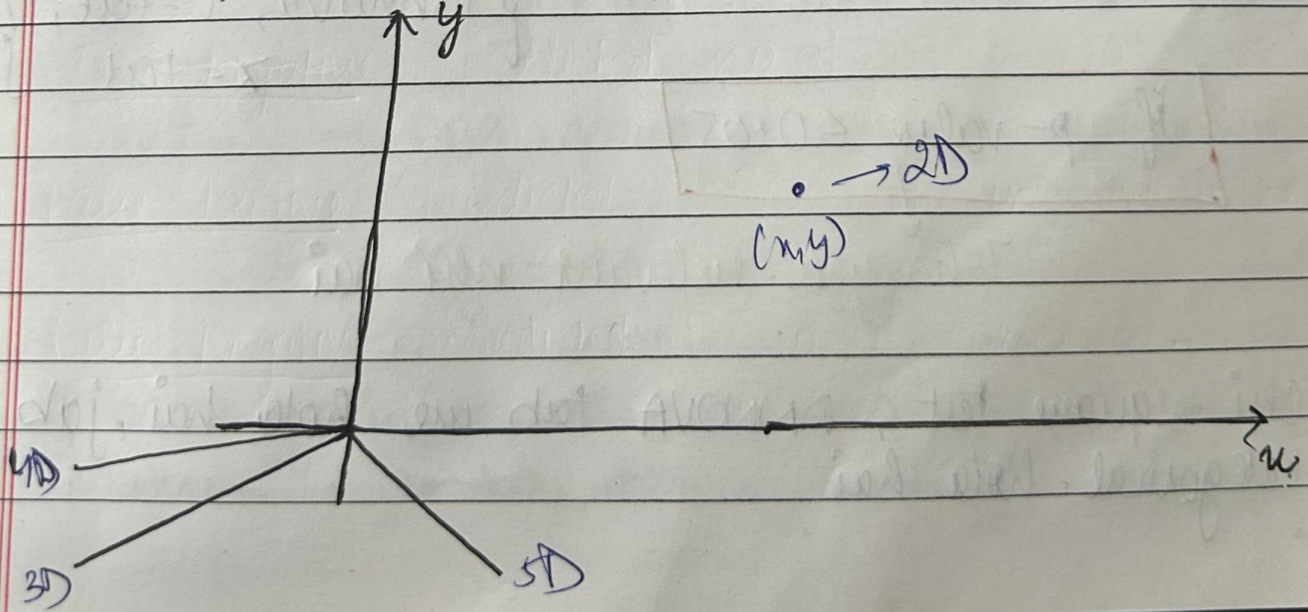
9/2/23

FEATURE SELECTION TECHNIQUES IN MACHINE LEARNING



Date / /
Page No.

⇒ If we're having 100 features, then 100 dimensions space.

⇒ Features which are really important, that we have to select.

⇒ Feature Selection tools → helps in selecting relevant features.

⇒ Unsupervised Techniques - can be used for unlabeled data.

⇒ FILTER METHOD

(Select K Best)

x_1	x_2	x_3	x_4	y

⇒ x_1 ki value y ke saath
x_2 y
x_3 y
x_4 y

} We'll get probability values (chi-square test) using { Anova, t-test, z-test }

If p-value ≤ 0.05

Toh yeh relevant nhi hai

⇒ Chi-square test, ANNOVA tab use hota hai, jab data categorical hota hai

- Filter methods pick up the intrinsic properties of the features measured via univariate statistics instead of cross-validation performance. These methods are faster & less computationally expensive than wrapper methods, when dealing with high-dimensional data, it is computationally cheaper to use filter methods.

- uses chi-square test
- recommended for large no. of features
  [Eg) 100, 200 etc, not for less no. of features.]

⇒ Correlation is the measure of the linear relationship b/w 2 or more variables
(we get this from dataframe) {.corr()}

⇒ Mostly in industries, < 0.7 (not imp)
> 0.7 (imp)

⇒ **WRAPPER METHOD**

x_1	x_2	x_3	x_4	y

⇒ x_1 & y → 90% accuracy
↳ $x_1 + x_2$ & y → 92% accuracy

If accuracy is inc, then we keep adding features

Now, $x_1 + x_2 + x_3$

↓

If accuracy ↑ (take the feature)
If accuracy does not increase
(that new feature added is not imp)

This process is called Forward Selection
(an iterative method)

Backward Feature Elimination

x_1	x_2	x_3	y

$x_1 + x_2 + x_3$ & y → accuracy = 92%

Now remove the features one by one

$x_1 + x_2$

↓

If accuracy is decreasing → then removed one is imp
increasing → then remove the last feature
{ $x_1 + x_2 +$ [x_3] → remove }

⇒ RECURSIVE FEATURE ELIMINATION

It assigns weights to the features.