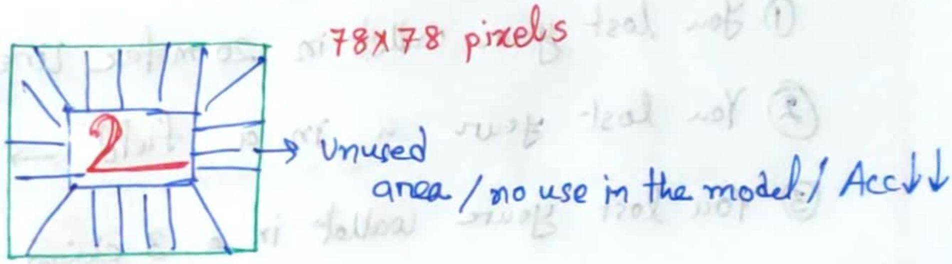
Machine Learning (CampusX Part)

Principle Component Analysis

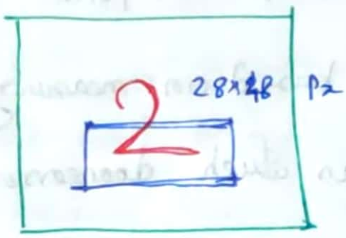
How can you explain Curse of Dimensionality easily?

In a machine learning model there is an optimal number of features/ dimensions you can have which provides the best accuracy. If you use less number or more number of features the accuracy will start decreasing.

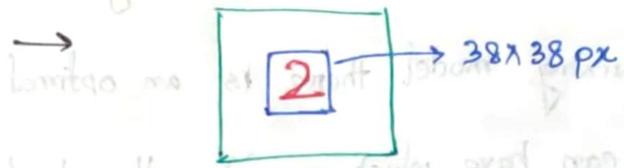
Example: Suppose in a digit handwritting classifier, a sample image is →



If we take all the pixels as a column / feature we will have 784 features which will take longer amount of time for processing, distance calculations and will be complex in nature. That will make our model accuracy lesser.



If we take the smaller part, that will also reduce the model accuracy as it is not ~~the~~ covers the whole area. So, we have to come to an optimal approach where only the digit area will be selected to train the model



"High demension data" also creates "sparsity" which means the data points are really very far from each other.

Here is an wallet example to understand that

① You lost your wallet in 20 meter line → can find easily → 1D

② You lost your " in a field → a little time to find. → 2D

③ You lost your wallet in a 3 storied building → a lot of time will be needed to find → 3D

The higher the dimension – the longer the distance – the longer the time complexity to find a data point

So, the algorithms which are based on measuring distance, they have to calculate for longer distances which decreases accuracy.

To solve such problems and increase acc we can use dimensionality reduction techniques.

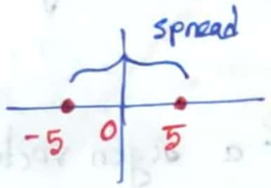
Dimensionality Reduction

Feature Selection
- Forward Selection
- Backward Elemination

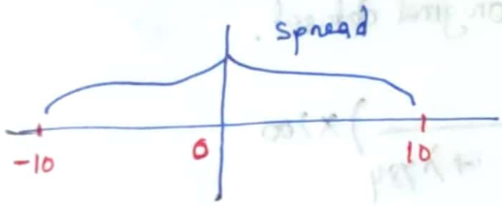
Feature Extraction
- PCA

Why Variance is important?

Variance is proportional to spread.



$$\text{variance} = \frac{(-5)^2 + 0^2 + 5^2}{3} = \frac{50}{3}$$



$$\text{variance} = \frac{(-10)^2 + 0 + 10^2}{3} = \frac{200}{3}$$

So spread and variance are proportional to each other, if the spread of the data increased, variance will also increase and vise versa.
But they are not same.

But you can't take spread in terms of measure. Because spread actually measured by ~~mod~~ MAD (Mean Absolute Deviation) which comes with a mod ~~function~~ $\frac{|x_i - \bar{x}|}{n}$ that can't divided by zero. So, can't apply optimization algorithms to that because that is not differentiable.

That's why variance is preferable $\rightarrow \sum_{i=1}^{n} \frac{(x_i - \bar{x}_i)^2}{n}$ which is differentiable.

When you convert the data from having higher dimension to be in lower dimensions, to keep the essence and relationship of datapoints of each other we need variance to be maximized.

How to find optimal number of principle components?

eigen value = λ, this eigen value tells what amount of variance it's eigen vector covers from the original data.

In a 784 dimension after fitting data we will get 784 eigen values like $\lambda_1, \lambda_2, \cdots\cdots \lambda_{784}$

From them we can calculate the percentage of a eigen vector, the percentage of holding variance of the original dataset.

Suppose for $\lambda_1 = \left(\dfrac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_{784}} \right) \times 100$

Total we have to explain about 90% of variance.

So, we will take the set of λ who together make 90%

Like suppose, $\lambda_1 = 30\%$, $\lambda_2 = 21\%$, $\lambda_3 = 15\%$, $\lambda_4 = 5\%$, $\lambda_5 = 9\%$, $\lambda_6 = 10\%$

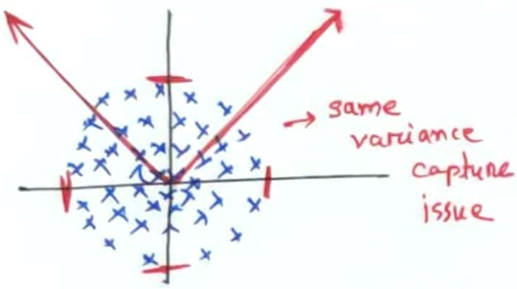
Here, taking 6 components we can come to 90%

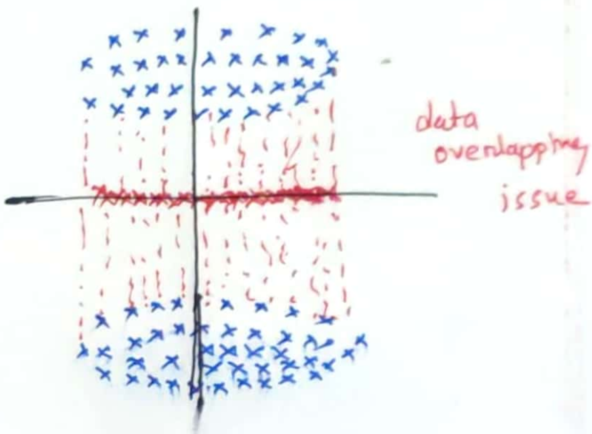
So, we will take the number of λ components which together can cross 90% of variance

When PCA doesn't Work:

If your data distribution is like this that after applying principle components in higher dimensions also every principle components holding same variance that time PCA won't work because we reduce dimensions who have a less variance capture.

Here are some situations where PCA won't work.





Then any kind of pattern data, if we reduce the dimensionality, we will lose the pattern.