UNIT-III

Building good training data sets

* **Dataset** :- A set of multiple features which are related to each other which provided meaningful information.

Record →

Data level element → (PQR)

SNO	SName	Ssec	Smarks
1	XYZ	A	20
2	PQR	A	25
3	STU	B	
4	ABC	B	30

– Attributes / Features.

Mid 4a)

* **Dealing with missing data** :-

* It is used to find the missing value or data.
* One way of handling missing values is the deletion of the rows or columns having null values.
* If any columns have more than half of the values as null then you can drop the entire column.
* In the same way, row can also be dropped if having one or more columns values as null.
* Naive Bayes method can tolerate missing data this method models the class-specific distribution of each predictor.

* There are 4 Techniques to deal with missing data.
  - Delete the data.
  - Imputing Averages
  - Assign New Category
  - Certain Algorithms.

- Deleting / Dropping the data :-
  * This is commonly used to handle null values.
  * It is easy to implement and there is no manip-ulation of data required.
  * If the data set information is valuable or training dataset has less no.of records then deleting rows might have negative impact on the analysis.
    1. deleting rows
    2. deleting columns
    3. pairwise deletion

Imputing missing values :-
* There exists many approach to missing-data imputation and they usually depend on your problem and how your data algorithm behaves.
* Missing data in 2 types 1. Time-Series problem
  2. General problem.

Time Series problem :-
Time series data sets may contain trends and seasonality. It is influenced by seasonal factors.
— Define in three categories :
* data without trend and without Seasonality.
* data with trend and without Seasonality
* data with trend and with Seasonality.

2. General problem :- Method of handling missing values between two data type such as continuous data and categorical data are different.

→ Mean and Median (Categories) :-

* This the most common method of imputing missing values of numeric columns. Median is the middlemost value.

Mode :-

* It is the most frequently occurring value.
* It is used in the case of categorical features.

Certain Algorithms :-

* KNN is a machine learning algorithm which works on the principle of distance measure.
* This algorithm can be used when there are nulls present in the dataset.
* Another algorithm which can be used is Random Forest. This model produces a robust result because it works on well on non-linear and the categorical data.

Mid Sub

* Handling Categorical Data :-

* After handling missing values in the dataset the next step was to handle categorical data.
* All machine learning models are some kind of models that needs numbers to work with.
* Types of categorical data :

1. Nominal Data
2. Ordinal Data.

1. Nominal Data :-

The nominal data called labelled / named data. Allowed to change the order of categories, Change in order doesn't affect its value.

2. Ordinal Data :-

Represent discretely and ordered units. Same as nominal data but have ordered / rank. Not allowed to change the order of Categories.

3 One hot Encoding :-

* One-Hot Encoding is a very handy and popular technique for treating categorical features.
* This based on Creating additional features by its unique value.
* Every unique value in this is added features and values are assigned as 1 or 0 based on presence.

Advantages :-

* Easy to use.
* Create no bias as assumption of any Ordering between the Categories.

Disadvantages :-

* Can result in, an increase in no. of features resulting in performance issues.

- Ordinal number Encoding :-

* Each label is assigned a unique integer based on alphabetical ordering.
* This is the easiest way and used in most of the data where there is natural relation between the Categories of ordinal values.

Advantages :-

1. Easy to use
2. Easily reversible.
3. Doesn't increase feature Space.

Disadvantages :-

1. May result in unexpected results if the ordering of number is not related in any order.

- Count or Frequency Encoding :-

* In this type of encoding the count of existence of each category in the variable is determined.
* Each category is then replaced by the frequency of it.

Advantages :-

1. Easy to implement
2. There will be no increase in ~~feature~~ feature Space.
3. work well with tree-based algorithms.

Disadvantages :-

1. It will not provide the same weight if the frequencies are the same.

- Ordinal Encoding as per Target :-

* Features are replaced as per the target feature.
* As per the Soating order of maximum tree.
* Ordering the categories as per target.

Advantages :

1. It makes a monotonic relationship with target.

Disadvantages :

1. May Cause overfitting

- Mean Encoding :-
  * Categories are assigned mean value as per the target value.
  * Each category mean is calculated as per target and the same value is assigned.

Advantages :-
1. It makes a monotonic relationship with target.
2. Doesn't affect the volume of the data and helps in learning faster.

Disadvantages :-
1. Model may overfit
2. Hard to validate the results
3. Fewer Splits, faster learning.

* Bringing features onto the same scale :-

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descents are updated at the same rate for all features, we scale the data before feeding it to the model.

Having features on similar scale can help the gradient descent converge more quickly towards the minima

Techniques to perform feature scaling :-

- Min-Max Normalization :-

This technique re-scales a feature or observation value with distribution value between 0 and 1.

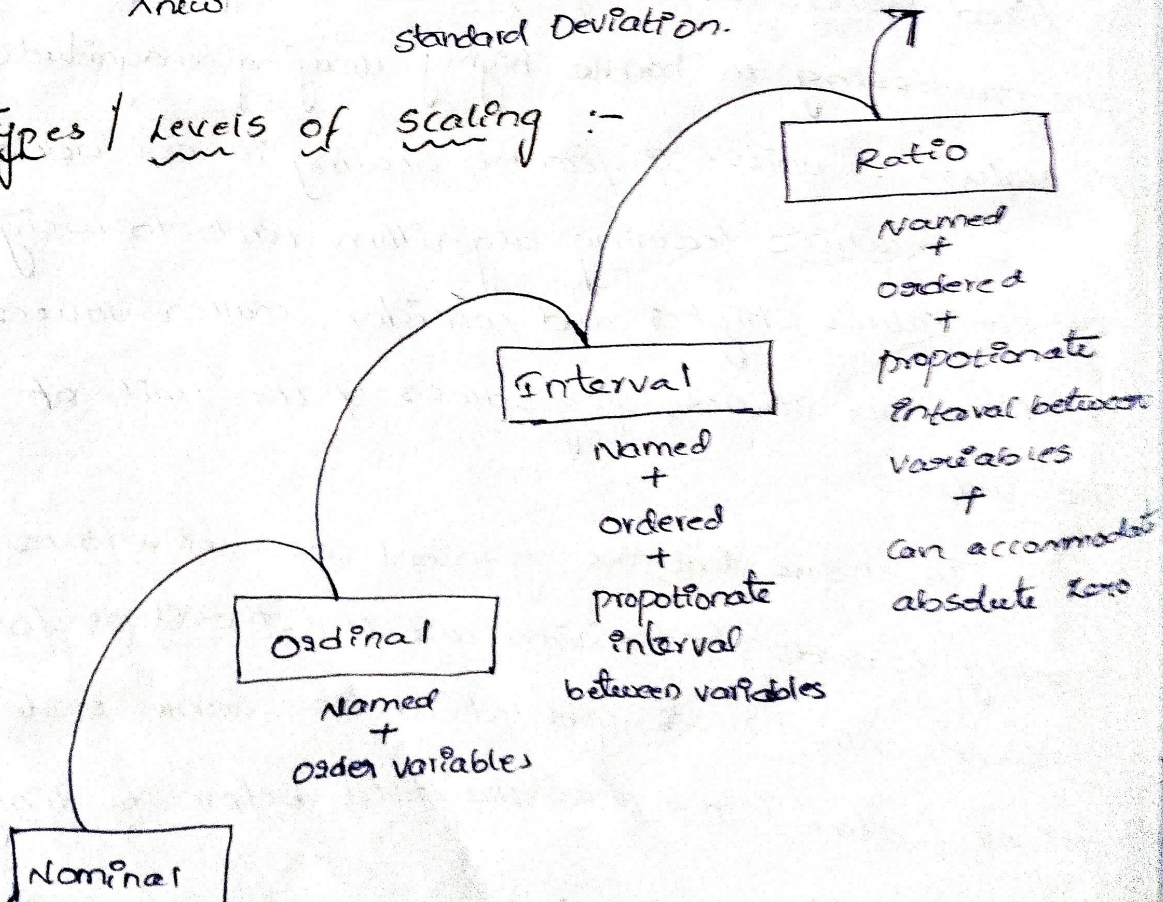
$$X_{new} = \frac{X_i - min(X)}{max(x) - min(X)}.$$

Standardization :-

It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{standard Deviation}}.$$

Types / Levels of scaling :-



Nominal Scale :-

* It is the first level of measurement scale in which the numbers serve as "tags" or "labels" to classify or identify the objects.
* A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.

characteristics :-

* A nominal scale variable is classified into two or more categories.
* It is qualitative. The numbers are used to identify the objects.
* The numbers don't define the object characteristics.

Ordinal Scale :-

* It is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them.
* Ordinal represents the "order". It is also known as qualitative data or categorical data.
* It can be grouped, named and also ranked.

characteristics :-

* The ordinal scale shows the relative ranking of the variables.
* It identifies and describes the magnitude of a variable.
* The interval properties are not known

Interval Scale :-

* The interval scale is 3rd level of measurement scale.
* It is defined as quantitative measurment scale in which the difference between the two variables is meaningful.

characteristics :-

* The interval scale is quantitative as it can quantify the difference between the values.
* It allows calculating the mean and median of the variables.
* It is the preferred scale in statistics as it helps to assign any numerical values to arbitrary assessment such as feeling etc.

Ratio Scale :-

* It is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale.
* It allows researchers to compare the differences or intervals. The ratio scale has a unique feature.
* It possesses the character of the origin or zero points.

Characteristics :-

* Ratio scale has a feature of absolute zero.
* It doesn't have negative numbers, because of its zero-point feature.

* It affords unique opportunities for statistical analysis
* the variables can be orderely added, Subtracted, multiplied, divided. Mean, median and mode can be calculated using the ratio Scale.
* Ratio Scale has unique and useful properties.
* It allows unit Conversions like Kilogram_calories, gram-Calories etc.
* Selecting meaningful features :-

It is the process of isolating the most consistent, non-redundant and relevant features to use in model construction.

* It aims to minimise both the classification error rate and the number of features.
* Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow.

Features :-

- Simpler models :- Simple models are easy to explain a model that is too complex and unexplainable is not valuable.

Shorter training times :- A more precise subset of feature decreases the amount of time needed to train a model.

Variance reduction :- Increase the precision estimates that can be obtained for a given simulation

- avoid the curse of high dimensionality :-

Dimensionality and the no. of features increases. the volume of space increases so fast that the available data become limited - pca feature selection may be used to reduce dimensionality.

selection methods :-

Feature selection algorithms are categorized as either supervised, which can be used for labeled data, or unsupervised which can be used for unlabeled data.

unsupervised techniques are classified as filter methods, wrapper methods, embedded methods or hybrid methods.

- Filter methods :- / subparts 4

* Filter methods select features based on statistics rather than feature selection cross-validation performance.
* A selected metric is applied to identify irrelavant attributes and perform recursive feature selection.
* Filter methods are either univariate, in which an ordered ranking list of features is established to inform the final selection of feature subset.

- Wrapper methods :- / subparts 4

* Wrapper feature selection methods consider the selection of a set of features as a search problem.
* This method facilitates the detection of possible interactions amongst variables.
* Examples are Boruta feature selection and

Embedded methods :- subparts 2

* The features that will contribute the most to each iteration of model training process are carefully extracted.
* Embedded feature selection methods integrate the feature selection machine learning algorithm as part of learning algorithm.
* Random forest feature selection, decision tree feature selection and LASSO feature selection are common embedded methods.

→ Choosing feature selection :-

- Numerical input, Numerical output :-

Feature selection regression problem with numerical input variables - use a correlation coefficient such as correlation coefficient or Spearman's rank coefficient.

- Numerical input, categorical output :-

Feature selection classification problem with numerical input variables - use a correlation coefficient taking into account the categorical target such as ANOVA correlation or Kendall's rank coefficient.

- Categorical input, Numerical output :-

Regression predictive modeling - use a correlation coefficient such as ANOVA correlation coefficient or Kendall's rank, but in reverse.

- Categorical input, categorical output :-

use a correlation coefficient such as chi-squared test or mutual information which is powerful method that is for data types.

Trees and Calle

filter, wrapper, embedded subparts.

* Filter methods :- Filter methods

* Chi-Square :- It is a technique determine the relationship between the categorical variables.

- The chi-square is calculated between each feature and the target variable and the desire number of features with the best chi-square value is selected.

Fisher's Score :-

- It is one of the Supervised technique of feature selection.

- It returns the rank of the variable on the fisher's criteria in descending order then we select the variable with a large fisher's score.

The value of the missing value ratio can be used for evaluating the feature set against the threshold va

$$\text{Missing value Ratio} = \frac{\text{No. of missing values} \times 100}{\text{Total no. of observations}}$$

Filter's method :-

* Information Gain :-

The information gain determines the reduction in entropy which transforming the dataset. It can be used as a feature selection technique by calculating the information gain of each variable with respect to the target variable.

Wrapper method :- In this method, selection of features is done by considering it as search problem, in which different combinations are made, evaluated and compared with other combinations, all the basis of the output of the model. and with this feature set, the model has

- If the trains algorithm by using the subsets of features iteratively, some techniques of wrapper methods are.

1. Forward selection :-

- It is an iterative process which begins with an empty set of features :
- Each iteration it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not.

Backward elimination :-

It is also an iterative approach but it is the opposite forward selection, the process begins by considering all the features and removes the least significant feature.

- this elimination process continues, until removing the features, does not improve the performance of this model.

Exhaustive feature selection :-

It evaluates each feature set or Brute force, it means this method tries and make each possible combinat of features and return the best performing feature

Recursive Feature elimination :-

It is a recursive greedy optimization approach where features are selected by recursively taking a smaller and smaller subset of feature.

- Embedded method :-

It combines the advantages of both filter & wrapper methods. These are fast processing methods similar to the filter method but more accurate than the filter method.

Some techniques of embedded are :

Regularization :- If codes are parameters the ML model for avoiding overfitting in the model.

The types of regularization techniques :

L1 (Regularization (LOSSO) elastic nets (L1, L2) Regularizati

L2 (Regularization) (Rite. Regularization)

Embedded methods :- 30/April/21

* The features that will contribute the most to each iteration of model training process are carefully extracted.
* Embedded feature Selection methods integrate the feature selection machine learning algorithm as part of learning algorithm.
* Random forest feature selection, decision tree feature selection and LASSO feature selection are common embedded methods.

→ Choosing feature selection :-

- Numerical input, Numerical output :-

Feature Selection regression problem with numerical input variables - use a correlation coefficient such as correlation coefficient or Spearman's rank coefficient.

- Numerical input, categorical output :-

Feature selection classification problem with numerical input variables - use a Correlation coefficient taking into account the categorical target such as ANOVA Correlation or Kendall's rank Coefficient.

- Categorical input, Numerical output :-

Regression predictive modeling - use a correlation Coefficient such as ANOVA correlation coefficient or Kendall's rank, but in reverse.

- Categorical input, categorical output :-

use a Correlation coefficient such as chi-Squared test or Mutual information which is powerful method that is for data types.

* Assesing feature importance with random Forests :-

Random Forests :- Random forests construct many individual decision trees at training. predictions from all trees are pooled to make final prediction, the mode of the classes for prediction classification or the mean prediction for regression.

Feature importance :-

* It is calculated as the decrease in node impurity weighted by the probability of reaching that node.
* The node probability can be calculated by the no. of samples that reach the node, divided by the total no. of samples. The higher the value more important to the feature.

scikit-learn :-

Scikit learn calculate a node importance

$$ni_j = w_j c_j - w_{left(j)} c_{left(j)} - w_{right(j)} c_{right(j)}$$

- ni sub (j) = importance of node j.
- w sub (j) = weighted no. of samples reaching node j.
- c sub (j) = impurity value of node j.
- left (j) = child node from left split on node j.
- right (j) = child node from right split on node j.

$$\longrightarrow fi_i = \frac{\sum_{j : \text{node } j \text{ splits on feature } i} ni_j}{\sum_{k \text{ ea ll nodes}} ni_k}$$

fi sub (i) = importance of feature i

ni sub (j) = importance of node j.

normalized to a value between 0 & 1.

$$normf_i = \frac{f_i}{\sum_{j \in \text{all features}} f_j}$$

Final feature importance, at random Forest level

$$RF\text{-}ff_i = \frac{\sum_{j \in \text{all tree}} normf_{ij}}{T}$$

where $RFf_i(Sub(i))$ = importance of feature i calculated in random Forest model.

$normf_i\ Sub(f_j)$ = normalized feature importance for i in tree j.

T = total no. of trees.

Spark :

$$f_{ij} = \sum_{j:\text{nodes } j \text{ splits on feature } i} S_j C_j$$

where, $f_i\ sub(i)$ = importance of feature i

$S\ sub(j)$ = no. of samples reaching node j.

$C\ sub(j)$ = impurity value of node j.

normalized Final feature importance at the Random Forest level.

$$norm\ f_{ij} = \frac{f_{ij}}{\sum_{j \in \text{all features}} f_{ij}}$$

where $normf_i\ sub(i)$ = normalized importance feature i

$f_i\ Sub(i)$ = importance of feature i.

Final Feature.

$$RF f_{ij} = \frac{\sum_j norm\ f_{ij}}{\sum_{j \in \text{all features}, k \in \text{all trees}} norm\ f_{ijk}}$$

$RFf_i\ sub(i)$ = importance of feature i calculated from all trees in RF.

normfi Sub(ij) = normalized feature importance for i in tree j.

Conclusion :-

This goal of this model was to explain how scikit learn and Spark implementation DecisionTrees and calculate Feature importance values.

filter, wrapper, embedded subparts.

* ~~In~~ Filter methods :- Filter methods.

* Chi-Square :- It is a technique determine the relationship between the categorical variables.

* Randomcolumns :-

1. It is under Supervised learning technique.
2. Mostly used for classification and regression problems.
3. It is based on ensemble learning (compiling multiple classifiers).

Random Forest Algorithm :-

Step 1 : Select random 'K' datapoints from the training set.

Step 2 : Build the decision trees associated with the selected data points (Subsets).

Step 3 : Choose the Number 'N' for decision trees that you want to build.

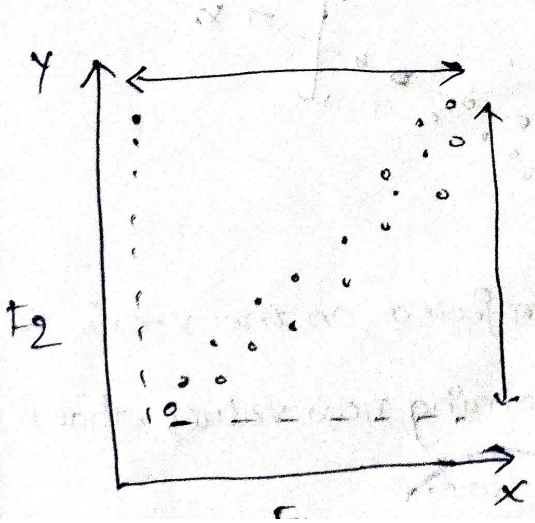
Step 4 : Repeat Step 1 & 2.

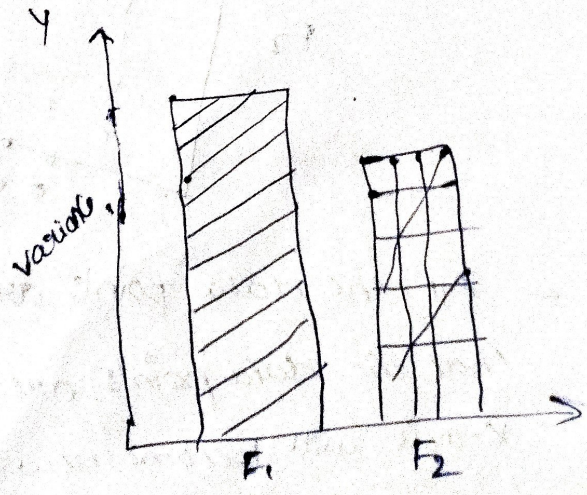
Step 5 : For new data points, find the predictions of each decision tree and assign the new data points to the category that wins the majority value.

* Unsupervised Dimensionality Reduction Via PCA :-

PCA - principle component analysis.

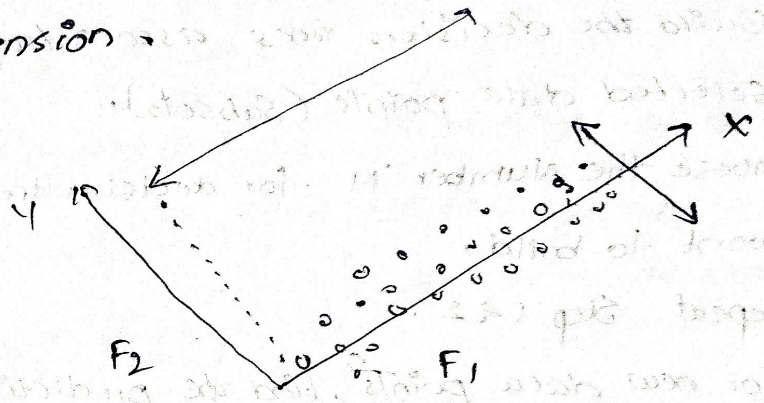
- It is under Supervised learning Algorithm.
- It is not a machine learning technique.





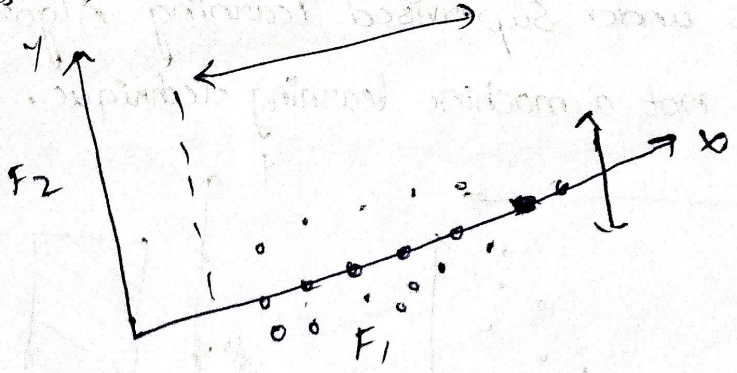
- PCA is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large data set of variables into a smaller one that still contains most of the information in the large set.

Based on the dataset, find a new set of orthogonal feature vector in such a way that the data spread is maximum in the direction of the feature vector or dimension.



Orthogonal Feature Vector :-

It is rotating feature which are on x-axis and y-axis at certain degree when compared to y-axis, x-axis variance is high. so, we project our data points on x-axis.



- All the data points are projected on the x-axis, so that the data points are becoming new values. that is x-axis will become new feature.

So that two dimensional data will become single dimensional data.

Note :-
PCA will be done only on independent variables but not dependent variables.

PCA Algorithm :-

Step 1 : Standardized the data points.

$$X_{new} = \frac{X - X\text{-}mean(x)}{std(x)}$$

Step 2 : Compute the two variance matrix of the whole dataset.

$$cov(x,y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 (y_i - \bar{y})^2$$

Step 3 : Compute Eigen vector and corresponding Eigen value

→ Let 'A' be a Square matrix, 'v' are a vector, 'λ' as a scalar that satisfies.

$Av = \lambda v$, then 'λ' is called Eigen value associated with eigen vector 'v' of 'A'.

Step 4 : Sort the eigen vector by decreasing eigen values and choose 'k' eigen vector with the largest Eigen values to form 'ad xk' dimensional matrix 'w'.

Where 'd' is the co-variance matrix.

K is the how many eigen vector we are taken.

* Feature Extraction :-

It tries to reduce the no. of features by creating new features from the existing ones then it discards the original features.

It refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set.

Feature Extraction offers three methods for Supervised Classification : K-Nearest Neighbor, Support Vector Machine or principal Component Analysis.

Eigen Vector of a matrix :-

$$AX = \lambda X$$

$$AX = I(\lambda X) = 0, \text{ where } I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$(A - I\lambda)\lambda = 0$$

$$\det |A - I\lambda| = 0$$

- The matrix 'A' represents a lineal transformation of the vectors (each column represents the vector).
- 'X' represents an eigen vector (Non-zero)
- λ is a scalar representing an eigen value.
- Let's find the eigen value and Eigen vector

$$\Rightarrow \left| \begin{bmatrix} 5 & -3 \\ -6 & 2 \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\Rightarrow \begin{vmatrix} 5-\lambda & -3 \\ -6 & 2-\lambda \end{vmatrix} = 0 \Rightarrow (5-\lambda)(2-\lambda) - (18) = 0$$

$$\Rightarrow 10 - 5\lambda - 2\lambda + \lambda^2 - 18 = 0$$

$$\Rightarrow \lambda^2 - 7\lambda - 8 = 0$$

$\Rightarrow \lambda(\lambda+1)(\lambda-8)=0$

$\lambda = -1, 8$

the eigen values are $\lambda = -1, 8$.

Let's first consider $\lambda=1$ & $(A-I\lambda)X=0$

$$\begin{bmatrix} 5-\lambda & -3 \\ -6 & 2-\lambda \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 6 & -3 \\ -6 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$6x_1 - 3x_2 = 0$ $\qquad 3(2x_1 - x_2) = 0$

$-6x_1 + 3x_2 = 0$ $\qquad 2x_1 - x_2 = 0$

The eigen vector $\lambda = -1$ is $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$

Eigen vector $\lambda = 8$ is $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$

$$\begin{bmatrix} 5 & -6 \\ -6 & 12 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$5x_1 - 6x_2 = 0$

$-6x_2 + 12x_2 = 0$

$\begin{bmatrix} -1 \\ -2 \end{bmatrix}$

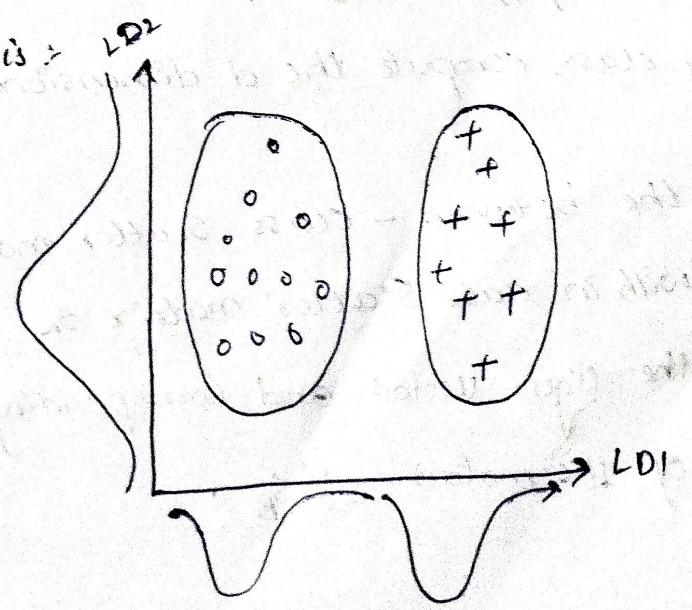
where $\lambda = 8$

$AX = \begin{bmatrix} 8 \\ -8 \end{bmatrix}$

$AX = -8\begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$\lambda \quad x$

* Supervised data Compression via linear discriminant Analysis :



A linear discriminant as shown on the x-axis, would separate the two normal distributed classes well. Although the exemplary linear discriminant shows on the y-axis, captures a lot of the variance in the dataset, it would fail as a good linear discriminant analysis since it does not capture any of the class-discriminatory information.

One assumption in LDA is that the data is normally distributed. Also we assume that the classes have identical covariance matrices and that the samples are statistically independent of each other. However, even if one or more of these assumptions are slightly violated, LDA for dimensionality reduction can still work reasonably well.

The inner workings of linear discriminant analysis:-

Before we divide into the code implementation, Main steps that are required to perform LDA:

1. Standardize the d-dimensional dataset (d is the number of features).
2. For each class, compute the d-dimensional mean vector.
3. Construct the between-class scatter matrix S_B and the within-class scatter matrix S_W.
4. Compute the eigen vectors and corresponding eigen values of the matrix $S_W^{-1} S_B$.

5. Sort the eigen values by decreasing order to rank the corresponding eigen vectors.

6. Choose the k eigen vectors that corresponds to the k largest eigen values to construct a dxk - dimensional transformation matrix W; the eigen vectors are the column of the matrix.

7. project the samples onto the new feature subspace using the transformation matrix W.