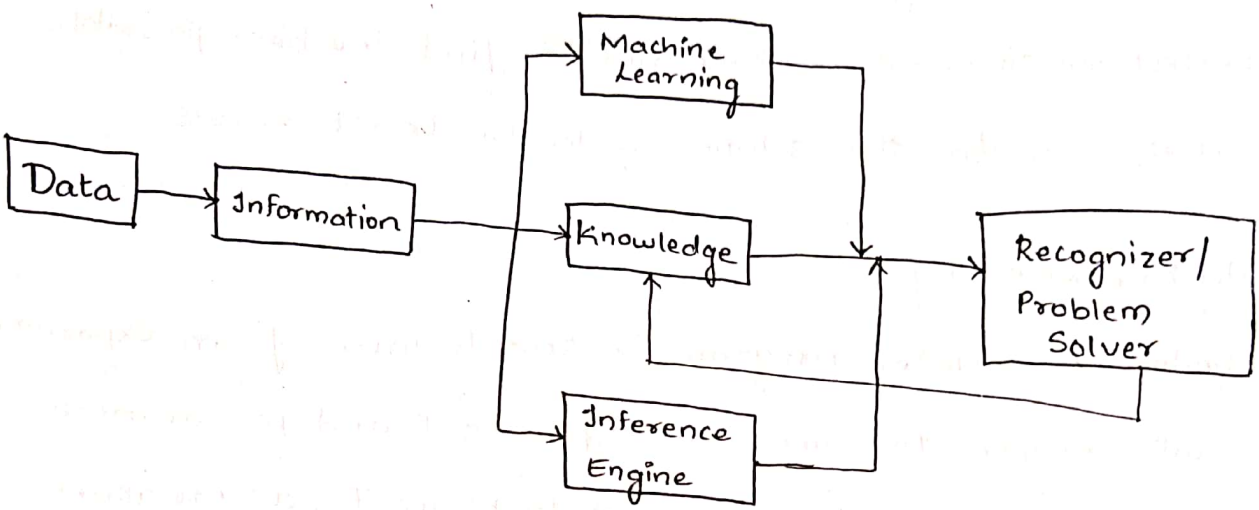
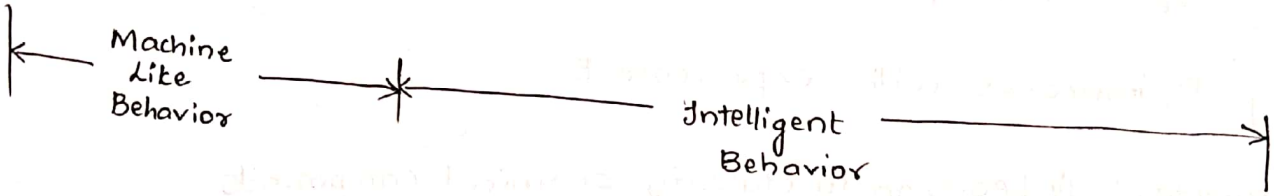
04/01/2020
Saturday

Advanced Machine Learning

Machine Learning Components





- Goal of machine learning is to build computer systems that can learn from their experience and adopt to their environments.
- Machine learning is an important aspect or component of intelligence.
- Intelligence is ability to learn.

Well-Posed Learning Problems

Jacques-Hadamard defined the term "well-posed problem". A problem that has a unique solution that changes continuously with the initial conditions.

Ill-Posed Problems

They are typically the subject of ML methods & AI, including Statistical learning. these methods do not aim to find the perfect solution; rather they aim to find the best possible solution and/or the solution with the least errors.

Machine Learning

Definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Examples: 1). Learning to classify chemical compounds.
2). Learning to drive an autonomous vechicle.
3). Learning to play bridge.

Concept Space

A concept is the representation of the problem wrt the given attributes. For eg, if we're talking about the problem scenario of concept SICK defined over the attributes T & BP, then the CS is defined by all the comb^{-n} of values of SK for every instance x.

- Concept Learning involves 2 algorithms:

Find S Algorithm

↳ most specific hypothesis.

↳ Consider only +ve examples.

Algorithm:

1). Initialize h to most specific hypothesis.

$h = \{'\phi', '\phi', \dots\ '\phi'\}$

2). For each +ve example:

For each attribute in the example:

if attribute value = hypothesis value:

Do nothing

else

Replace hypothesis with more general constraint '?'.

↳ General Hypothesis.

↳ Specific Hypothesis.

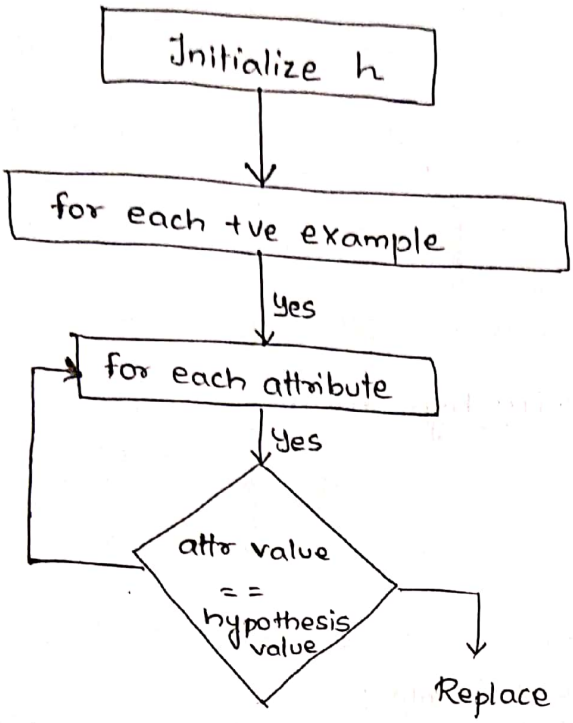
$G = \{'?', '?', '?', \dots\ '?'\}$

↳ No. of. attributes.

$S = \{'\phi', '\phi', '\phi', \dots\ '\phi'\}$

↳ No. of. attributes.

Flowchart



Database

Concept: Days on which person enjoys sport.

Sky	Temp	Humidity	wind	Water	Forecast	Enjoy
Sunny	warm	Normal	Strong	Warm	Same	yes
Sunny	warm	High	Strong	Warm	Same	yes
Rainy	cold	High	Strong	Warm	Change	No
Sunny	warm	High	Strong	Cool	Same	yes

Step 1

$h_0 = \{'\phi', '\phi', '\phi', '\phi', \ldots\ldots '\phi'\}$

$h_0 = \{Sunny, warm, \ldots\ldots warm, same\}$

$h_0 = \{'Sunny', 'warm', '?', 'Strong', 'warm', 'Same'\}$

After Comparison,

$h_0 = \{'Sunny', 'warm', '?', 'strong', '?', 'Same'\}$

Candidate Elimination Algorithm

↳ Uses version space

↳ Consider both +ve & -ve results.

↳ We have both specific & general hypothesis.

↳ For a +ve eg,

we tend to generalise specific hypothesis.

↳ For a -ve eg,

we tend to specify mate general hypothesis more specific.

↳ General Hypothesis $G = \{?, ?, ---- ?\}$

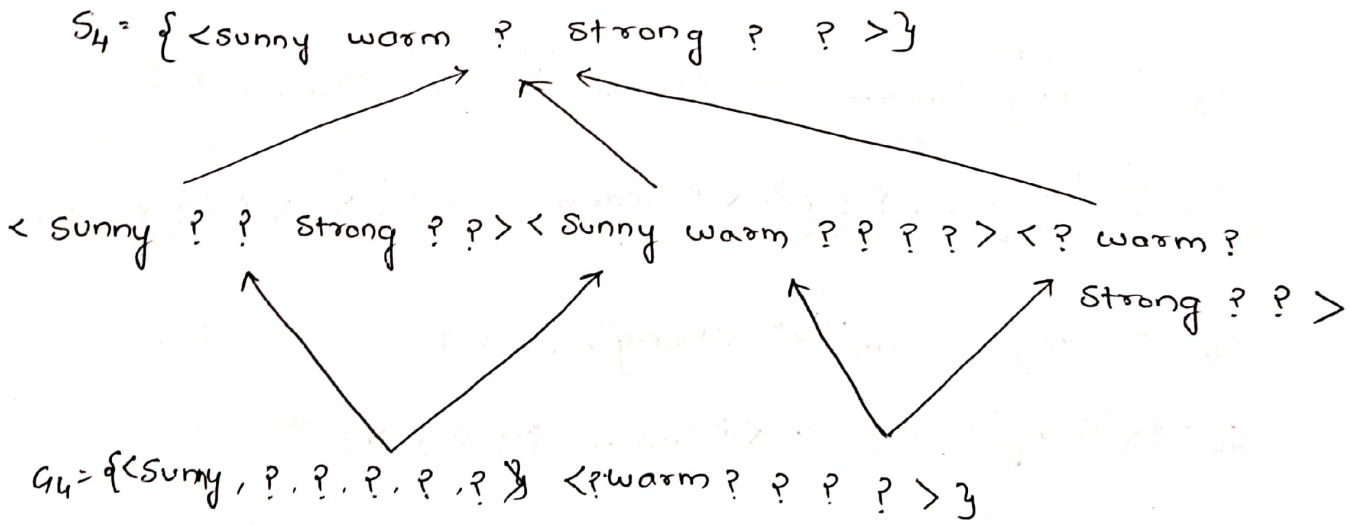
↳ Specific Hypothesis $S = \{\phi, \phi, ---- \phi\}$

↳ Version Space

+ve ↓ $S = \{\phi, \phi, ---- \phi\}$

⋮

$G = \{?, ?, ---- ?\}$ ↑ -ve



Candidate Elimination Algorithm

> Initialize G and S as most general & specific hypothesis.

> For each example e:

if e is +ve:

→ make specific hypothesis more general

else:

→ make general hypothesis more specific.

$S_0 = \{\phi, \phi, \phi, \phi, \phi, \phi\}$ $G_0 = \{?, ?, ?, ?, ?, ?\}$

$S_1 = \{$"Sunny", "warm", "Normal", "Strong", "warm", "same"$\}$

$G_1 = \{? ? ? ? ? ?\}$

$S_2 = \{$"Sunny", "warm", "?", "Strong", "warm", "same"$\}$

$G_2 = \{? ? ? ? ? ?\}$

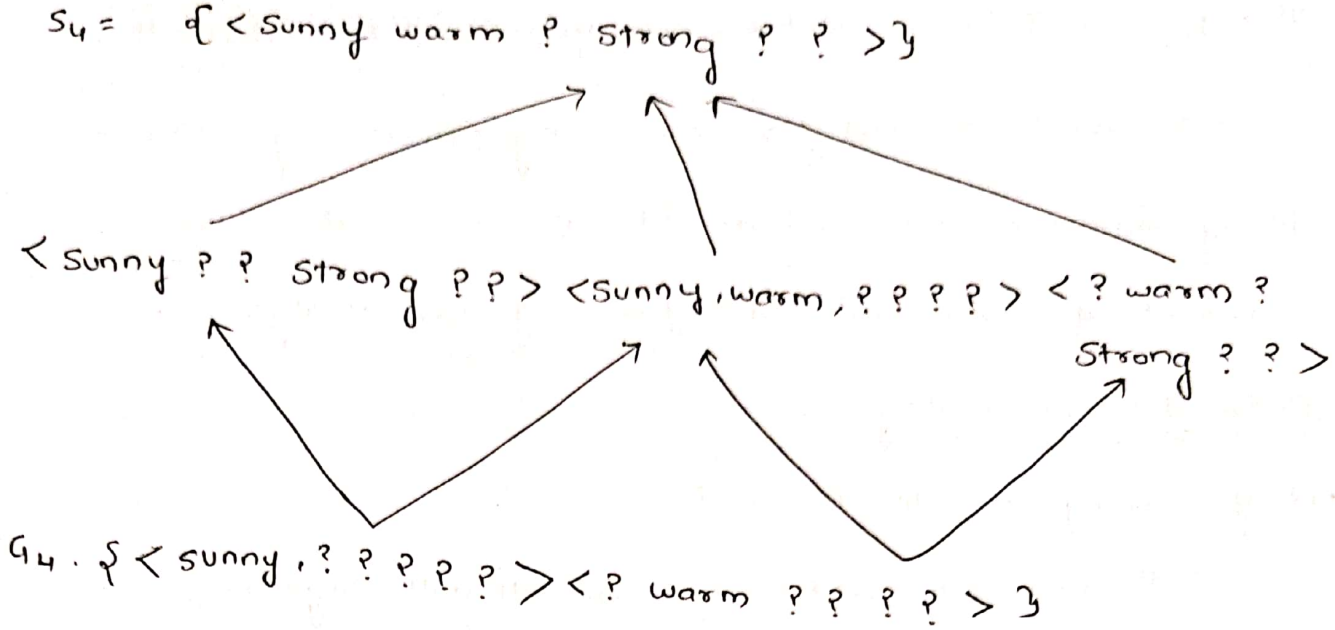
$S_3 = \{$"Sunny", "warm", "?", "Strong", "warm", "same"$\}$

$G_3 = \{<$"Sunny" ? ? ? ? ?>, <? "warm" ? ? ? ?>, <? ? ? ? ? same$\}$

$S_4 = \{$"Sunny", "warm", "?", "Strong", ?, ?$\}$

$G_4 = \{<$"Sunny" ? ? ? ? ?>, <? "warm" ? ? ? ?>$\}$

Version Space



Feature Extraction And Selection

Feature extraction aims to reduce the number of features in a dataset by creating new features from the existing ones. These new reduced set of features should then be able to summariz most of the information ~~combined~~ contained in the original se of feature. In ML, pattern recognition & in image processing feature extraction starts from an initial set of measured data & builds derived values (features) intended to ~~the~~ be informative & non-redundant, facilitating the subsequent learning and generalization steps, & in some cases leading to better human.

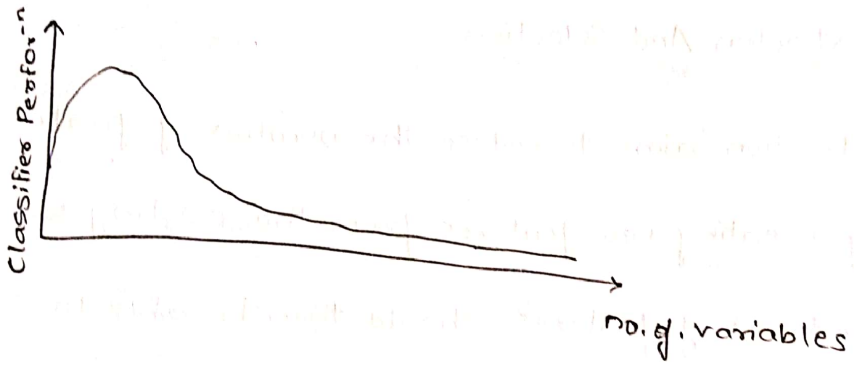
Why Feature Extraction?

The process of FE is useful when you need to reduce the no. of resources needed for ~~resourcing~~ processing without losing important information. FE can also reduce the amount of redundant data for a given analysis.

Curse of Dimensionality

- No. of training examples is fixed

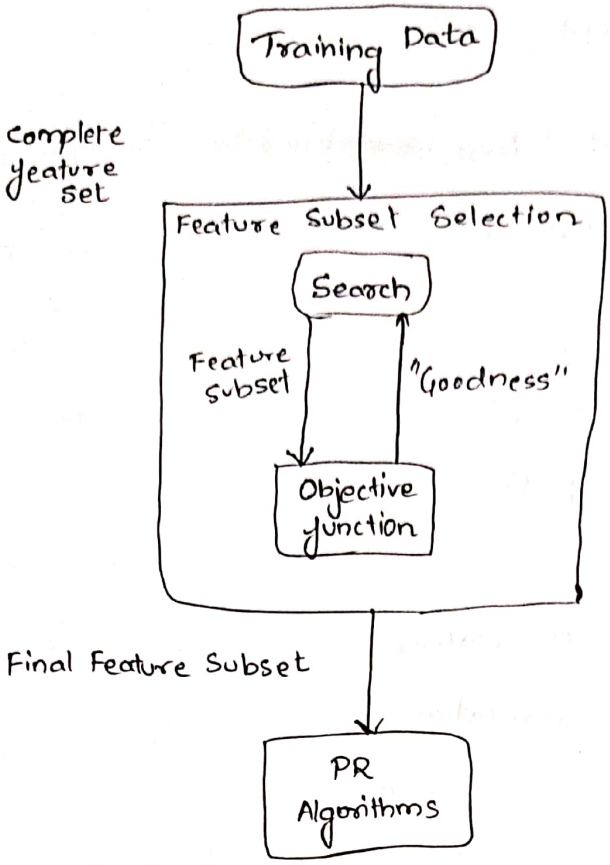
  ⇒ the classifier's performance usuall will degrade for a large no. of features.



Steps of Feature Selection

- Feature selection is an optimization problem.
  - Step 1). Search the Space of all possible feature subsets.
  - Step 2). Pick the subset that is optimal or near-optimal wrt some objective function.

- Search Strategies
  - Optimum
  - Heuristic
  - Randomized.
- Evaluation Strategies
  - Filter Methods
  - Wrapper Methods



Evaluation of subsets

- Supervised (Wrapper Method)
  - Train using selected subset
  - Estimate error on validation dataset.
- Unsupervised (Filter method)
  - Look at input only.
  - Select the subset ~~&~~ the most information.

Feature Selection

Univariate (looks at each ~~variate~~ feature indp^ly of others)

- Pearson correlation coefficient
- F-Score
- Chi-square
- Signal to noise ratio.
- Mutual Information.

Pearson Correlation Coefficient

- Measures the correlation bet^{n} two variables.
- Formula for Pearson correlation.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

r value +1 → predict +ve correlation.

r value -1 → predict -ve correlation.

Signal to Noise Ratio

- Difference in means divided by difference in standard deviation between the two classes.

$$S2N(X, Y) = (\mu_x - \mu_y) / (\sigma_x - \sigma_y)$$

- Large values indicate a strong correlation.

Methods of Feature Selection

All the methods of feature selection involve searching for the best subset of features, we have two approaches.

i). Forward Search.

ii). Backward Search.

- After removing a feature or adding a feature the resultant set of feature is evaluated & decision is taken to keep the feature or not.

- The evaluation function is called the objective function J.
- The evaluation usually employed the classification accuracy obtained on a validation set of points.
- This can also be a classification error.
- If E is the % error then objective function J = 100 - E.

Feature Selection

11/01/2020
Saturday

$T_D = F = \{x_1, x_2, \ldots x_N\}$

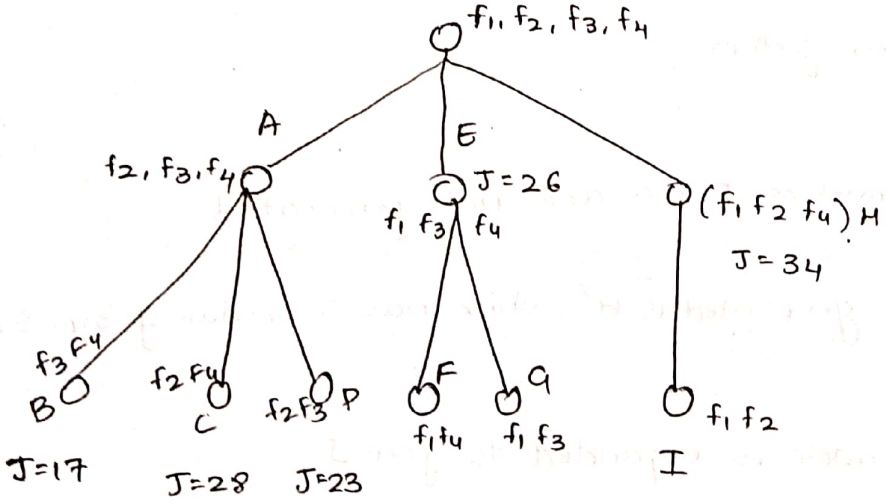
$F' \subseteq F = \{x_1, x_2, \ldots x_m\}$

$m < N$

↳ Irrelevant Features.
↳ Redundant Features.

Branch and Bound

- Method of Feature selection.
- Consider a TD of 4 feature use Branch & Bound method to select 2 features.



Step I)
- When the leftmost node B reached which corresponds to feature subset $f_3 f_4$, let evaluation of node J be 17

Step II)
At this stage, since this is a leaf node the bound B = 17.

Step III)
The next node generated is C having J value of 28.

Step IV)
Since, it is greater than value of B, the bound B is updated to 28, the best subset so-far is $f_2 f_4$.

Step V)
The next node is D corresponding to the feature subset $f_2 f_3$.

Step VI)
This has a J value of 23 which is smaller than B, so B remains unchanged.

Step VII)
The next node generated is E, which is found to have J value of 26.

Step VIII). Since the J value of this node is less than B, this branch is not expanded further.

Step IX)
This means that branches F & G are not generated.

Step X)
The next node to be generated is H, which has J value of 34. Since

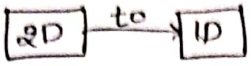
Step XI)
Since, 34 > B this node is expanded to give I.

Step XII
The J value of I is 20, which is lower than bound.

Step XIII
Now, the entire tree is completed & the node with the lower bound B has the criterian func^n, is the best subset which is $f_2 f_4$. this is based on the fact they are selecting two out of four features.

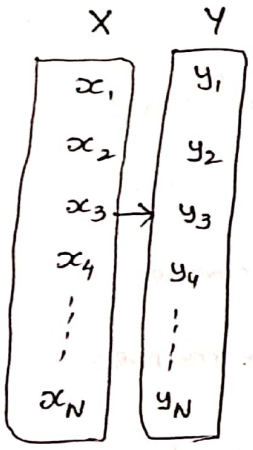
Feature Extraction

- Transformation of data from higher dimensional to lower dimensional.
- It transforms or projects a original set of features into a new space, which smaller no. of dimension.



- the process of identifying certain features of training data. This is a preprocessing step of
- Before Classification is carried it is necessary to decide what attributes of training data are measured or recorded.
- The features choosen should be discriminating features of training data.
- It is an important stage of preprocessing. The feature extraction training data classification carried out.

Training Data:



$Y = W^T X$

W^T is the orientation of the separator/line/plane/hyperplar

- We have two methods of feature extraction,

1). Fisher's Linear Discrement

2). PCA

1). Fisher's Linear Discrement

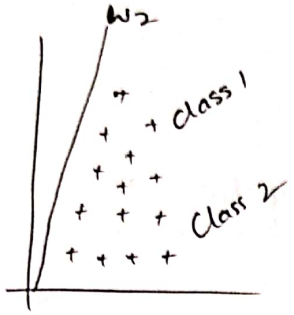


Fig: Class Separated (No overlap)

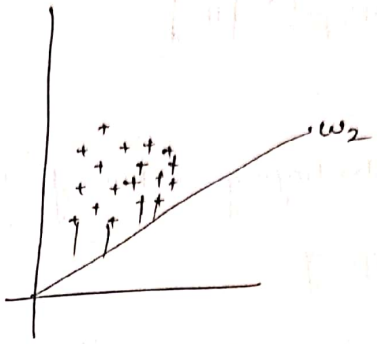


Fig: Class Overlap

- Methods which are manageable & dimensions are low become highly manageable
- This is on account of large computational time required.
- To overcome this problem, we tried to reduce dimensions.
- One method of dimensionality reduction is to map the N dimensional samples on to a line.
- But here the orientation of the line is very important for class separation
  this are explained in fig-① and fig-②.
- In fig-① if w_1 of the line on which the samples of two classes are mapped. It is clear by the two classes are that the mapped samples of w_1 overlapped with one-another & there is no class separation.
- Fig-② shows the line having orientation having w_2 when the samples of the two classes are mapped on w_2, we find that the samples of two classes are well separated. This is a desired mapping.

- Our aim in mapping will be to find good orientation such as w_2.

-

Let, $x_1\ x_2\ x_3\ x_4 \dots x_N$

The sample x can be mapped on to a line w by $y = w^T x$

$$= [w_1, w_2, \dots w_d] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$w_1, w_2, w_3, w_4 \dots w_d$ is found that there is a, there is a good class separation. The fisher's linear discrement func^n is that func^n $w^T x$ for which the criterian func^n,

$$J(w) = \frac{w^T S_B w}{w^T S_w w} \quad \text{is maximum.}$$

where, S_B = Class Scatter matrix

$$= (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T.$$

where, $\mu_i = \sum x_i$

$x_i \in i^{th}$ class.

S_w = within class scatter matrix

$= S_1 + S_2$

where $S_i = \sum (x - \mu_i)(x - \mu_i)^T$

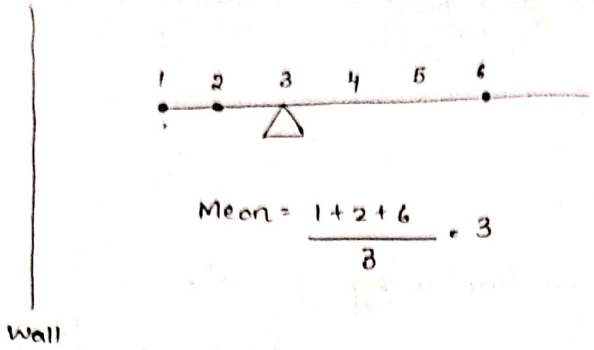
$x_i \in i^{th}$ class

The solution of this is,

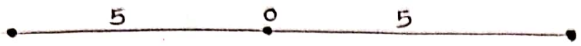
$$w = S_w^{-1}(\mu_1 - \mu_2)$$

//.

15/01/2020
Wednesday

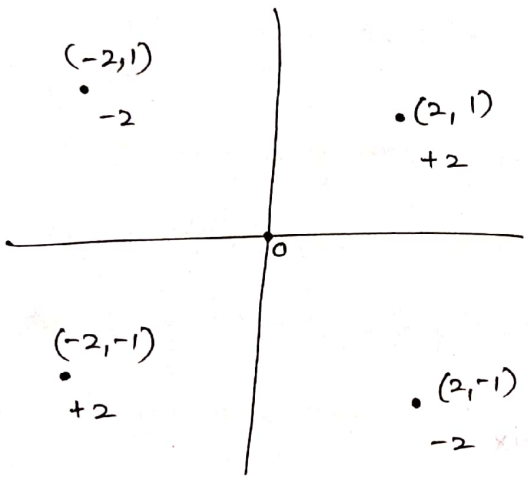


$$\text{Mean} = \frac{1+2+6}{3} = 3$$

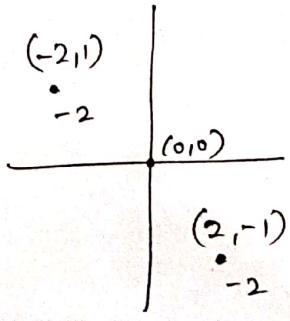


$$\text{Variance} = \frac{5^2+0^2+5^2}{3} = 50/3$$

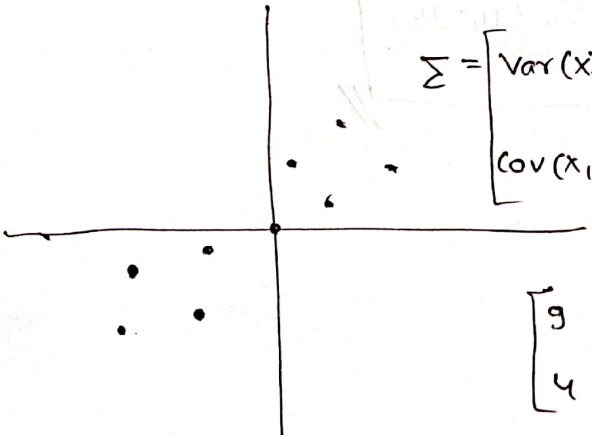
Product of Coordinates



$$\text{covariance} = \frac{(-2)+0+(-2)}{3} = -4/3$$



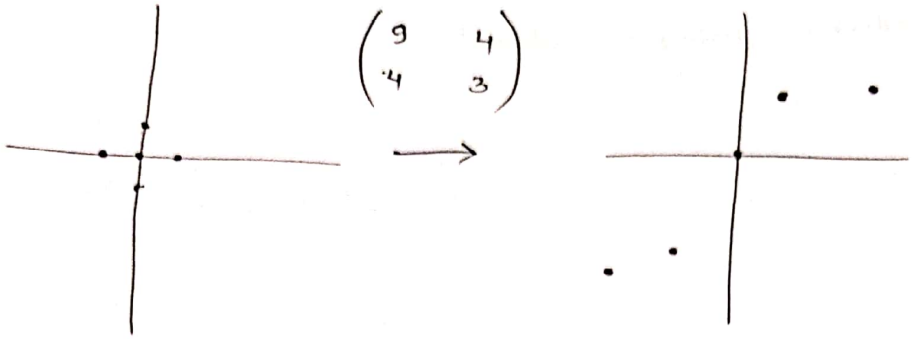
Covariance Matrix



$$\Sigma = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(Y) \end{bmatrix}$$

$$\begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$$

Linear Transformation



$(x, y) = (9x + 4y, 4x + 3y)$

$(0,0) \rightarrow (0,0)$

$(1,0) \rightarrow (9,4)$

$(0,1) \rightarrow (4,3)$

$(-1,0) \rightarrow (-9,-4)$

$(0,-1) \rightarrow (-4,-3)$

Principle Component Analysis

- PCA is a procedure by which no. of attributes are reduced. the attempt here is to find the smaller no. of attributes which are uncorrelated.
- The first principle component is the most important & accounts for as much of the variability as possible.
- The second principle component comes next.
- the no. of attributes are reduced by projecting the data in the dirⁿ of maximum variance.
- The method involves finding the eigen vectors & the corresponding eigen values of correlation matrix.
- If the eigen vectors are ordered in descending order of the eigen values, the first eigen vector value gives the dirⁿ of the largest variance of data.

- By excluding the direc^{ns} giving very low eigen values, we can reduce the no. of attributes being considered.