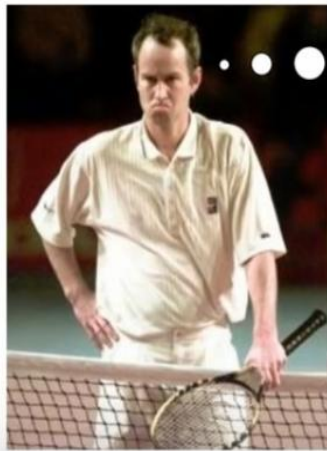


## Gini And Entropy

### Decision Trees



To play or  
not to play?

## Impurity Criterion

### Gini Index

$$I_G = 1 - \sum_{j=1}^c p_j^2$$

$p_j$ : proportion of the samples that belongs to class  $c$  for a particular node

### Entropy

$$I_H = - \sum_{j=1}^c p_j \log_2(p_j)$$

$p_j$ : proportion of the samples that belongs to class  $c$  for a particular node.

\*This is the the definition of entropy for all non-empty classes ( $p \neq 0$ ). The entropy is 0 if all samples at a node belong to the same class.

## 1. Concept learning: an example

Given the data:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

predict the value of PlayTennis for

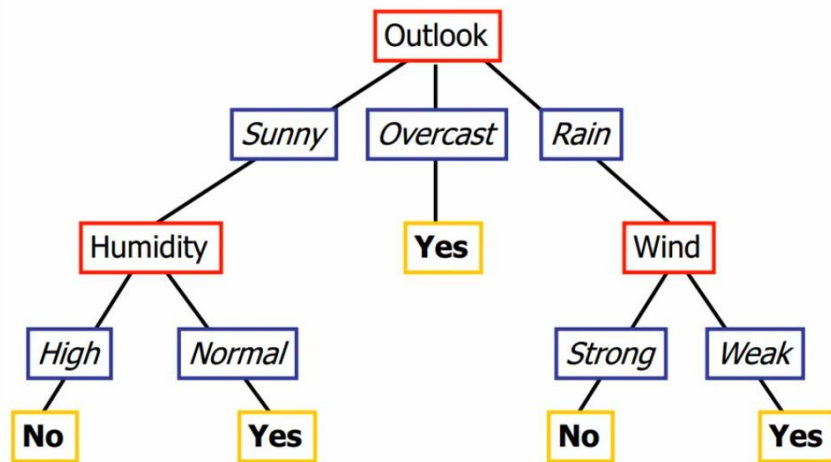
*(Outlook = sunny, Temp = cool, Humidity = high, Wind = strong)*



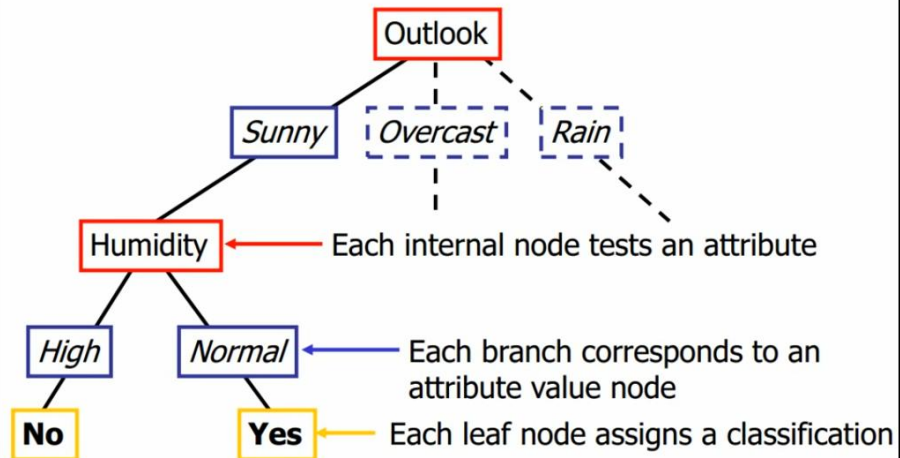
## Decision Tree for PlayTennis

- Attributes and their values:
  - Outlook: *Sunny, Overcast, Rain*
  - Humidity: *High, Normal*
  - Wind: *Strong, Weak*
  - Temperature: *Hot, Mild, Cool*
- Target concept - Play Tennis: *Yes, No*

## Decision Tree for PlayTennis

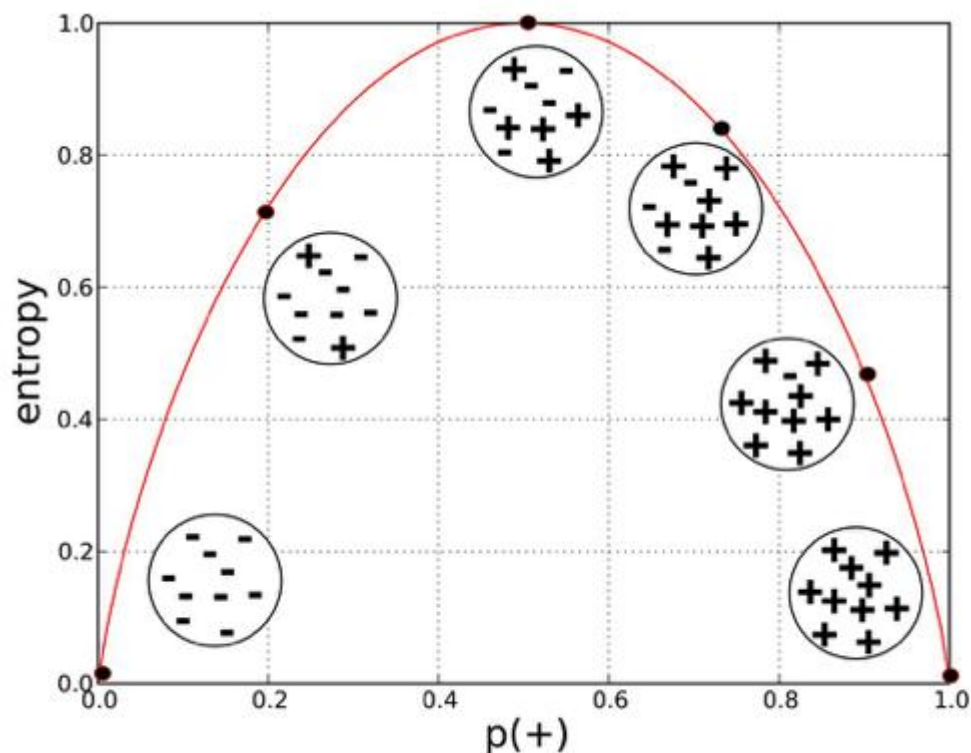


## Decision Tree for PlayTennis



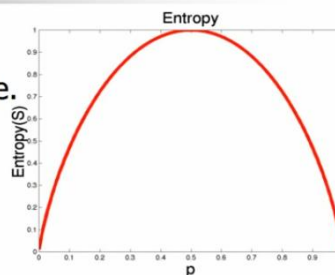
**Entropy**, as it relates to **machine learning**, is a measure of the randomness in the information being processed.

In case of diabetes dataset ,if all patients has no diabetes,  
Then no information is gained ,it is 0. Or if all people  
diabetes then also information gain is 0 .But when  
50% people has diabetes then entropy is 1.



## Entropy

- Entropy is the measure of the homogeneity of a sample in a node.
  - If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.
  - $S$  is a sample of training examples
  - $p_+$  is the proportion of positive examples
  - $p_-$  is the proportion of negative examples
  - Entropy measures the impurity of  $S$
- $$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$



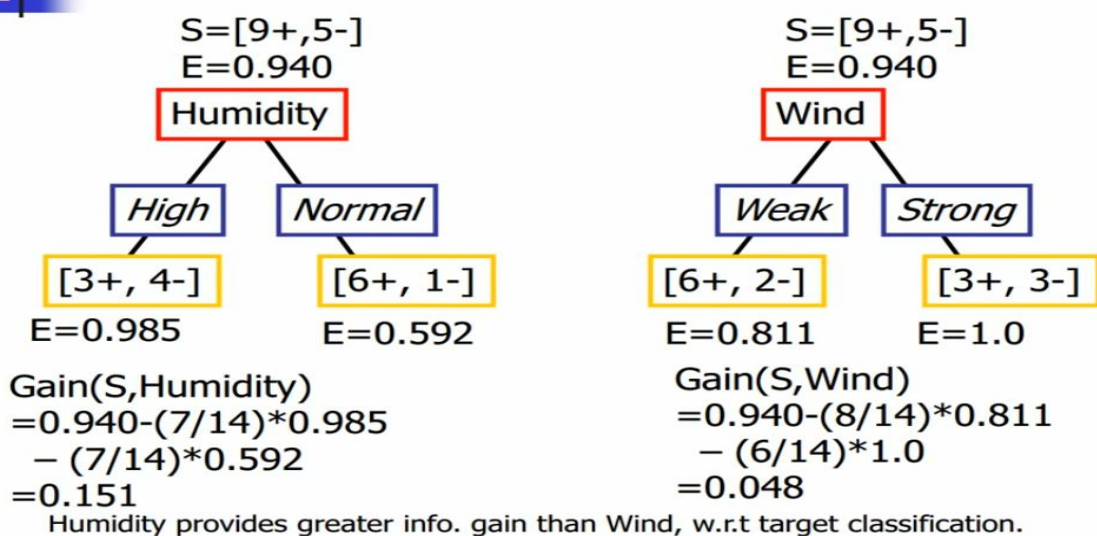
## Information Gain

- Gain(S,A): expected reduction in entropy due to sorting S on attribute A
- The information gain is based on the decrease in entropy after a dataset is split on an attribute.

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} |S_v|/|S| \text{Entropy}(S_v)$$

- Calculate entropy of the target.
- $\text{Entropy}(\text{PlayTennis}) = -p_+ \log_2 p_+ - p_- \log_2 p_-$   
 $\rightarrow (-0.36)\log_2(0.36) - (0.64)\log_2(0.64) \rightarrow 0.94$

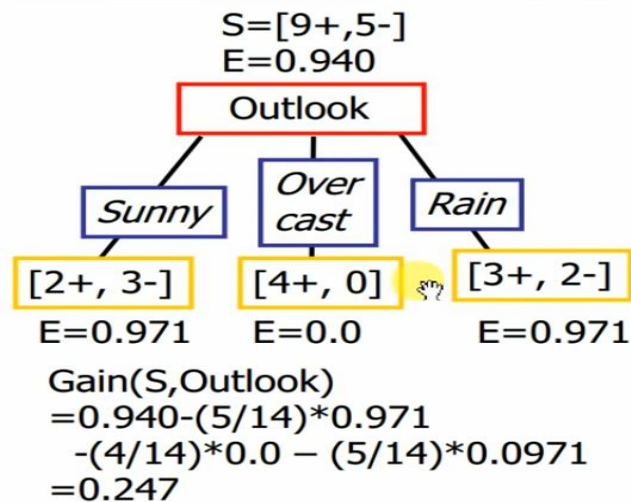
## Selecting the Next Attribute







## Selecting the Next Attribute



Here, overcast is 0.0 as it is very homogeneous sample.



## Selecting the Next Attribute

The information gain values for the 4 attributes are:

- $\text{Gain}(S, \text{Outlook}) = 0.247$
- $\text{Gain}(S, \text{Humidity}) = 0.151$
- $\text{Gain}(S, \text{Wind}) = 0.048$
- $\text{Gain}(S, \text{Temperature}) = 0.029$

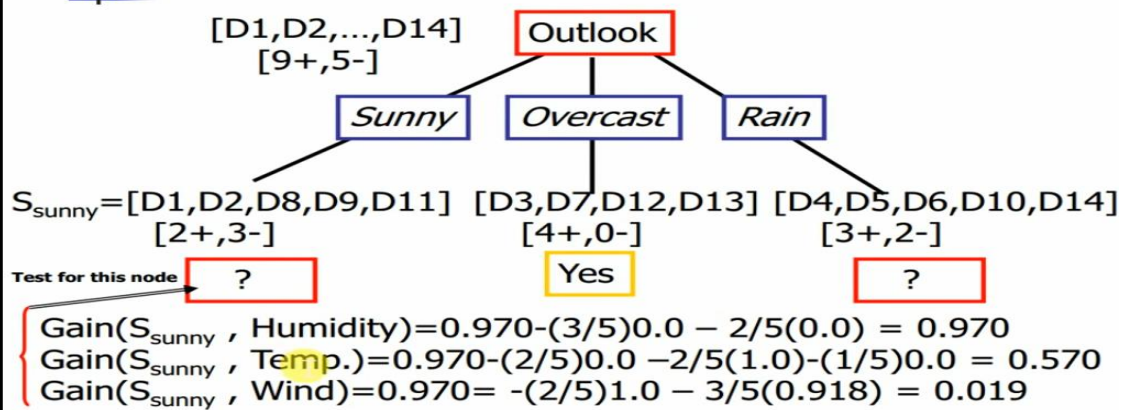
where  $S$  denotes the collection of training examples



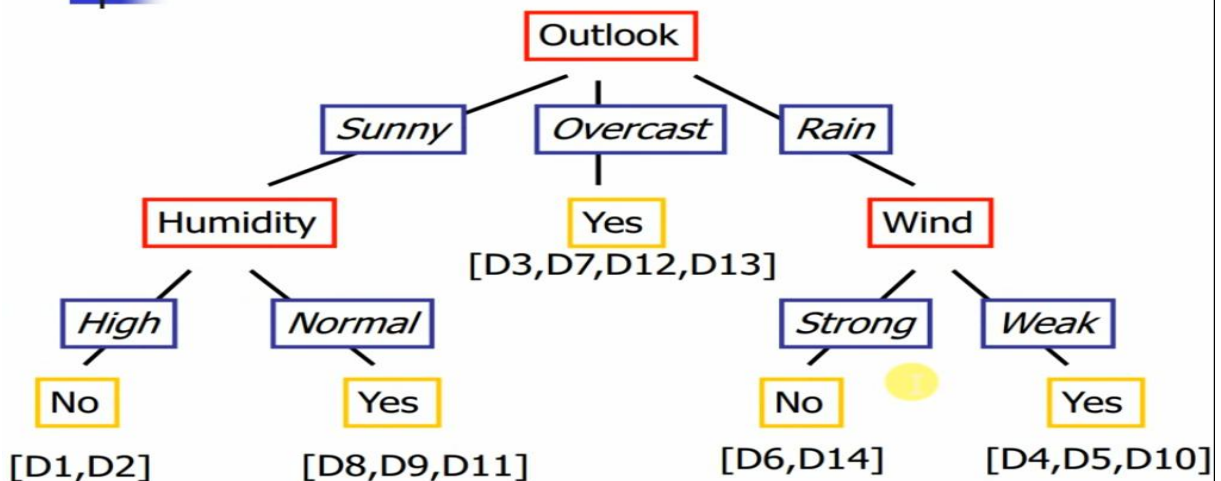


## ID3 Algorithm

Note:  $0\log_2 0 = 0$



## ID3 Algorithm



Gini Index



## Gini Index

- If a data set  $D$  contains examples from  $n$  classes, gini index,  $gini(D)$  is defined as:

$$gini(D) = 1 - \sum_{j=1}^n p_j^2$$

where  $p_j$  is the relative frequency of class  $j$  in  $D$

- If a data set  $D$  is split on  $A$  into two subsets  $D_1$  and  $D_2$ , the gini index  $gini(D)$  is defined as

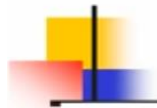
$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Reduction in Impurity:  $\Delta gini(A) = gini(D) - gini_A(D)$

## Training Examples

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



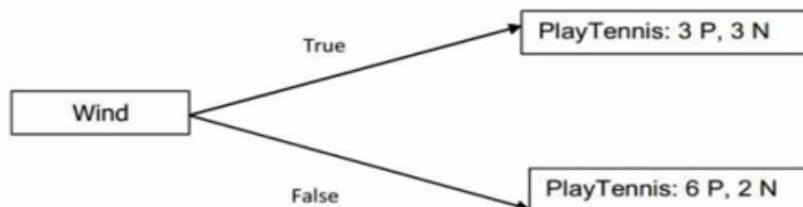


# Gini Index I

## Gini index calculation:

There are 5 Ns and 9 Ps, so the

- Calculate the information gain after the Wind test is applied:



$$\text{Gini (PlayTennis|Wind=True)} = 1 - (3/6)^2 - (3/6)^2 = 0.5$$

$$\text{Gini (PlayTennis|Wind=False)} = 1 - (6/8)^2 - (2/8)^2 = 0.375$$

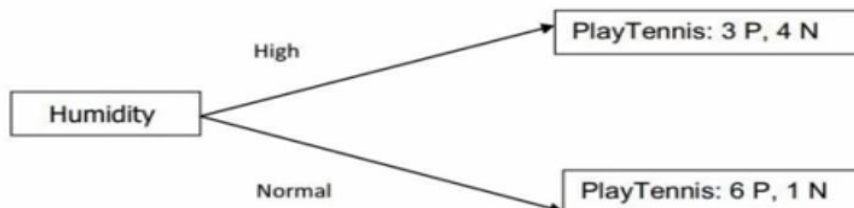
Therefore, the Gini index after the Wind test is applied is

$$6/14 \times 0.5 + 8/14 \times 0.375 = 0.4286$$



# Gini Index II

- Calculate the information gain after the Humidity test is applied:



$$\text{Gini (PlayTennis|Humidity=High)} = 1 - (3/7)^2 - (4/7)^2 = 0.4898$$

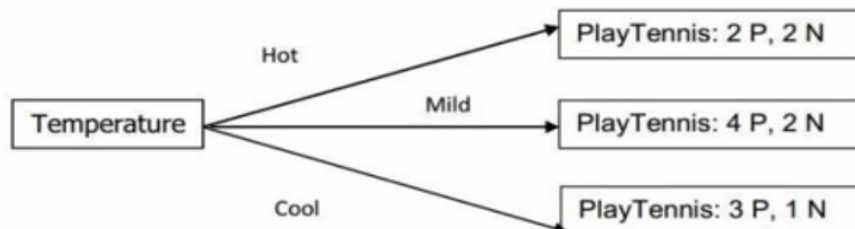
$$\text{Gini (PlayTennis|Humidity=Normal)} = 1 - (6/7)^2 - (1/7)^2 = 0.2449$$

Therefore, the Gini index after the Humidity test is applied is

$$7/14 \times 0.4898 + 7/14 \times 0.2449 = 0.3674$$

## Gini Index III

- Calculate the information gain after the Temperature test is applied:



$$\text{Gini (PlayTennis | Temperature = Hot)} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$\text{Gini (PlayTennis | Temperature = Mild)} = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.4444$$

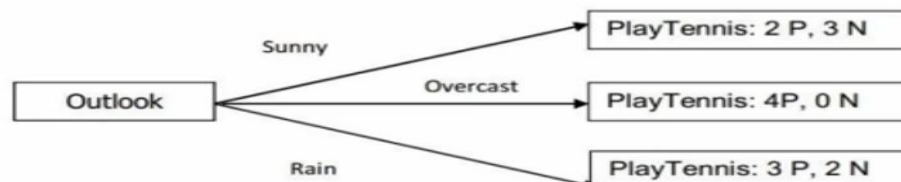
$$\text{Gini (PlayTennis | Temperature = Cool)} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

Therefore, the Gini index after the Temperature test is applied is

$$\frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.4444 + \frac{4}{14} \times 0.375 = 0.4405$$

## Gini Index IV

- Calculate the information gain after the Outlook test is applied:




$$\text{Gini (PlayTennis | Outlook = Sunny)} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$\text{Gini (PlayTennis | Outlook = Overcast)} = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$\text{Gini (PlayTennis | Outlook = Rain)} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

Therefore, the Gini index after the Temperature test is applied is

$$\frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 = 0.3429$$



# Gini Index v

After calculating all attributes:

- $\text{gain}(\text{outlook}) = 0.3429$
- $\text{gain}(\text{temperature}) = 0.4405$
- $\text{gain}(\text{humidity}) = 0.3674$
- $\text{gain}(\text{windy}) = 0.4286$

Here we are getting temperature as a highest gain 0.4405 ,hence root node will be temperature.

## Examples of SVM Kernels

Let us see some common kernels used with SVMs and their uses:

### 4.1. Polynomial kernel

It is popular in image processing.

Equation is:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$$

*Polynomial kernel equation*

where d is the degree of the polynomial.

## 4.2. Gaussian kernel

It is a general-purpose kernel; used when there is no prior knowledge about the data. Equation is:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

*Gaussian kernel equation*

## 4.3. Gaussian radial basis function (RBF)

It is a general-purpose kernel; used when there is no prior knowledge about the data.

Equation is:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

*Gaussian radial basis function (RBF)*

, for:

$$\gamma > 0$$

*Gaussian radial basis function (RBF)*

Sometimes parametrized using:

$$\gamma = 1/2\sigma^2$$

*Gaussian radial basis function (RBF)*

## 4.4. Laplace RBF kernel

It is general-purpose kernel; used when there is no prior knowledge about the data.

Equation is:

$$k(x, y) = \exp \left( -\frac{\|x - y\|}{\sigma} \right)$$

*Laplace RBF kernel equation*

## 4.5. Hyperbolic tangent kernel

We can use it in neural networks.

Equation is:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa \mathbf{x}_i \cdot \mathbf{x}_j + c)$$

*Hyperbolic tangent kernel equation*

, for some (not every)  $\kappa > 0$  and  $c < 0$ .

## 4.6. Sigmoid kernel

We can use it as the proxy for neural networks. Equation is

$$k(x, y) = \tanh(\alpha x^T y + c)$$

*Sigmoid kernel equation*

## 4.7. Bessel function of the first kind Kernel

We can use it to remove the cross term in mathematical functions. Equation is :

$$k(x, y) = \frac{J_{v+1}(\sigma \|x - y\|)}{\|x - y\|^{-n(v+1)}}$$

*Equation of Bessel function of the first kind kernel*

where  $J$  is the Bessel function of first kind.

## 4.8. ANOVA radial basis kernel

We can use it in regression problems. Equation is:

$$k(x, y) = \sum_{k=1}^n \exp(-\sigma (x^k - y^k)^2)^d$$

*ANOVA radial basis kernel equation*

## 4.9. Linear splines kernel in one-dimension



It is useful when dealing with large sparse data vectors. It is often used in text categorization. The splines kernel also performs well in regression problems. Equation is:

$$k(x, y) = 1 + xy + xy \min(x, y) - \frac{x + y}{2} \min(x, y)^2 + \frac{1}{3} \min(x, y)^3$$

*Linear splines kernel equation in one-dimension*

If you have any query about SVM Kernel Functions, So feel free to share with us. We will be glad to solve your queries.