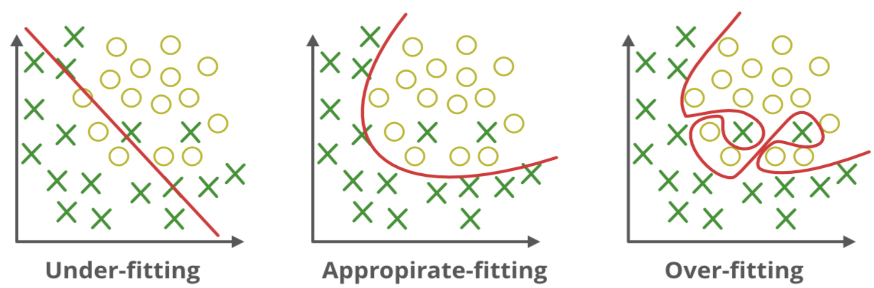
Cs19btech11020
Sharanya Gupta

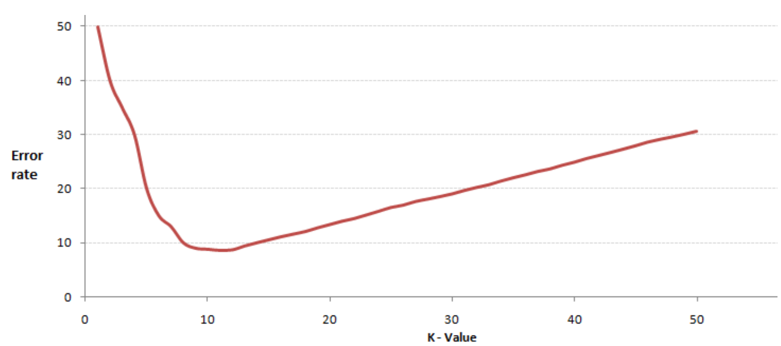
1.

a. The training error in a KNN classifier decreases as we change k from n to 1, the training error is in fact zero when k=1 as the model overfits on the training data. When k=1, the nearest neighbour is the point itself present in training data so it will never make any error and the training error is 0.

b. The generalization error in a KNN classifier decreases as we increase k from 1 , it reaches optimum and again increases.
When k is small; the model overfits and the generalization error is high, because the data is overlapped and the points in the overlapped region have a high chance of misclassification.
As we increase k,this error decreases and reaches a minimum(the model tries to tune the boundary without overfitting) and increases as we increase k further(the model underfits as a large neighbourhood is considered, the capacity to classify is lost) .

K → n n > k >1 k → 1





c. KNN is undesirable when the input dimension is too high. The KNN classifier classifies based on density of class points in the neighbourhood. But if the input dimension is too high, the points are scattered among classes and the density decreases (insufficient data), thereby making the model unable to make correct classification. Moreover if the input dimension is too high the computation also increases making the model slower.

d. It depends on the data; if data is one dimensional, then there is a possibility that the 1nn classifier and complete decision tree give the same results.

In more than 1 dimension, **we can't** . The decision boundary for nearest neighbours are Voronoi diagrams, and are in arbitrary directions. Whereas the decision boundary of a decision tree are lines parallel to the axis. So we can't recreate a 1NN classifier using a univariate decision tree.

2.

2) a) Variance $C_1 = 0.0149$

Variance $C_2 = 0.0092$

$$\text{Mean for } C_1 \ (\mu_1) = \frac{0.5+0.1+0.2+0.4+0.3+0.2+0.2+0.1+0.35+0.25}{10}$$

$$= \frac{2.6}{10} = 0.26$$

$$\sigma_1 = \sqrt{Var} = \sqrt{0.0149} = 0.122$$

$$\text{Distribution of a class} = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\text{Distribution of class}_1 = 3.27\, e^{-1/2\left(\frac{x-0.26}{0.122}\right)^2}$$

μ_2 of $C_2 = 0.8625$

$\sigma_2 = 0.0959$

$$\text{Distribution of class 2} = 4.16\, e^{-1/2\left(\frac{x-0.8625}{0.0959}\right)^2}$$

P_1 = prob of $C_1 = \frac{10}{14} = 0.714$

P_2 = prob of $C_2 = \frac{4}{14} = 0.286$

Given, $x = 0.6$

$$P(x/C_1)P(C_1) = 3.27e^{-1/2\left(\frac{0.6-0.26}{0.122}\right)^2} \times 0.714 = 0.048$$

$$P(x/C_2)P(C_2) = 4.16e^{-1/2\left(\frac{0.6-0.8625}{0.0959}\right)^2} \times 0.286 = 0.028$$

from bayes theorm

$$P(C_1/x) = \frac{P(x/C_1)P(C_1)}{P(x/C_1)P(C_1)+P(x/C_2)P(C_2)} = \frac{0.048}{0.048+0.023} = \underline{0.6315}$$

(b) Given

x = (goal, football, golf, defence, offence, wicket, office, strategy)

To find prob that $x = (1,0,0,1,1,1,1,0)$ is about politics

$$P(C_P/\vec{x}) = \frac{P(\vec{x}/C_P)\,P(C_P)}{P(\vec{x}/C_P)P(C_P) + P(\vec{x}/C_S)P(C_S)}$$

$$P(\vec{x}/C_S) = P(x_1/C_S) \times P(x_2/C_S) \cdots P(x_8/C_S)$$

$$\Rightarrow P(x_7 = 1/C_S) = 0 \Rightarrow P(\vec{x}/C_S) = 0$$

$$\Rightarrow P(C_P|\vec{x}) = 1$$