1. K-NN (8 marks): In k-nearest neighbors (k-NN), The classification is achieved by majority vote in the vicinity of data. Given n points, imagine two classes of data each of n/2 points, which are overlapped to some extent in a 2-dimensional space

(a) (1 mark) Describe what happens to the training error (using all available data) when the neighbor size k varies from n to 1.

Solution:

Training error is the error which is obtained by testing the model with the training data itself. Under such circumstances, when k=1 in k-NN model, the test point will be the training point itself. Hence when k=1, the error will be zero. But when k increases, we may not guarantee data points with class label same as the test point at the vicinity ∴ The error increases. Hence in short,

* The training error is zero when k=1
* The training error increases as the k increases

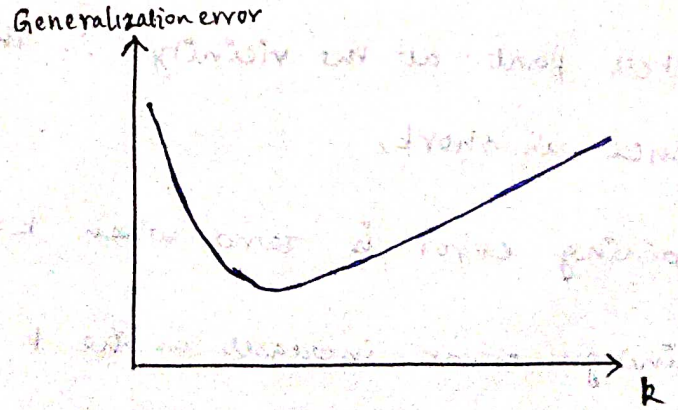
b) Predict and explain with a sketch how the generalization error (ex: holding out some data for testing) would change when k varies ? Explain your reasoning.

* The generalization error is the error ~~the~~ we get on the test data

* when k value is very small, the test point may be suceptible to noise points and hence the error will be high.

* when k value is large (≈ size of data set), then the model simply returns the majority class as the classification irrespective of the location of the test point. Hence in this case also, the error is very high.

* we get the least error for some value of k which is neither too small nor too large



c) Give two reasons (with sound justification) why k-NN may be undesirable when the input dimension is high.

* kNN is a transductive model which requires the knowledge of entire data set for each classification(ie) We need to calculate the distances from the test point to all the points in the data set to determine the nearest neighbors. With the increase in input dimension, the size of the data space increases and hence computing distances is a computationally intensive task which requires sophisticated hardware.

* Also, when the data space increases at the higher dimensions, the amount of data needed to maintain density also increases. Without dramatic increase in the size of the data set, K-nearest neighbors loses all the predictive power. (ie) The concept of distance disintegerates at larger dimensions.

d) Is it possible to build a univariate decision tree (with decisions at each node of the form "is $x > a$", "is $x < b$", "is $y > c$" (or) "is $y < d$" for any real constants a, b, c, d) which classifies exactly similar to a 1-NN using the Euclidean distance measure? If so, explain how. If not, explain why not

Solution

1 d)

- No. It is not possible Because,

+ For the 1NN scheme, the decision boundary could be identified with the help of voronoi diagram. There we could observe that the boundaries are slant and not parallel to either x (or) y axes.

+ But the decision at each node for the decision tree is given as lines parallel to x axis and y axis [i.e $x>a$, $x<b$, $y>c$, $y<d$].

+ Hence for the decision tree to approximate the gradient, it could take uncountable number of decisions which is not practically feasible.

2. Bayes classifier:

given,

$$\sigma_1^2 = 0.0149 \quad ; \quad \sigma_2^2 = 0.0092$$

$$P(C_i/x) = \frac{P(x/C_i) \cdot P(C_i)}{P(x)}$$

$$P(C_1/0.6) = \frac{P(0.6/C_1) \cdot P(C_1)}{P(0.6/C_1) \cdot P(C_1) + P(0.6/C_2) \cdot P(C_2)}$$

Since we need to fit a one dimensional Gaussian,

$$p(x/C_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{\frac{-(x-\mu_i)^2}{2\sigma_i^2}}$$

$$\mu_i = \frac{1}{N_i} \sum_{i=1}^{N_i} x_i$$

$$\mu_1 = \frac{1}{10}(0.5 + 0.1 + 0.2 + 0.4 + 0.3 + 0.2 + 0.2 + 0.1 + 0.35 + 0.25)$$

$$\boxed{\mu_1 = 0.26}$$

$$\mu_2 = \frac{1}{4}(0.9 + 0.8 + 0.75 + 1.0) = 0.8625$$

$$\Rightarrow \boxed{\mu_2 = 0.8625}$$

$$P(\text{class } 1) = \frac{\text{No. of samples in class 1}}{\text{Total No. of samples}} = \frac{10}{14} = 0.7142$$

$$P(\text{class } 2) = P(C_2) = 1 - P(C_1) = 0.2857$$

Now, $$p(0.6/C_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}}$$

$$p(0.6/C_1) = \frac{1}{\sqrt{2\pi} \times (0.122)} e^{\frac{-(0.6-0.26)^2}{2 \times 0.0149}}$$

$$\Rightarrow \boxed{P(0.6/C_1) = 0.06758}$$

$$P(0.6/C_2) = \frac{1}{\sqrt{2\pi} \times 0.0959} e^{\frac{-(0.6-0.8625)^2}{2 \times 0.0092}}$$

$$\Rightarrow \boxed{P(0.6/C_2) = 0.07833}$$

Now,

$$P(C_1 / 0.6) = \frac{0.06758 \times 0.7142}{(0.06758 \times 0.7142) + (0.09833 \times 0.2857)}$$

$$\Rightarrow \boxed{P(C_1 / 0.6) = 0.632}$$

2b)

$$x = (1, 0, 0, 1, 1, 1, 1, 0) = (A_1, A_2 \ldots\ldots, A_8)$$

$$P(\text{politics} / x) = ?$$

$$P(\text{politics} / x) = \frac{P(x / \text{politics}) \cdot P(\text{politics})}{P(x)}$$

$$= \frac{P((1,0,0,1,1,1,1,0) / \text{politics}) \cdot P(\text{politics})}{P(1,0,0,1,1,1,1,0)}$$

Let 1 represent ⇒ yes

0 represent ⇒ No.

Then, $x = [$Goal = yes, Football = No, Golf = No, Def. = No, offence = yes, Wicket = Yes, office = Yes,

Goal	Politics	Sport
yes	2/6	4/6
No	4/6	2/6

Football	Politics	Sports
yes	1/6	4/6
No	5/6	2/6

Golf	Politics	Sports
yes	1/6	1/6
No	5/6	5/6

Defence	Politics	Sports
Yes	5/6	4/6
No	1/6	2/6

offence	Politics	Sports
yes	5/6	1/6
No	1/6	5/6

Wicket	Politics	Sports
yes	1/6	1/6
No	5/6	5/6

Office	Politics	Sports
Yes	4/6	0/6
No	2/6	6/6

Strategy	Politics	Sports
Yes	5/6	1/6
No	1/6	5/6

$$p(x/\text{politics}) = 2/6 \times 5/6 \times 5/6 \times 5/6 \times 5/6 \times 1/6 \times 4/6 \times 1/6$$

$$= \frac{50 \times 25 \times 4}{6^8} = 0.0029$$

$$p(x/\text{sports}) = 4/6 \times 2/6 \times 5/6 \times 4/6 \times 1/6 \times 1/6 \times 0/6 \times 5/6$$

$$= 0.$$

$$p(\text{politics}) = 1/2 \qquad p(\text{sports}) = 1/2$$

$$p(\text{politics}/x) = \frac{p(x/\text{politics}) \cdot p(\text{politics})}{p(x/\text{politics}) \cdot p(\text{politics}) + p(x/\text{sports}) \cdot p(\text{sports})}$$

Strategy = No)

$$= \frac{0.0029 \cdot (1/2)}{0.0029 \cdot (1/2) + 0 \cdot (1/2)}$$

$$\boxed{p(\text{politics}/x) = 1}$$

Submitted by;
PRASANNA KUMAR. R
sm21mtech14001