AI21MTECH14011

Foundations of Machine Learning.

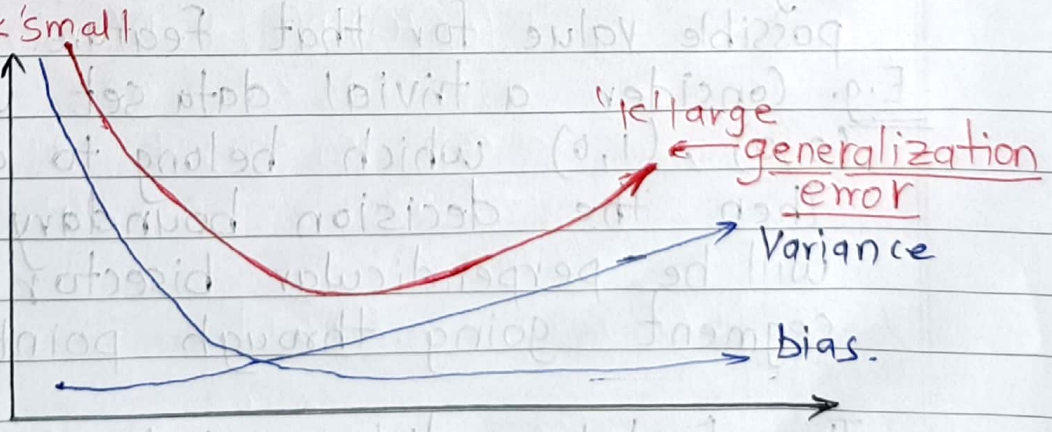
Assignment-1.

1. k-NN

a) Training error occurs when training set is loaded to k-NN as test set. By changing the values of k from n to 1, it keeps track of all accuracy values. At one point of k, one can get the least loss, then it can be conclude which class the given query point belongs to.

b) i) Initially k is very small then error is high because there are few points. As k increase, the error decreases. But as k tends to cover all points, then error increases again as the point gets classified with class of majority points. Hence Generalizatn error increases again

ii) In underfitting, higher the bias, more error between predictions and ground truth. In overfitting, variance is higher, hence error is more due to small fluctuations in the dataset.



c) i) In high dimensions, the points which are similar, may have large distances between them. The points can be far away from each other. Mostly, it suffers from a problem of irrelevant features.

ii) In higher dimensions, as k varies we can classify according to features but speed will be very slow. and hence computation cost will increase.

iii) Increasing dimensions, means increasing the number of distance functions exponentially. Hence number of outliers can be increased. Due to this, knn loses its predictive power. Hence k-NN may be undesirable when the input dimension is high.

d) Yes.
In univariate decision tree, each test uses one input dimension or feature. The tree contains only one branch, hence it indicates each possible value for that feature.
E.g. Consider a trivial data set with points (0,0), (1,0) which belong to different class. Then the decision boundary of 1-NN will be perpendicular bisector of the segment going through point $(\frac{1}{2}, 0)$.
The Euclidean distance between them is same.

The line will be parallel to Y-axis which separates the instances of two different classes into two different regions.

So, one dimentional data with proper separation between two classes produce same decision boundary for 1-NN and univariate decision tree.

2. Bayes Classifier.

a) $C_1 \Rightarrow$ Class 1, $C_2 \Rightarrow$ Class 2

$\sigma_1^2 = 0.0149$, $\sigma_2^2 = 0.0092$

$P(C_1) = ?$, $P(C_2) = ?$, $P(C_1/0.6) = ?$

$n(C_1) = 10$, $n(C_2) = 4$

i) $$P(C_1) = \frac{n(C_1)}{n(C_1)+n(C_2)} = \frac{10}{14}$$

$$\boxed{P(C_1) = 0.7142}$$

ii) $$P(C_2) = \frac{n(C_2)}{n(C_1)+n(C_2)} = \frac{4}{14}$$

$$\boxed{P(C_2) = 0.2857}$$

Classification probability using Bayes Thm.

$$P(C_i/x) = \frac{P(x/C_i)\,P(C_i)}{P(x)} = \frac{P(x/C_i)\cdot P(C_i)}{\sum_{N=1}^{n} P(x/C_n)\,P(C_n)}$$

In this problem, the Gaussian Likelihood is given by, $$P(x/C_i) = \frac{1}{\sqrt{2\pi\,\sigma_i^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$

Mean ⇒

$$\mu_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_i = \frac{0.5+0.1+0.2+0.4+0.3+0.2+0.1+0.2+0.35+0.25}{10}$$

$\therefore \mu_1 = 0.26$

$$\mu_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} x_i = \frac{0.9+0.8+0.75+1}{4} = 0.8625$$

Variance ⇒

$$\sigma_i^2 = \frac{1}{N_i}\sum_{i=1}^{N} (x_i - \mu)^2$$

$\therefore \sigma_1^2 = 0.0149$, $\sigma_2^2 = 0.0092$

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

$$P(0.6/C1) = \frac{1}{\sqrt{2\pi \times 0.0149}} e^{-\frac{1}{2}\left(\frac{0.6-0.26}{\sqrt{0.0149}}\right)^2}$$

$$= 0.0675$$

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

$$= 0.098$$

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

$$= \frac{0.0675 \times 0.7142}{0.0675 \times 0.7142 + 0.098 \times 0.2857}$$

$$= \frac{0.048}{0.076} = 0.6341$$

$\therefore$ **$P(C_1/0.6) = 0.6341$**

$\therefore$ Ans. i) $P(C_1) = 0.7142$ ii) $P(C_2) = 0.2857$

iii) $P(C_1/0.6) = 0.6341$

b) Sport or politics

$X = (1, 0, 0, 1, 1, 1, 1, 0)$

Probability = prob.(politics) × [prob.(goal/politics) .... prob(strategy/politics)]

Here we have 8 attributes. $(A_1, A_2 \ldots A_8)$
But the vector X is given to us has only 5 attributes

So P(goal/politics) · p(defense/politics) · p(offense/poli) p(wicket/politics) · p(office/politics)

Here, we have to calculate maximum likelihood,
∴ hence nelglecting p(politics)

$$\Rightarrow \frac{2}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{1}{6} \times \frac{4}{6} = 0.0257.$$

$$\therefore \boxed{p(A_1, A_4, A_5, A_6, A_7 / \text{politics}) = 0.0257}$$

3] Decision Trees:

a] Code is implemented in DecisionTree-1.

b] Decision tree is evaluated using 10 fold cross validation. Data is randomly shuffled using the function random.shuffle(). Accuray for each 'k' has noted. Finally average accuracy is computed. It was come out as 78.36%.

c] Improved the decision tree algorithm using i) Gini index ii) Pruning the tree.

Evaluated it also using 10-fold cross validation Accuracy for each 'k' has been noted. Average accuracy is increased from 78.36% to 81.16%.

In this part, results are obtained much faster than entropy method. i.e. computation is fast. Basically, pruning reduces the complexity of the tree. It limits the depth of tree. Gini index selects the best features in the tree, hence results obtained are better as compared to entropy. Hence, the accuracy increases in improved algorithm.