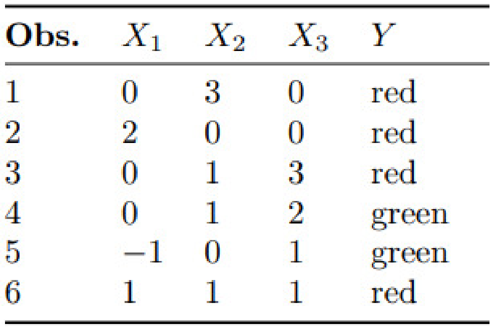
Q6

Rahul Atre

2023-09-30

Q6 [Statistical Learning]

The table below provides a training dataset containing six observations, three predictors, and one qualitative response variable.



Obs.	X_1	X_2	X_3	Y
1	0	3	0	red
2	2	0	0	red
3	0	1	3	red
4	0	1	2	green
5	-1	0	1	green
6	1	1	1	red

Suppose we wish to use this dataset to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbours.

1. Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

The Euclidean distance between two points in three dimensions is defined as follows:

$$d(V_1, V_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

for $V_1 = (x_1, y_1, z_1)$ and $V_2 = (x_2, y_2, z_2)$.

In this case, the test point is the origin, so let $V_2 = (0, 0, 0)$. The formula simplifies to:

$$d(V_1, V_2) = \sqrt{x_1^2 + y_1^2 + z_1^2}$$

Now, we can compute the distance for each observation V_1 and the test point V_2.

Obs. 1: $d((0,3,0), V_2) = \sqrt{0^2 + 3^2 + 0^2} = 3$
Obs. 2: $d((2,0,0), V_2) = \sqrt{2^2 + 0^2 + 0^2} = 2$
Obs. 3: $d((0,1,3), V_2) = \sqrt{0^2 + 1^2 + 3^2} = \sqrt{10} \approx 3.162$
Obs. 4: $d((0,1,2), V_2) = \sqrt{0^2 + 1^2 + 2^2} = \sqrt{5} \approx 2.236$
Obs. 5: $d((-1,0,1), V_2) = \sqrt{(-1)^2 + 0^2 + 1^2} = \sqrt{2} \approx 1.414$
Obs. 6: $d((1,1,1), V_2) = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3} \approx 1.732$

2. What is the prediction with k = 1? Why?

For a prediction with k=1, we take the single nearest neighbor from the training dataset, which in this case is X_5, $\sqrt{2}$ distance away from the origin. Since the observation's (X_5) response is green, the prediction for the test point will be green.

3. What is our prediction with k = 3? Why?

For a prediction with k=3, we take the three closest neighbours from the training dataset. In our case, the three closest (in order) are X_5 with distance $\sqrt{2}$, X_6 with distance $\sqrt{3}$, and X_2 with distance 2 from the test point. Observations X_2, X_6 have a prediction of red, and X_5 has a prediction of green. Since the total red predictions are more than green, we conclude that the overall prediction for the test point will be red.

4. If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for k to be large or small? Why?

If the Bayes decision boundary in this problem is highly non-linear, we would expect the best value for k to be small because a larger value of k would end up producing a decision boundary that is more linear. This makes the classifier inflexible and inaccurate since it is averaging out the predictions of the nearest neighbours with a wider scope. As k decreases in size, the classification can follow the non-linear boundary more closely, thus leading to better predictions of the test points.