Predictive Modelling Tutorial 2: kNN

Dr Liew How Hui

Jan 2021

Tut 2: kNN

kNN is discriminative, non-parametric predictive model

- For kNN classifier, the prediction is

$$\hat{h}(\boldsymbol{x}) = \underset{j \in \{1,\cdots,K}{\operatorname{argmax}} \frac{1}{k} \sum_{\boldsymbol{x}_i \in N(\boldsymbol{x})} I(y_i = j)$$

- For kNN regressor, the prediction is

$$\hat{h}(\boldsymbol{x}) = \frac{1}{k} \sum_{(\boldsymbol{x}'', y'') \in N(\boldsymbol{x})} y''.$$

Tutorial 3, Q3

The table on the right provides a training data set containing six observations, three predictors and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbours.

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

Tutorial 3, Q3 (cont)

1. Compute the Euclidean distance between each observation and the test point (TP).
2. What is our prediction with $k = 1$? Why?
3. What is our prediction with $k = 3$? Why?
4. If the Bayes decision boundary in this problem is highly non-linear, then would we expect the optimum value for k to be large or small? Why?

Tutorial 3, Q3 (cont)

5. By considering X_1 and X_2 only, sketch the 3-nearest neighbours decision boundary for range $-1 \leq X_1 \leq 3$ and $-1 \leq X_2 \leq 3$, with the distance measure used in (a). Assume that X_1 and X_2 can only take integer values.