Tut 2: kNN

Jan 2022

kNN Classifier

kNN is discriminative, non-parametric predictive model

- For kNN classifier, the mathematical formulation 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 mathematical formulation is

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

One popular choice of distance in kNN is the *Minkowski distance*:

$$d(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_r = \left(\sum_{i=1}^{p} |x_i - z_i|^r \right)^{\frac{1}{r}}, \quad \boldsymbol{x},\ \boldsymbol{z} \in \mathbb{R}^p. \tag{2.1}$$

Note that $\| \cdot \|^r$ is called the ℓ^r norm.

When $r = 1$, we have the *Manhattan distance*:

$$\|\boldsymbol{x} - \boldsymbol{z}\|_1 = |x_1 - z_1| + |x_2 - z_2| + \cdots + |x_p - z_p|.$$

When $r = 2$, we have the *Euclidean distance*:

$$\|\boldsymbol{x} - \boldsymbol{z}\|_2 = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_p - z_p)^2}.$$

There are other distance / dissimilarity functions which are used in specific cases:

- Gower
- Tanimoto
- Jaccard
- Mahalanobis

1. The given table 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

(a) Compute the Euclidean distance between each observation and the test point (TP).

(b) What is our prediction with $k = 1$? Why?

(c) What is our prediction with $k = 3$? Why?

(d) 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?

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

More Performance Evaluation

2. What are the advantages of k-fold cross validation relative to

   (a) Validation set approach

   (b) Leave-one-out cross validation (LOOCV)

3. (May 2019 Final Q3)

   (a) Supervised learning includes classification and regression.

      (i) State the difference between classification and regression in term of response variable. (1 mark)

      (ii) Explain the sampling methods used in splitting data for classification and regression respectively. (4 marks)

   (b) (i) State an issue that comes along with split validation, which can be overcome by using cross validation. (1 mark)

      (ii) Describe the process of a 5-fold cross validation. (4 marks)

(c) A sample of 500 males and 800 females had been collected to test on a model of gender prediction. The model resulted that 380 males and 510 females were predicted correctly.

(i) Assume male as positive class and female as negative class, calculate the count of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) for the model's result.

(2 marks)

(ii) Construct the confusion matrix for the model. State the classification error, specificity and sensitivity of the model. (4 marks)

(iii) Compare the recall and precision for both male and female. Interpret your results.

(4 marks)