k-Nearest Neighbors

Exercises

1. *Briefly* explain the basic idea behind the k-Nearest Neighbors model to someone who does not have a technical background.

2. For the following training data, compute the output for a test observation $\mathbf{x}_{test} = (4, 2)$ using the Manhattan distance function (1-norm), and 1 and 3 nearest neighbors. To make a prediction, use the mean of the nearest neighbors.

x	**Target**
(1,6)	7
(2,4)	8
(3,7)	16
(6,8)	44
(7,4)	50
(8,5)	68

Suppose you have a dataset of dummy variables - below consider the simple example where only one variable is present:

Obs	**Married**	Single	Other
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1

Putting this data into a matrix, we have:

$$\mathbf{X} = \begin{matrix} obs1 \\ obs2 \\ obs3 \\ obs4 \\ obs5 \\ obs6 \end{matrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Compute the matrix $\mathbf{X}\mathbf{X}^T$. How might this computation help you to create a distance matrix for categorical attributes, avoiding loops?

3. List 2 major advantages of the kNN algorithm. List 2 major disadvantages.

List of Key Terms

k

Nearest Neighbors

Euclidean Distance

Manhattan Distance

Matching Coefficients

kNN advantages

kNN disadvantages

antecedent

consequent