Homework Set 4, CPSC 6430/4430

LastName, FirstName

Due 04/03/2024, 11:59PM EST

Implementation and Application

Please refer to Jupyter Notebook.

K-NN and Decision Boundary

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 data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors (K-NN).

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

2. What is our prediction with $K = 1$ and why?

3. What is our prediction with $K = 3$ and why? Please determine the probability assigning to each class (Red or Green) by using uniform weight or distance weight (where each neighbor's weight is determined by $w_i = \frac{exp\{-d_i^2\}}{\sum_{j=1}^{K} exp\{-d_j^2\}}$ and d_i denotes the Euclidean distance of testing data to i-th nearest neighbor data) respectively.

4. Now if we only use the first two features X_1 and X_2, please scatter plot the six observation points and plot the contour for decision boundary by referring to here.

Objective for k-means

Given the objective of k-means:

$$\min_{\mu_1,\mu_2,\ldots,\mu_k} \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \|x - \mu_i\|_2^2. \tag{1}$$

Please prove that k-means algorithm will monotonically make the above objective non-increasing.

If we change the objective to:

$$\min_{\mu_1,\mu_2,\ldots,\mu_k} \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \|x - \mu_i\|_1, \tag{2}$$

please design an algorithm (similar to k-means) to make the objective non-increasing and prove it.