KNN for Classification

Data: (X_i, y_i), $i = 1, 2, \cdots n$ $\quad y_i \in \{1, 2, \cdots L\}$

Goal: predict y_0 from a new obs X_0

Step 1: (Find the K-nearest neighbors)

compute the distance between X_i and X_0:

$$d_i = \|X_i - X_0\|_2 = \sqrt{\sum_{j=1}^{p} (X_{ij} - X_{0j})^2}$$

$X_i = (X_{i1}, \cdots X_{ip})$

$X_0 = (X_{01}, \cdots X_{0p})$

$X_1 = (10, 1)$, $X_2 = (8, 2)$, $X_3 = (6, 3)$

$y_1 = 1$, $y_2 = 2$, $y_3 = 2$

For $X_0 = (7, 2)$

$d_1 = \|X_1 - X_0\|_2 = \sqrt{9+1} = \sqrt{10}$

$d_2 = \|X_2 - X_0\|_2 = \sqrt{1+0} = 1$

$d_3 = \|X_3 - X_0\|_2 = \sqrt{1+1} = \sqrt{2}$

$2\text{-}NN \Rightarrow N_0 = \{2, 3\}$

$$P(y_0 = 2 \mid X = X_0) = \frac{1}{2} = \frac{1}{2} \sum_{i \in N_0} 1\{y_i = 2\}$$

$$P(y_0 = 1 \mid X = X_0) = 0 = \frac{1}{2} \sum_{i \in N_0} 1\{y_i = 1\}$$

If $P(Y=1 \mid X) = \dfrac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} > 0.5$, $\hat{Y} = 1$

$\Leftrightarrow P(Y=1 \mid X) > P(Y=0 \mid X)$

$\Leftrightarrow \hat{Y} = \underset{j}{\arg\max}\, P(Y = j \mid X)$

$X_1 = (1000, 2)$

$X_2 = (1000, 8)$

$X_3 = (900, 2)$

$d(X_1, X_2) \ll d(X_1, X_3)$