29.18 K-NN for Regression

The dataset for a binary classification task is represented as

$\mathbf{D = \{(x_i, y_i)_{i=1}^n \mid x_i \in R^d, y_i \in \{0,1\}\}}$

The dataset for a regression task is represented as

$\mathbf{D = \{(x_i, y_i)_{i=1}^n \mid x_i \in R^d, y_i \in R\}}$

Procedure of K-NN for Regression

1) Given '$\mathbf{x_q}$', find the 'K' nearest neighbors. Let them be
   $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots\ldots, (x_k, y_k)$
2) Find $\mathbf{y_q}$' from $y_1, y_2, y_3, \ldots., y_k$ using the formula
   $\mathbf{y_q' = mean(y_i)_{i=1}^k}$ **(or)** $\mathbf{y_q' = median(y_i)_{i=1}^k}$
   (The median is less prone to the outliers when compared to the mean)

Note: In classification using K-NN, we go with the majority vote, whereas in regression using K-NN, we go with either the mean or the median of the output values of the 'K' nearest neighbors.

It is recommended to go with the median, because median is less prone to the outliers/noise when compared to the mean.

AppliedRoots (Draft Copy)