Notes on K-Nearest Neighbors

February 1, 2016

1 Formal Definition

Let \( x \) be our test data point, and \( N_K(x) \) be the indices of the \( k \) nearest neighbors of \( x \) in \( \mathbb{R}^D \).

1.1 Classification

In classification, we are trying to predict some discrete target label, e.g. 1/0 for “spam” / “no spam”, or in multi-class problems, some class ID (e.g. this movie is a drama i.e. class ID = 1, action i.e. class ID = 2, indie i.e. class ID = 3, etc.) Then \( y = \arg\max_c \#(y_i = c) \) where \( c \) is the class ID. This can also be written as \( y = \arg\max_c \sum_{i \in N_K(x)} I(y_i = c) \).

1.2 Regression

In regression, we want to predict some continuous label, e.g. price of a house based on its size, year of construction, etc. Then \( y = \frac{1}{K} \sum_{i \in N_K(x)} y_i \).

2 Distance Metrics

2.1 Euclidean

For some \( x, z \in \mathbb{R}^D \), their distance is defined as \( d(x, z) = \left[ \sum_{i=1}^{D} (x_i - z_i)^2 \right]^\frac{1}{2} \).
2.2 Mahalanobis

We want to weigh different dimensions differently, e.g. because in some dimensions, points are naturally spread out, so distances are on average larmed. Then we can define a distance as

\[ d(x, z) = \sum_{i=1}^{D} \frac{(x_i - z_i)^2}{\sigma_i^2} \]

where \( \sigma_i \) is the variance in the \( i \)-th dimension.

More generally, we can write

\[ d(x, z) = (x - z)^T A (x - z) \]

where \( A \) is a diagonal matrix with the \( \sigma_i \) along the diagonal. This is an example of a Mahalanobis distance. For a distance of this form to be a Mahalanobis distance, \( A \) has to be positive semi-definite, i.e. it has to be a symmetric matrix for which \( x^T A x \geq 0 \) for all \( x \in \mathbb{R}^D \).

2.3 Minkowski

We define

\[ d(x, z) = \left[ \sum_{i=1}^{D} |x_i - z_i|^P \right]^{\frac{1}{P}}. \]

Note that if \( P = 1 \) this becomes the Manhattan distance, and if \( P = 2 \) it is the Euclidean distance.

3 Weighted K-NN

For this generalization, we will use all \( n \) training points, i.e. \( K = n \). Let \( w_i \) be the weight of the \( i \)-th instance. Note that this \( w_i \) depends on the particular query sample \( x \). One choice of a weighting function is a Gaussian kernel: \( w_i = e^{-\frac{||x - x_i||^2}{\sigma^2}} \). We need a minus sign because we want a point close to \( x \) (i.e. a point with a small distance to \( x \)) to get a high weight. Note that \( \sigma \) is the bandwidth parameter and it expresses how quickly our weight function “drops off” as points becomes further and further from the query \( x \).

3.1 Classification

The weighted prediction becomes \( y = \arg \max_c \sum_{i=1}^{n} w_i 1(y_i = c) \).

3.2 Regression

The weighted average becomes \( y = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \).
Volume of a shell with width $\epsilon$ (to be discussed on Wednesday)

The volume of a “sphere” with radius $r$ in $D$ dimensions is $K_D r^D$ (see Bishop Section 1.4). Say we have a sphere with radius 1, and a thin shell of the sphere with thickness $\epsilon$. Then $V(\text{shell}) = V(\text{sphere}) = K_D (1)^D - K_D (1-\epsilon)^D = 1 - (1-\epsilon)^D$. This tends to 1 as $D$ tends to infinity. Thus, most of the volume of the sphere is in that thin outer shell, which is counter-intuitive.

Acknowledgement: These notes are based on Dhruv Batra’s notes.