K-nearest neighbors

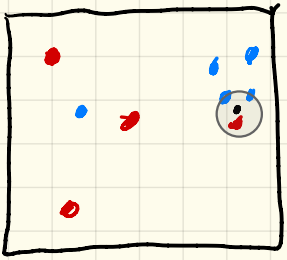
Let $D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$ w/ $\vec{x}_i \in \mathbb{R}$ our training set
$D_i \subseteq D$ pts w/ label c_i
labels $\{c_1, \dots, c_k\}$
$n_i = |D_i|$

Given test point $\vec{x} \in \mathbb{R}^d$
and $k \in \mathbb{Z}^+$ # of neighbors to consider

Let $r \in \mathbb{R}$ be the dist of $\vec{x}$'s k-th nearest neighbor in D

$$B_d(\vec{x}, r) = \{p \in \mathbb{R}^d \mid \|p - \vec{x}\| \leq r\}$$

($\|p - \vec{x}\|$: Euclidean dist)



K is the k-nearest neighbors of $\vec{x}$
$K \subseteq B(\vec{x}, r)$

$$K_i = \{\vec{x}_j \in K \mid y_j = c_i\}$$ ← all pts in radius r ball w/ label c_i

and $k_i = |K_i|$

Idea estimate $P(c_i \mid \vec{x})$
estimate the conditional prob density at $\vec{x}$

- $\vec{x}$
- c_1 = red
- c_2 = blue pts
- D_1 = set of all red pts
- D_2 = set of all blue pts
- $k = 3$
- $k_{red} = 1$
- $k_{blue} = 2$

Step 1: estimate density

Let $V = \text{vol}\left(B_d(\vec{x}, r)\right)$ ← vol of the d-dim ball centered at $\vec{x}$ w/ radius r

$= \text{vol}\left(B_{d-2}(\vec{x}, r)\right) * \frac{2\pi r^2}{d}$

w/ $\text{vol}\left(B_2(\vec{x}, r)\right) = \pi r^2$

$\text{vol}\left(B_3(\vec{x}, r)\right) = \frac{4}{3}\pi r^3$

Estimate density at $\vec{x}$

$$\hat{f}(\vec{x} \mid c_i) = \frac{\frac{k_i}{n_i}}{V} = \frac{k_i}{n_i V}$$

Step 2:

$$P(c_i \mid \vec{x}) = \frac{\hat{f}(\vec{x} \mid c_i)\,\hat{P}(c_i)}{\sum_{j=1}^{k} \hat{f}(\vec{x} \mid c_j)\,P(c_j)}$$

$\hat{P}(c_i)$ ← $\frac{n_i}{n}$

$P(c_j)$ ← $\frac{n_j}{n}$

$$\hat{f}(\vec{x} \mid c_i)\,\hat{P}(c_i) = \frac{k_i}{n_i V} \cdot \frac{n_i}{n} = \frac{k_i}{Vn}$$

$$\hat{P}(c_i \mid \vec{x}) = \frac{\frac{k_i}{nV}}{\sum_{j=1}^{k} \frac{k_j}{nV}} = \frac{\frac{1}{nV} k_i}{\frac{1}{nV} \sum_{j=1}^{k} k_j} = \frac{k_i}{\sum_{j=1}^{k} k_j} = \frac{k_i}{k}$$

$$\hat{y} = \underset{c_i}{\arg\max} \left\{ P(c_i \mid \vec{x}) \right\} = \underset{c_i}{\arg\max} \left\{ \frac{k_i}{k} \right\} = \underset{c_i}{\arg\max} \left\{ k_i \right\}$$

pick label w/ majority vote from knn