

Supervised Binary Classification

$$\text{Data set } D = \{(x_i, y_i)\}_{i=1}^N$$

$x_i \in \mathbb{R}^D$ (D random numbers)

$$y_i = \{-1, 1\}$$

"design matrix"

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nD} \end{bmatrix}$$

row: data point (x_i) vector
column: j th element of x

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

$f(x_i) \rightarrow y_i$
take (x_i) vector (row)
and go to y_i

supervised learning
(training)

have y , have x , estimate $f()$

Testing Phase

$$D_{\text{test}} = \{x_i\}_{i=1}^T$$

$$x_{\text{test}} = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

(prediction)

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_T \end{bmatrix}$$

(actual, hidden)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_T \end{bmatrix}$$

$f_{\theta}(x)$

K Nearest Neighbours [KNN]

Training: store x, y , choose K (K is positive integer)

Testing: for ($x_i \in X_{\text{test}}$) { // row in matrix

find K closest example $[x_1, x_2, \dots, x_K]$ from training set
look at corresponding $[y_1, y_2, \dots, y_K]$
choose majority

}

$$O(n \cdot d \cdot K)$$

Discuss:

Issues

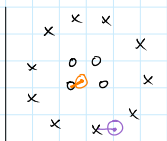
- ties
- weights

Advantages

- only 1 choice
- simple
- training is fast
- parallelizable
- fast update

Disadvantages

- large memory for data set
- choose K
- needs data "everywhere",
it may be too far
- testing slow
(curse of dimensionality)



- near o
- near x

"feature space" (two columns?)

Weird example:

$$\begin{bmatrix} x & o & x & o & x & o & x \\ o & x & o & x & o & x & o \\ x & o & x & o & x & o & x \\ o & x & o & x & o & x & o \end{bmatrix}$$

$$\begin{bmatrix} x & o & x & o & x & o & x \\ o & x & o & x & o & x & o \\ x & o & x & o & x & o & x \\ o & x & o & x & o & x & o \end{bmatrix}$$

missing data: nearest are x's,
global structure
would dictate o