Supervised binary classification

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

$x_i \in \mathbb{R}^D$ $(.5, .6, 0, 1, -18, \ldots)$

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

"Design Matrix"

each row is a datapoint

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{i1} & x_{i2} & \cdots & x_{iD} \\ \vdots & & & \\ x_{N1} & \cdots & \cdots & x_{ND} \end{bmatrix} \qquad \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

column is feature values across examples

Training

"discriminant function"

$$f(x_i) \mapsto y_i$$

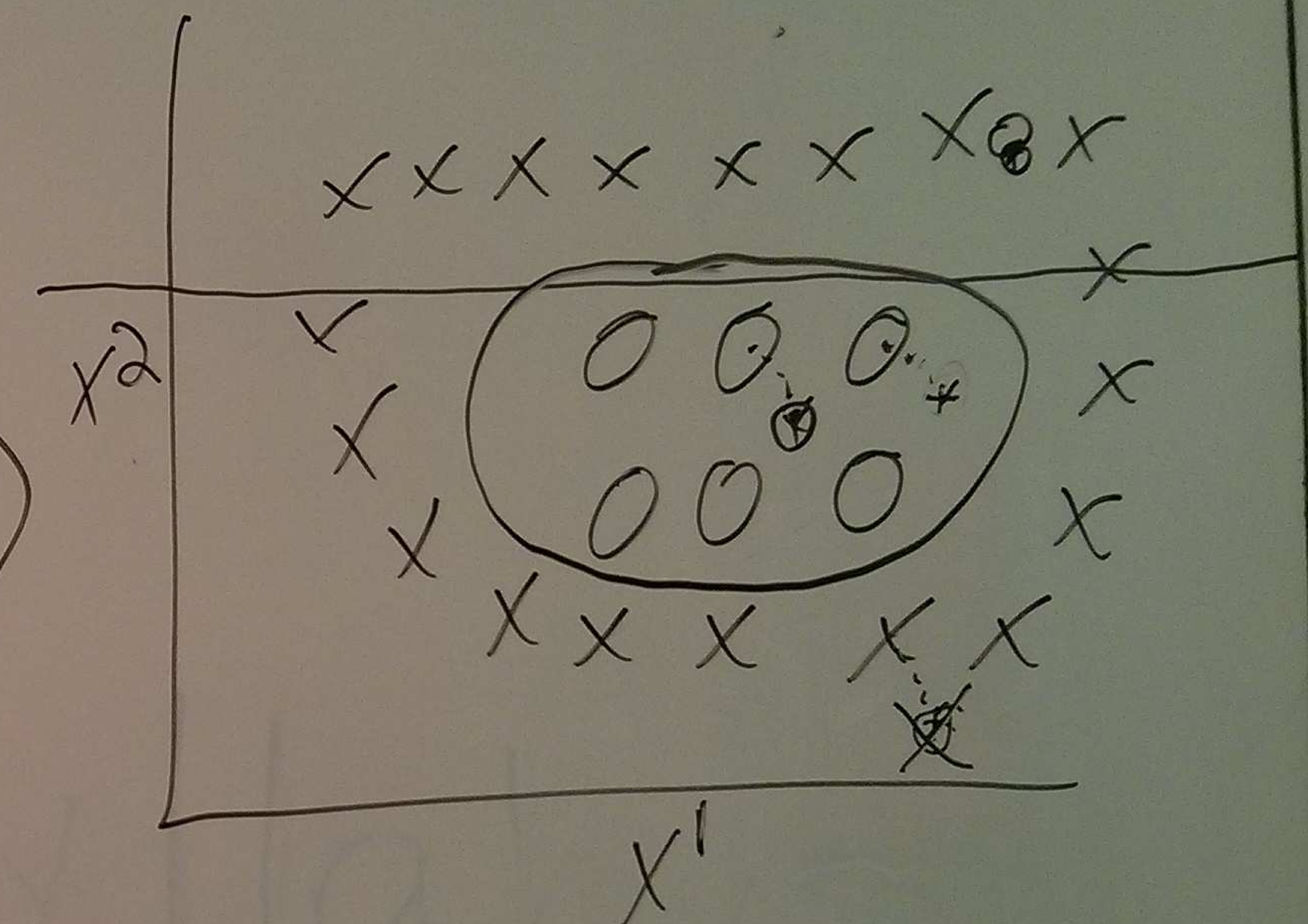
Supervised learning is about finding 'f'

Testing phase

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

$$X_{test} = \begin{bmatrix} & \\ & \end{bmatrix} \xrightarrow{f(x_i)} \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_T \end{bmatrix} \qquad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}$$

K-Nearest Neighbours



Training: store X and y, choose K.

Testing:

for $x_i \in X_{test}$

find k 'closest' examples $\{x_1, x_2, \dots, x_k\}$ from training set.

look at corresponding $\{y_1, y_2, \dots, y_k\}$ $O(NDT)$

choose majority

Issues

Ties

Weight neighbours

'closest'

Advantages

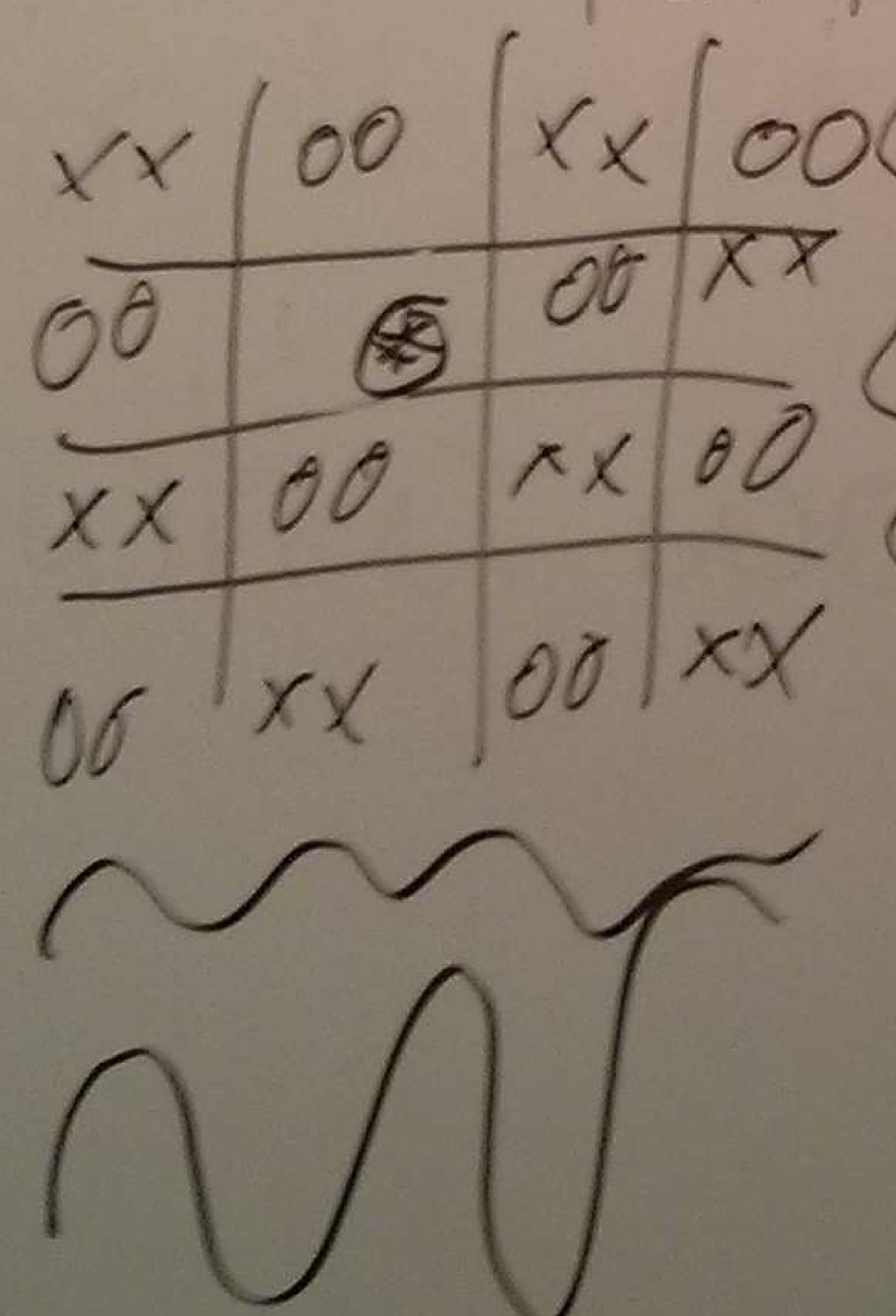
only 1 choice

simple, parrallelizable

training fast

fast updates

highly flexible/non-linear.



Disadvantages

huge memory

choose k

class imbalance

need to cover the space } 'curse of dim

testing slow

how to define 'distance'?

Uniform distance measure?