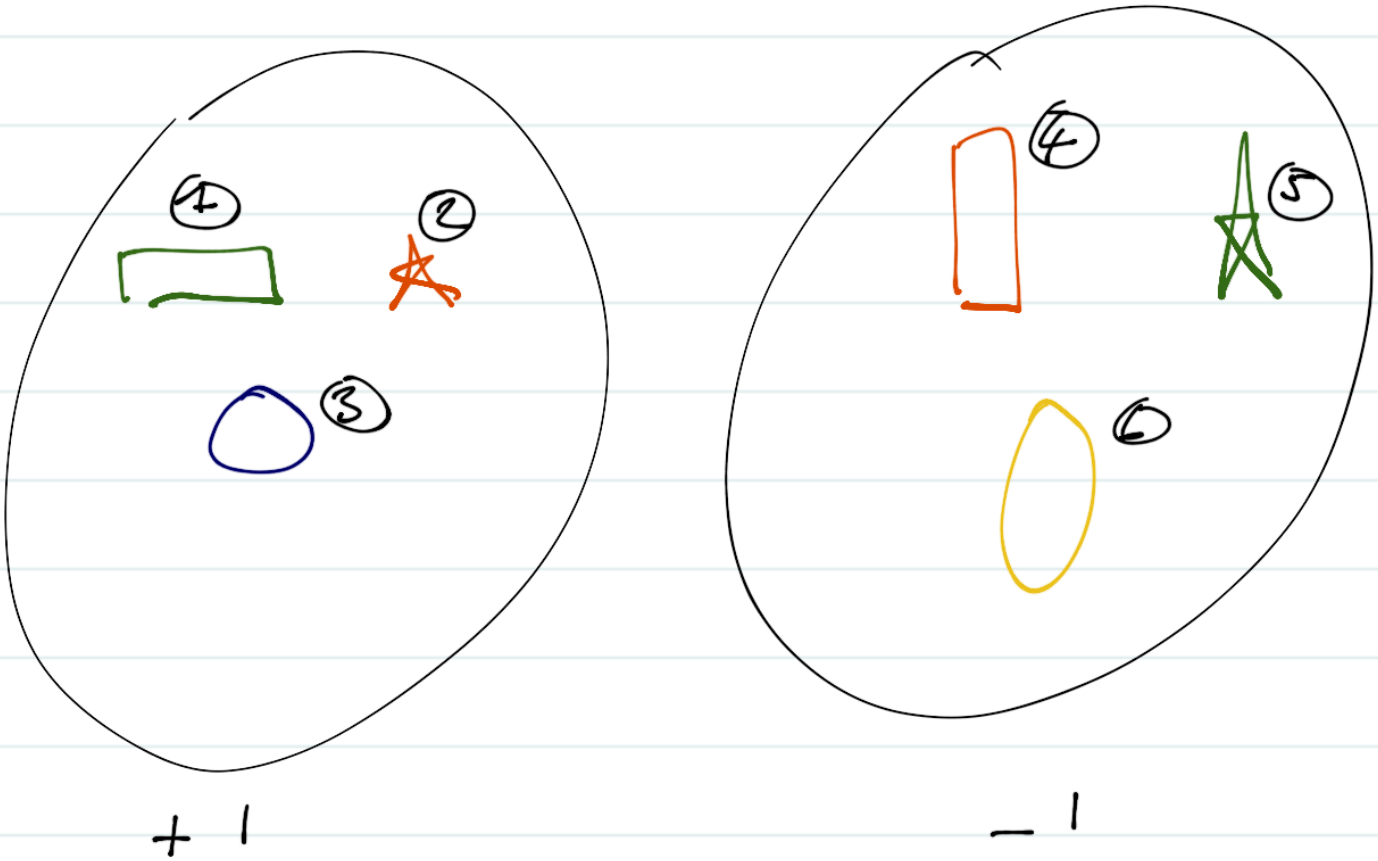
Lecture 1

TRAINING SET	datapoints	$x_1, \ldots, x_n \in \mathbb{R}^d$ $y_1, \ldots, y_n \in \{-1, +1\}$
UNSEEN DATA	new datapoint	$x \longrightarrow$ what is y?

Learn the dependency

$$x \longrightarrow y$$



+ 1 — 1

	Color	Height	Width
1	green	small	large
2	red	small	small
3	blue	small	small
4	red	large	small
5	green	large	small
6	yellow	large	small

① $x_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^3$

green	0	height	small	0
red	1		large	1
blue	2			
yellow	3	width	small	0
			large	1

$x_1, \ldots, x_n \in \mathbb{R}$

$n = 6 \quad d = 3$

$$p(y = \pm 1 \mid x, \mathcal{D})$$

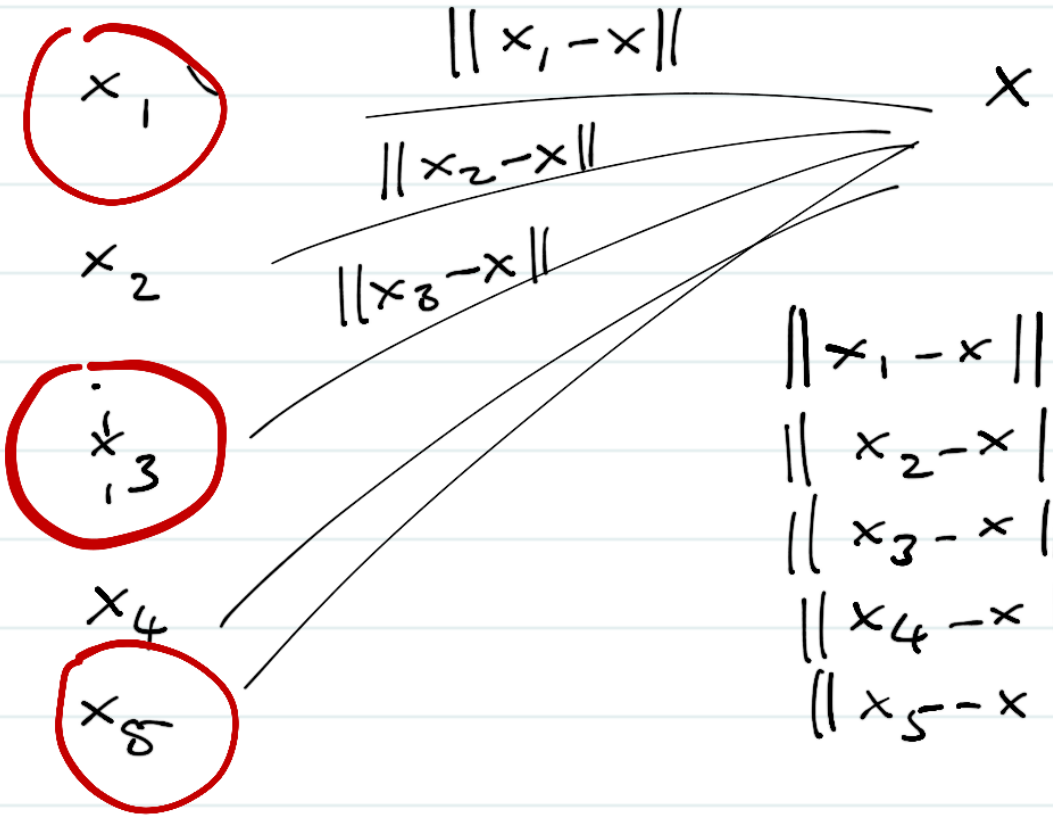
x: new datapoint

$\mathcal{D}$: dataset $\{(x_1, y_1) \dots (x_n, y_n)\}$

$$= \frac{1}{K} \sum_{i \in \mathcal{NN}_K(x, \mathcal{D})} \mathbb{1}(y_i = +1)$$

$\mathcal{NN}_K(x, \mathcal{D})$ the set of K nearest-neighbors of x in $\mathcal{D}$

$\| x - x_i \|_2$ distance between x and x_i



$\|x_1 - x\| = 0.01$ (1)

$\|x_2 - x\| = 0.9$ (5)

$\|x_3 - x\| = 0.02$ (2)

$\|x_4 - x\| = 0.6$ (4)

$\|x_5 - x\| = 0.04$ (3)

$K=3$ $\mathcal{N}_3(x, \mathcal{D}) = \{1, 3, 5\}$

$$P(y = +1 \mid x, \mathcal{D}) = \frac{1}{3} \sum_{i \in \mathcal{N}_3(x, \mathcal{D})} \mathbb{1}(y = +1)$$

$\mathcal{N}_3(x, \mathcal{D}) = \{1, 3, 5\}$

$x_1 \qquad y_1 = +1$

$x_3 \qquad y_3 = +1$

$x_5 \qquad y_5 = -1$

$$p(y=+1 \mid x, \alpha)$$

$$= \frac{1}{3} \sum_{i \in \{1,3,5\}} \mathbb{1}(y_i = +1)$$

$$= \frac{1}{3} \times \{1 + 1 + 0\}$$

$$= \frac{2}{3}$$

Binary supervised classification

TRAINING SET

$$(x_1, y_1), \ldots, (x_n, y_n)$$

example
observation
input

label
response
output

$$x_i \in \mathbb{R}^d \qquad x_i = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} \; d$$

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

$$\{0, 1\}$$

1 - Nearest - neighbor

$n \longrightarrow +\infty$ large sample
training set

$d \ll n$ asymptotic setting

error of 1-NN
number of mistakes

is

2x best error possible

Error = 1% 1-NN error = 2%

15% 30%

- (K) wrt n

$$\frac{K}{n} \xrightarrow[n \to +\infty]{} 0$$

$$K \ll n \qquad K = \frac{\sqrt{n}}{2}$$

- K - NN with $K = o(n)$

as $n \to +\infty$ K-NN achieves the best possible error

Nearest-Neighbors $n \gg 1$, $d \gg 1$

Approximate NN search

$x \in \mathbb{R}^d \quad x_i \in \mathbb{R}^d$

- $\| x - x_i \|_2$ $O(d)$ time, $O(1)$ space
- $\| x - x_i \|_2$ for all $i = \{1, \dots, n\}$

$O(nd)$

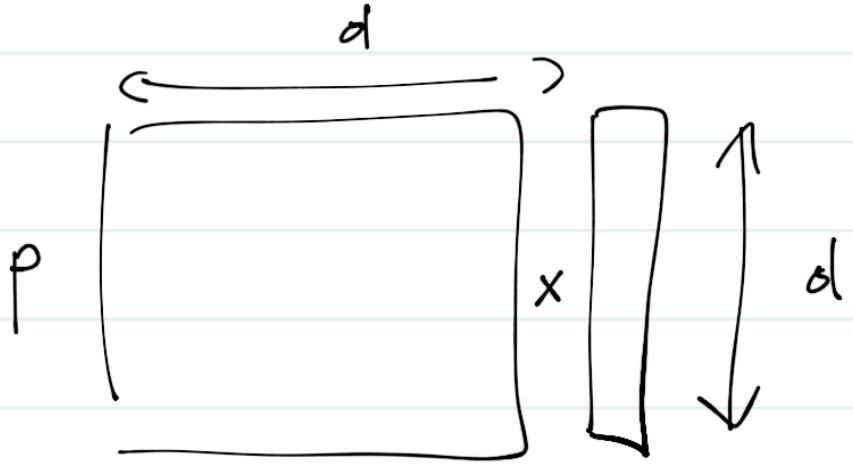
$\tilde{d}_i = \| x - x_i \|_2$

Hashing techniques

$x \in \mathbb{R}^d \qquad Ax$

$A_{ij} \overset{iid}{\sim} \mathcal{N}(0,1) \qquad \forall i,j \in \{1,\dots,p\}$

$\times \{1,\dots,d\}$



$p \ll d$

$\tilde{x} \in \mathbb{R}^p$

$d = 10^6$

$p = 10^2$

Issues with K-NN :

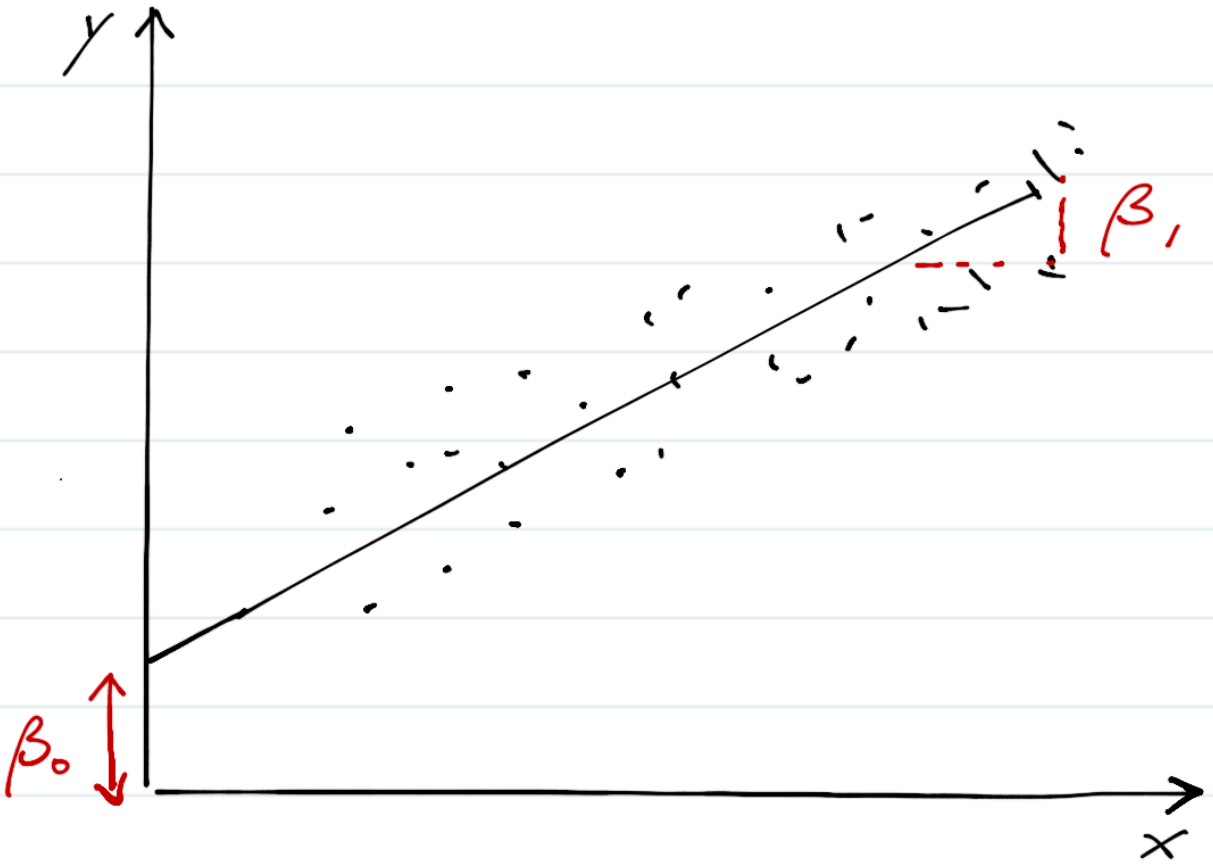
- choice of distance
- choice of k

Linear models

$$y = x^T\beta + \beta_0 \qquad x \in \mathbb{R}^{d-1}$$

$$\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{d-1} \end{pmatrix} \in \mathbb{R}^{d-1}$$

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{d-1} \end{pmatrix} \in \mathbb{R}^{d}$$



$$\min_{\beta_1, \beta_0} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(x_i^T \beta + \beta_0 \right) \right)^2 + \boxed{\frac{\lambda}{2} \|\beta\|_2^2}$$

$\lambda = 0$ Ordinary Least-Squares

$\lambda > 0$ L_2-regularized Least-Squares

$$y = x^T\beta + \beta_0$$ ← offset intercept

$$x^T\beta = \begin{bmatrix} | & | & | & | \\ x^1 & x^2 & x^3 & x^4 \\ | & | & | & | \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_4 \end{bmatrix}$$

$$= \sum_{j=1}^{4} \beta_j x^j$$

$$\underset{\beta, \beta_0}{\text{Min}} \; \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{4} \beta_j^2$$ REGULARIZATION

- $\|\beta\|_2^2$ L_2 reg° penalty
- $\|\beta\|_1$ L_1 reg penalty

$\longrightarrow$ Sparsity
select features

LASSO $\underset{\beta, \beta_0}{\text{Min}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (x_i^T \beta + \beta_0)\right)^2 + \lambda \|\beta\|_1$

Model evaluation

$$x \xrightarrow{f} y$$

Empirical risk

$$\frac{1}{n} \sum_{i \in \mathcal{D}_{TRAIN}} \ell(y_i, f(x_i))$$

↑ sum over training set

Training error

$$\frac{1}{n} \sum_{i \in \mathcal{D}_{TRAIN}} Err(y_i, f(x_i))$$

Test error

$$\frac{1}{m} \sum_{i \in \mathcal{D}_{TEST}} Err(y_i, f(x_i))$$