L6 L6看lecture6_note in PDF_expert.

① k·nearest neighbour classifier

② Perception

Structure

① training samples $X = \{(x_1,y_1),(x_2,y_2)\cdots\}$

binary classification

$x_i \in \mathbb{R}^d$. $y_i \in \mathbb{R}$ $\{0,1\}$. $\{-1,1\}$

② test point $\hat{x} \in \mathbb{R}^d$

Aim: find $\hat{y}$

find function $f: x \to y$

$\mathbb{R}^d \to \mathbb{R}$

any function | any guess is as good as any other

underlying assumptions above: Inductive bias

Applications:

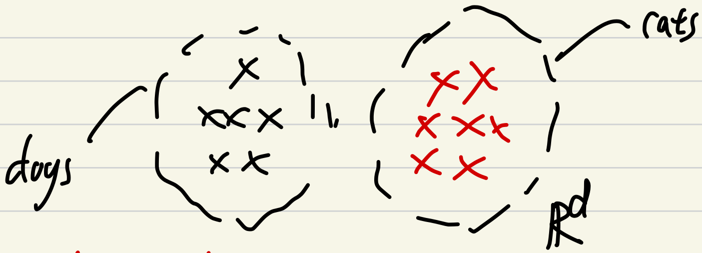
- Imagine classification | Tagging
- Spam classification
- mp3 — audio

⋮

⋮

① k-nearest neighbour

Bias/assumption: similar data points generally lie near each others



Algorithm/pesudo:

① For test point $\hat{x}$ compute $d(\hat{x}, x_i)$ $\forall i \in [1 \cdots n]$

distance between $\hat{x}, x_i$

② $i^* = \underset{\text{取最小}}{\arg\min}\, d(\hat{x}, x_i)$

$\| \ \|_2$

$\| \ \|_1$

③ $\hat{y} = y_i$

1-nearest neighbour classifier

k-nearest

Running time each test

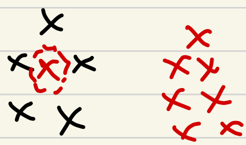
n: training points

$O(nd)$

problems?

① outliers 异常值

② label noise



③ $O(nd)$ large

木 it up generalizing very "complex" decision

概括

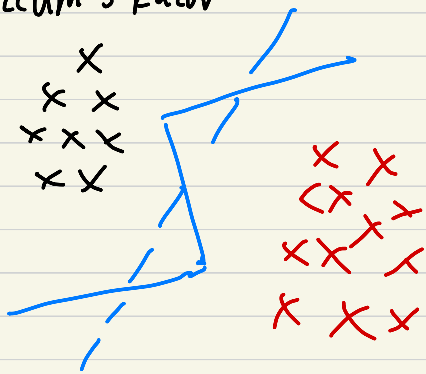
boundaries 推论.归纳

are not best [do not generalize well / overfit]

简约法则:

" Occam's Razor " 奥卡姆剃刀

同等情况下取最简单的解法(用)



for i in range(n):

$d(\hat{x}_i, x_i) = \|x - x_i\|_2$

if $d \le$ current best

i-star = i

$\hat{y} = y_i$

$k = 1$

k:

③ perception: 感知器

① early attempts to solve classification

② 1980's

Aim/goal:

learn a linear seperator b/w two classes

Setting: $X = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$

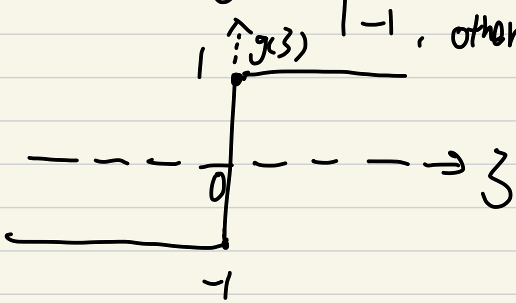
Binary: $y_i = \{-1, 1\}$

Aim

$$y = w^T x + b$$
$$= w^T x$$

problem: y is not really labels, y为 z (分类的输入)

$$g(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ -1, & \text{otherwise} \end{cases}$$



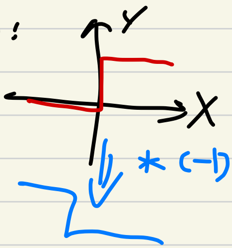
Loss: (simplest possible)

$L = 1$ if classify wrong

$L = 0$ if correct

$L = \text{step}(-y_i f(x_i))$ $\quad f(x_i) = W^T x$ 以正负号来分类：

Step:



y_i	$f(x_i) = W^T x$	Loss
1	0.5	0
-1	-0.5	0
1	-0.5	1

同号，loss=0

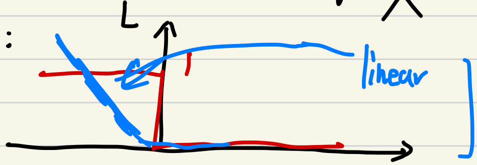
$y_i f(x_i)$

Update (gradient descent)

$$\frac{\partial L}{\partial w} = \begin{cases} 0 & \text{everywhere} \\ X & \text{label fails} \end{cases}$$

修改loss function

modification:



$$L(y, f(x_i w)) = \begin{cases} -y_i f(x_i) & \\ 0 & \text{correct} \end{cases}$$

$$\frac{\partial L}{\partial w} = \begin{cases} -y_i x_i & \\ 0 & \text{otherwise} \end{cases}$$

$w^{k+1} := w^k + \eta(y_i x_i)$

Algorithm ① : $w_0 = 0$

② : for epoch in range (10000)

for sign (x_i, y_i) in X:

错误点出现update. if $f(x_i) \neq y_i$:

update

③ all data are correctly classified : terminate.

Problems:

① previously correctly classified example can be misclassified

② No. of iteration

Theorem

assumption :

① $\|x\| \leq 1$

② $b = 0$ (bias)

③ data is perfectly seperable

If there exists a solution then perception algorithm is guaranteed to find it.

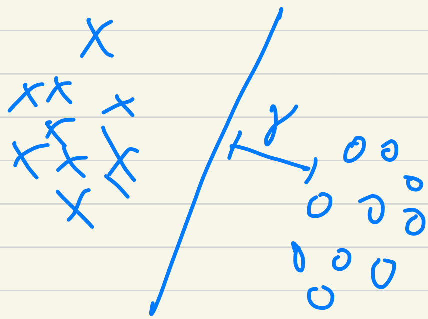
[converge to 0 error in finite time]

Proof : $w^0 = 0$ $\quad w^* = \text{optimum}$

$\|X\| \le 1$ $\quad \gamma > 0$

$X = \arg\min_i | w^T x$

y_i	$w^T x$
1	0.1
1	0.5
1	10



① $(w^*)^T w^t = (w^*)^T (w^{t-1} + y_i x_i)$

$= (w^*)^T w^{t-1} + y_i (w^*)^T x_i$

$\ge \underbrace{(w^*)^T w^{t-1}}_{\text{recursively}} + \gamma$

$(w^*)^T w^* \ge t\gamma$

② $\|w^t\|^2 = \|w^{t-1} + y_i x_i\|^2 = \|w^{t-1}\| + \|y_i x_i\| + \underline{2 y_i w^{t-1} x_i}$

$2 y_i w^{t-1} x_i < 0$

$\le \underline{\|w^{t-1}\|^2} + \underbrace{\|x_i\|^2}_{=1}$

$\leq \quad t$

③ Angle b/w optimize w^* & current solution

$$1 \geq \cos(w^*, w^t) = \frac{(w^*)^T w^t}{\|w^*\| \, \|w^t\|}$$

$(w^*)^T w^t$ → from ①

$\|w^t\|$ → from ②

$t\gamma \leq \|w^*\| \, \|w^t\|$

$\leq \|w^*\| \sqrt{t}$

$\sqrt{t} \leq \frac{\|w^*\|}{\gamma}$ $\quad t \leq \frac{1}{\gamma^2} \|w^*\|^2$

if $\|w^*\| = 1$ $\quad \boxed{t \leq 1/\gamma^2}$

No

Perception Vs. logistic :

① differences :

→ output - probability

① if solution exists, both of them will find it

② logistic regression →