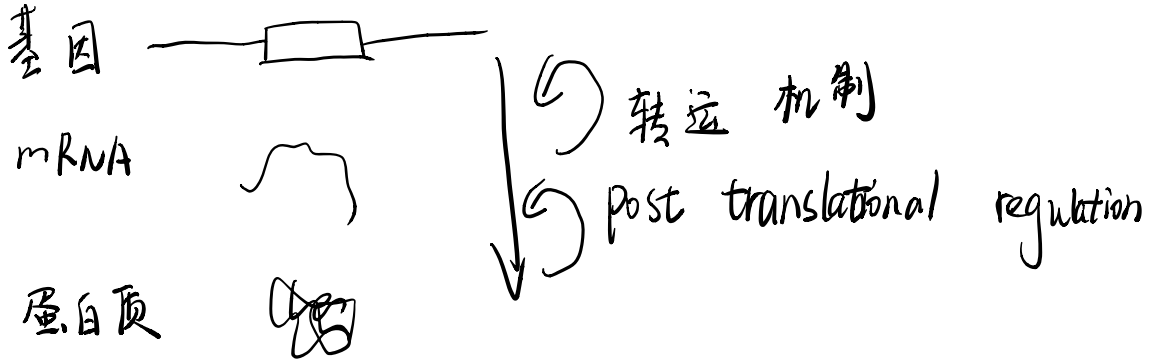
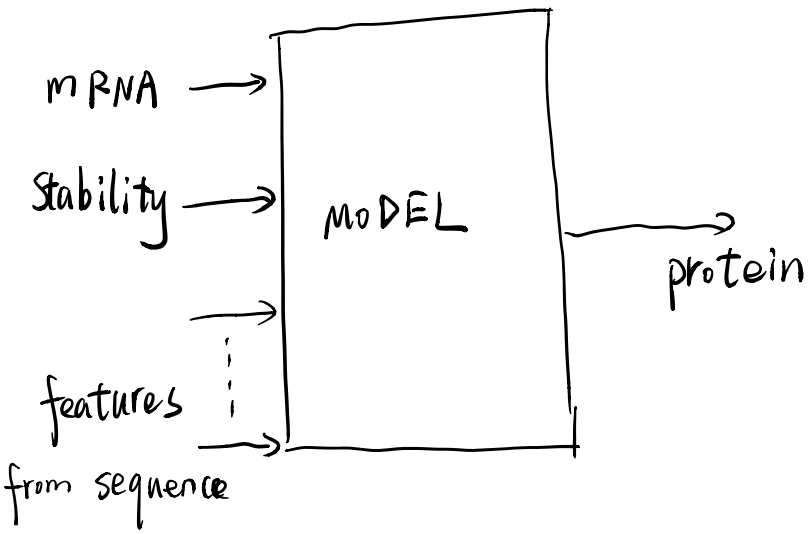
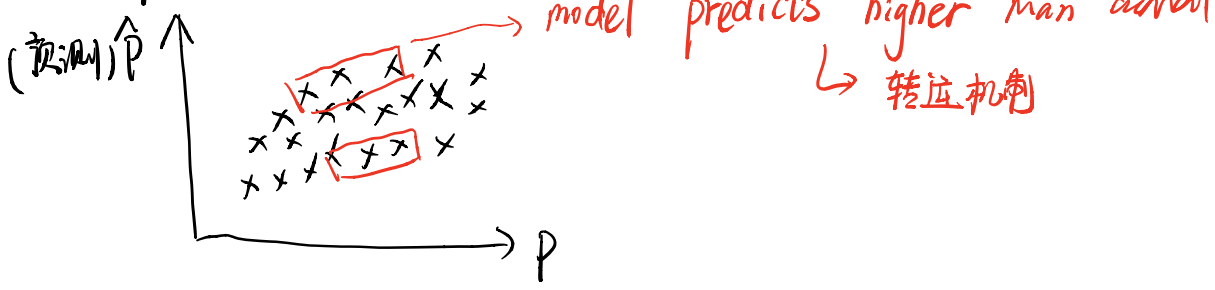
RIS HOR
PR ML

1. Genomics 基因组学





(2). Careful error analysis



1) build model → predict

measurements + derive features

⇒ regression model (回归模型)

2) Error analysis (statement about problem)

2. Finance

time series analysis

预测: $\hat{S}(n) = S(n-1), S(n-2)$

tomorrow today yesterday

residual (error) 剩余 ⇐ $r(n) = S(n) - \hat{S}(n)$

true predict

unexplained information ⇐ 例如 macro economics

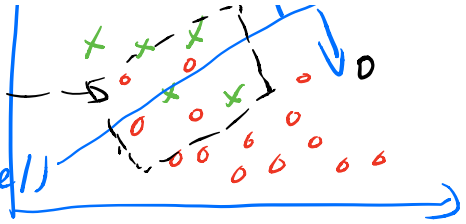
不可预测的原因 ⇒ 会导致 r(n) 的存在

3. classification 分类

① Features proxies 代理 vs knowledge

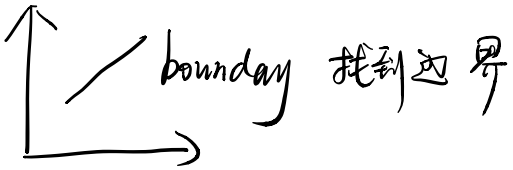
代替考试的形式

② Distribution 分布 x o



overlap ← - - - - → Training a model (Learning a model)

避免 ⇒ error function
Loss function
↳ minimise those 缩到最小 (将 error . loss)

use date ⇒ model ⇒  boundary 找到边界

点估：

① Regression 回归

② Time series 时间序列

③ classification 分类

Part 2. Chapter 1 of the book

Data → Regulationtios → predicts

Re presentation Learning

Training $\{x_n, t_n\}_{n=1}^{N}$ ⇒ Test $\hat{t}_{N+1} = f(x_{N+1})$

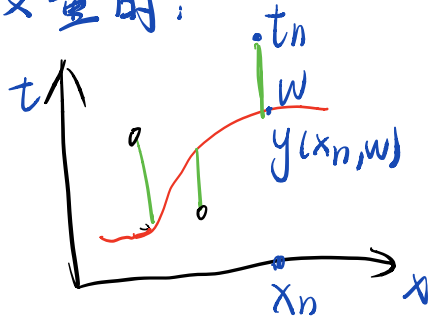
1.1 Polynomial Curve Fitting

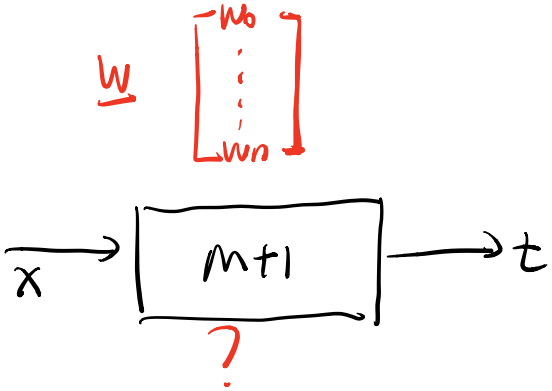
Model: Polynomial

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_m x^m$$

多个w变量时的模型表达式.

单变量时:





? $E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2$

误差

$E(0) \neq \frac{1}{2} x^2$?

问题: over fitting

原因: 跨度太大 / 数据量小

解决: constraining (约束) the weight magnitude (量级)

hyper parameter

tune 入 (超参数)

$$\tilde{E}(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2} \|w\|^2$$

$w_0^2 + w_1^2 + \cdots + w_n^2$ 和

1. $\lambda = 0$，不影响
2. λ 很大，$\|w\|^2 = 0$，才不影响

1.2 Probability Theory

$p(x)$ $p(y)$ $p(x,y)$

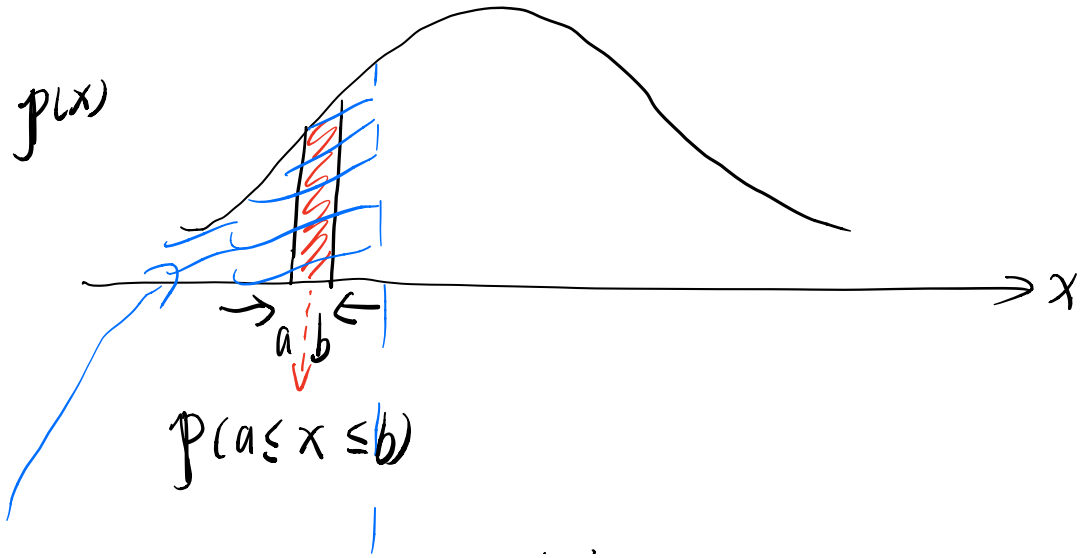
$p(x) = \sum_y p(x,y)$

$p(x,y) = p(x|y) \cdot p(y)$

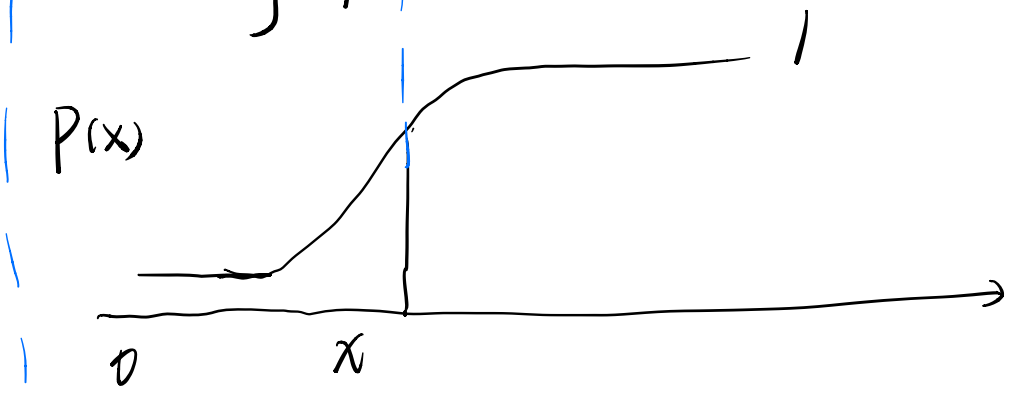
$= p(y|x) \cdot p(x)$

$$\Rightarrow \quad p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}$$

概率密度



cumility function 累积函数



$$P[X] = P[x < x] = \int_{-\infty}^{x} P(z)\,dz$$

Exportation X x f(x)

$$E[f] = \sum_{x} p(x) \cdot f(x) \qquad -(1.33)$$

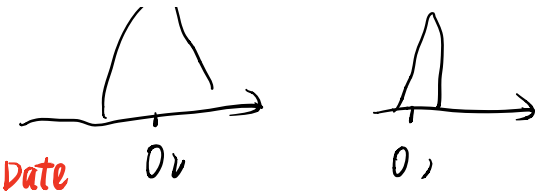
probability * integrate over the whole space

$$= \int p(x) f(x)\,dx \qquad -(1.34)$$

协方差 covariance

Frequentist → frequency of occurence 时间点

Bayesian 贝叶斯 → quantifying uncertainly 量化



(估算)

likelihood of Date

1.43. $$P(\underline{w}|D) = \frac{P(D|\underline{w})\,P(\underline{w})}{P(D)}$$

Posterior — Bayes inference on $\underline{w}$

conditional of Data (数据后验条件)

D: Data $\{x_n, t_n\}_{n=1}^{N}$

coefficients 系数 of the polynomial 多项式

1.44. Posterior $\propto$ likelihood $\times$ prior

$$P(D) = \int P(D|\underline{w})\,P(\underline{w})\,d\underline{w}$$

- 高斯

多1:- Decisions

- Model selection
- Curse of dimensionality 维度灾难?
- Information Theory (Entropy)

熵

Part 3

独立同分布 (1.54)

IID : independently identically distributed ?

likelihood : $p(\mathbf{x} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2)$ 已知 参数

对参数的函数 (输入 $\mathbf{x}$) (μ, σ^2) 高斯函数 $\sigma\sqrt{2\pi}$ ....

需要给出 : $\mathbf{x} = [x_1, x_2, \ldots, x_n]$

$$\ln p(\mathbf{x} \mid \mu, \sigma^2) = \frac{-1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln 2\pi$$

$$\frac{\partial \ln p(\mathbf{x} \mid \mu, \sigma^2)}{\partial \mu} = 0$$

$$\frac{\partial}{\partial \sigma^2} = 0$$

作业 max likelihood (ML)

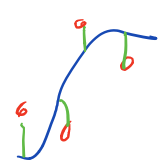
得到 ⇓

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \quad \text{平均值}$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$$

1.56
↓ 自己学(作业)
1.57

$\min \sum \text{error}^2$ 尽量减小误差



条件 condition prior input

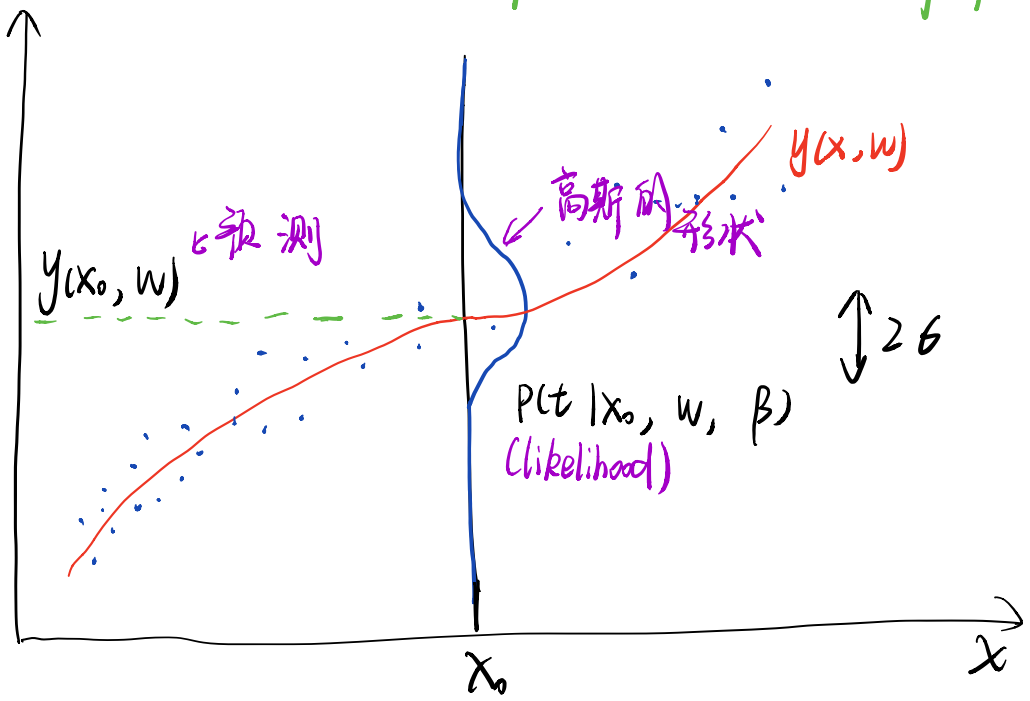
1.60 $p(t|x, w, \beta)$

probability over targets

noise term $\beta = \frac{1}{\sigma^2}$ → 变量 variance

precision parameter

weights (coefficients of polonomial)



预测的 $y(x_0, w)$ 与实际有偏差，是因为有噪音的存在 (2σ)

$$p(\underline{t}|\underline{x}, \underline{w}, \beta) = \prod_{n=1}^{N} N(t_n | y(x_n, \underline{w}), \beta^{-1})$$

原始高斯: $N(x|\mu, \sigma^2)$

$$\ln p(\underline{t}|\underline{x}, \underline{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \underline{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi$$

可能性最大化

precision 最小化:

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \underline{w}_{ML}) - t_n\}^2$$

贝叶斯公式:

$\underline{w}$ prrior distribution (先验散布)

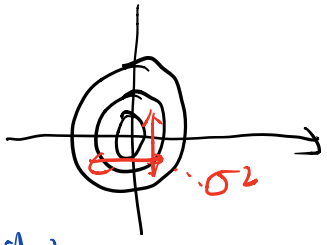
$\{x_n, t_n\}$ $\Rightarrow$ 得到 likelihood

$y(x, \underline{w})$

it 等于 = posterior distribution

并且最大化 maximize posterior ⇓

$$p(\underline{w}|\alpha) = \mathcal{N}(\underline{w}|\underline{0}, \alpha^{-1}I) \quad \cdots\cdots (1.65)$$



$m_i = 0$

$\sigma^2 I$

$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

改0: 先验的 α ↓ likelihood 的 $\alpha^{-1}I$

$$p(\underline{w}|\underline{x}, \underline{t}, \alpha, \beta) \propto p(\underline{t}|\underline{x}, \underline{w}, \beta)\, p(\underline{w}|\alpha)$$

posterior

若要取最大值 则 ⇒ 取最大值 ⇒ $\mathcal{N}(\underline{w})$

所以 max: $\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \underline{w}) - t_n\} + \frac{\alpha}{2}\underline{w}^T\underline{w}$

(w)

和 1.4 提到的公式：$\sum_{n=1}^{N} \{ \text{Data} \cdots \}^2 + \frac{\lambda}{2}\|w\|^2$ 一样

即：maximizing posterior

= minimizing regularized ? square

Week 4

overfitting:

1. posterior $\propto$ likelihood $\times$ prior

$$p(\underline{w}|\underline{x},\underline{t},\alpha,\beta) \propto p(\underline{t}|\underline{x},\underline{w},\beta) \times p(\underline{w}|\alpha)$$

↑max ⇑ $\underline{w}$

Bayesian

new point (input) $\leftarrow x$

predictive distribution on the targets $\rightarrow t$

$$p(t|x,\underline{x},\underline{t}) = \int p(t|x,\underline{w})\, p(\underline{w}|\underline{x},\underline{t})\, dw$$

$\underline{x}$, $\underline{t}$: training data

$p(t|x,\underline{w})$: 概率密度函数

$p(\underline{w}|\underline{x},\underline{t})$: 已知

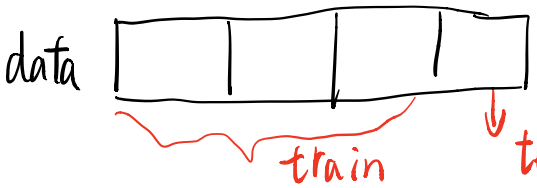
test input

posterior distribution 计算出 $\underline{w}$

2. 模型选择

问题 {
- high over polynomial → overfitting
- low ------ → not modelling data well

1) Cross validation



对比 model m2 m3.

2) Akaike information criterion AIC

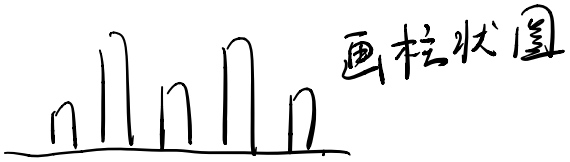
$$\ln p(D|w_{ML}) - M$$

↑ best value of w　　↑ model (polynomial order)

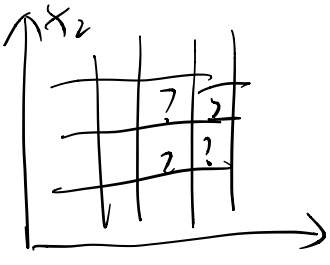
(重要概念)

3、Curse OF dimensionality (多维)

① 一维：



② 二维：

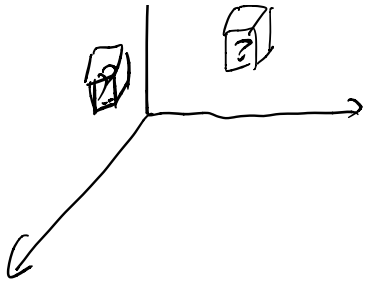


均分的各个格子各有多少数据

③ 三维：

↑

均分为多个立方体. 各有多少个data



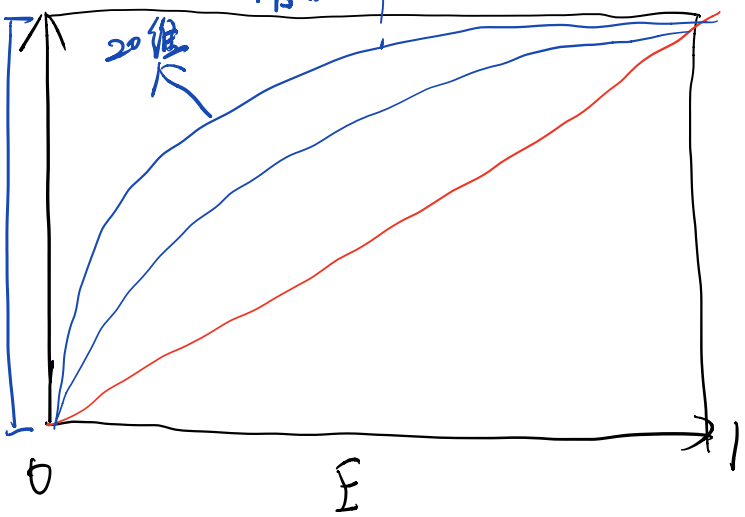
多维. Volume contain in high dimentions

二维 ⇒  计算圆形的 volume

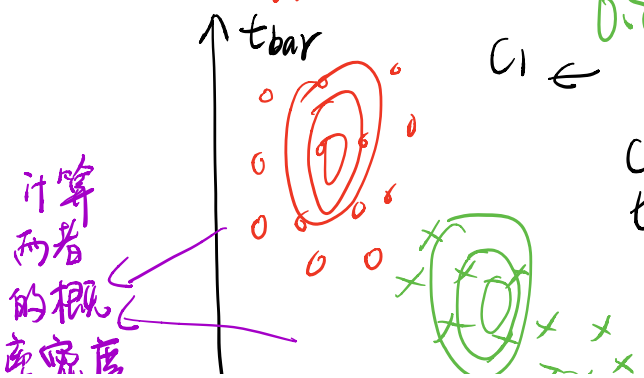
随着维度的增加

r=1

r=1-E



4. Decisions.



P(x|C2) Pass

P(x|C1) fail

$$P(C_k|\underline{x}) = \frac{P(\underline{x}|C_k)\cdot P(C_k)}{P(\underline{x})}$$

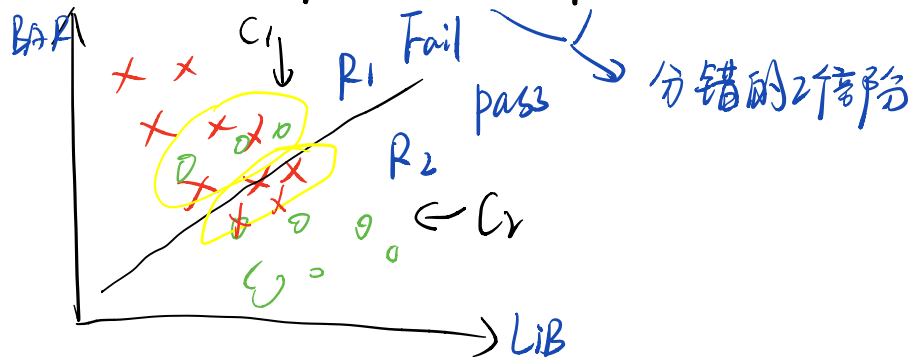
$P(\underline{x}|C_k)$: likelihood; $P(C_k)$: prior

所以，当给出 $\underline{x} = \begin{pmatrix} 3 \\ 0.1 \end{pmatrix}$（$C_1$的time；$C_2$的t）可计算出 C_k（pass or fail）

$$\therefore \quad P(C_k|\underline{x}) \geq 0$$

$$\sum_{k=1} P(C_k|\underline{x}) = 1$$

$$P(\text{error}) = P(\underline{x} \in R_1, C_2) + P(\underline{x} \in R_2, C_1)$$



分错的2情形

$$P(\underline{x}, C_1) > P(\underline{x}, C_2) \Rightarrow \text{class } C_1 \text{ ✓}$$

最小化 P(error)

L_{kj} : how much it cost to misclassify class j as class k

(Loss of k_j)

期望：

$$\sum_k \sum_j \int L_{kj}\, p(x, C_k)\, dx$$

L_{kj}：mispredicted as class j；$p(x, C_k)$：true class

如果不能计算，则可统计出错的次数

Reject

if $p(C_1|x) > p(C_2|x)$

在上述例子中，概率可分2种情况：

① 一种特别明显的 pass 0.8 > fail 0.2

② 一种是边界处 pass 0.501 ≥ fail 4.99

∴ 可reject第2种情况

if $p(C_1|x) > p(C_2|x)$ → class 1

else reject

week 5

Probability distribution

Bernoulli distribution ∈ binary

$p(x=1) = \mu$ $(0 \le \mu \le 1)$

$p(x=0) = 1-\mu$

$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$

⇓

当 $x=1$ 时 $\Rightarrow \mu$

当 $x=0$ 时 $\Rightarrow (1-\mu)$

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

D ∈ Data

$$= \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n}$$

$$\ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu)$$

$$= \sum_{n=1}^{N} x_n \ln \mu + (1-x_n) \ln(1-\mu)$$

最大化：则 $\frac{\partial \ln p(D|\mu)}{\partial \mu} = 0$

得：$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$ (x_n 由 0 或 1)

$= \frac{m}{N}$ → 多个 1

2、$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$

$E[\mu] = \frac{a}{a+b}$

$\text{Var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$

$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$

predictions：数据 ↑

$p(x=1|D) = \frac{m+a}{m+a+l+b}$

若 data only（只有数据）⇒ $\frac{m}{N}$

prior ⇒ $\frac{a}{a+b}$

posterior ⇒ $\frac{a+m}{(a+b)+N}$

3、 Likelihood 最大值.

$\mu \in [0,1]$

Week 6. chapter 4

备注： exp(-a) 表示 e^{-a}

4.1

决策面　　超平面

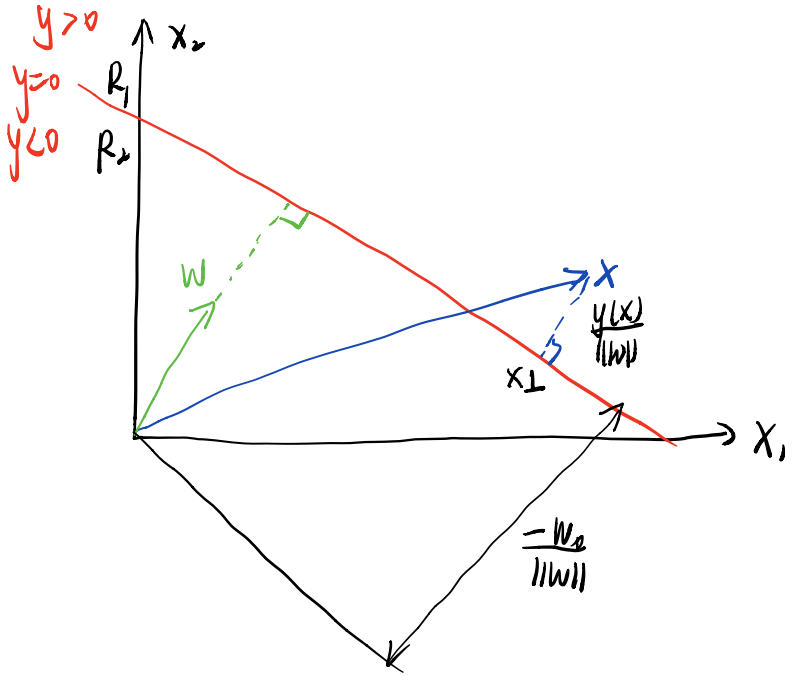
Decision surfaces are hyperplanes:

$$y(x) = w^T x + w_0$$

weight　　变量

① $y(x) \geq 0 \Rightarrow C_1$

②



有：$\frac{w^T x}{\|w\|} = -\frac{w_0}{\|w\|}$

③

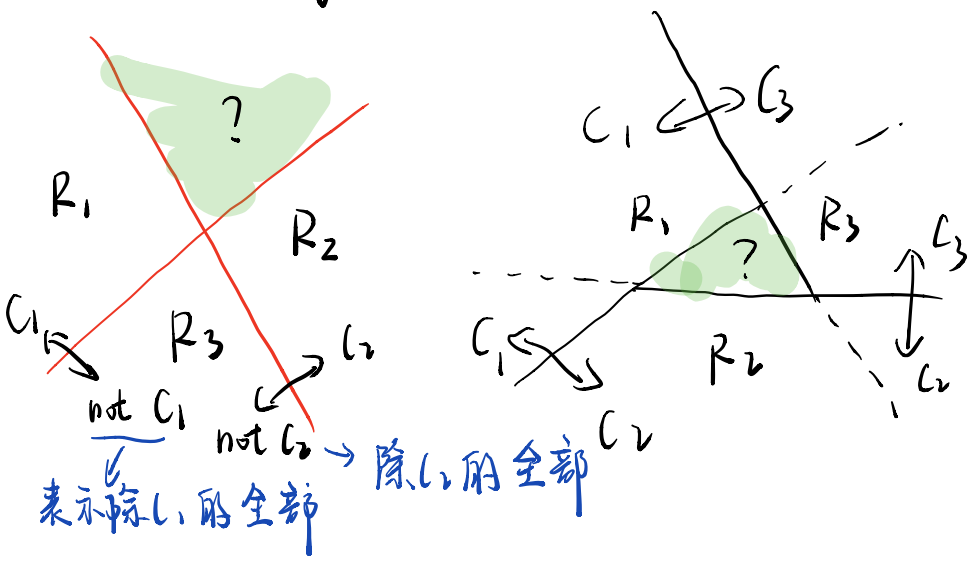
设 K classes $C_1, C_2 \cdots C_K$

(K-1) one versus rest (相对的) // $K\frac{(K-1)}{2}$ one versus one classifiers

$C_2 \Rightarrow C_3$
$C_3 \Rightarrow C_4$
$C_4 \Rightarrow C_5$
以及36对。

若K=10，则可 build 9 个 classfiers，

④ Ambiguous regions



定义了关于 K 的函数：

$$y_k(x) = w_k^T x + w_{k0}$$

指定 x → C_k if $y_k(x) > y_j(x)$, $\forall j \neq k$

则有：Boundary between C_k and C_j：

$$(w_k - w_j)^T x + (w_{k0} - w_{j0})$$

⑤ 指定 $t = [0, 0, 1, 0, 0]$

$$\tilde{w}_k = [w_{k0}, w_k^T]^T, \quad \tilde{x} = [1, x^T]^T$$ (x^0)

输入 class C_k if $y_k = \tilde{w}_k^T \tilde{x}$ is largest

vector of outputs as weight matrix times input. $y(x) = \tilde{w}^T \tilde{x}$

在以上条件的情况下，

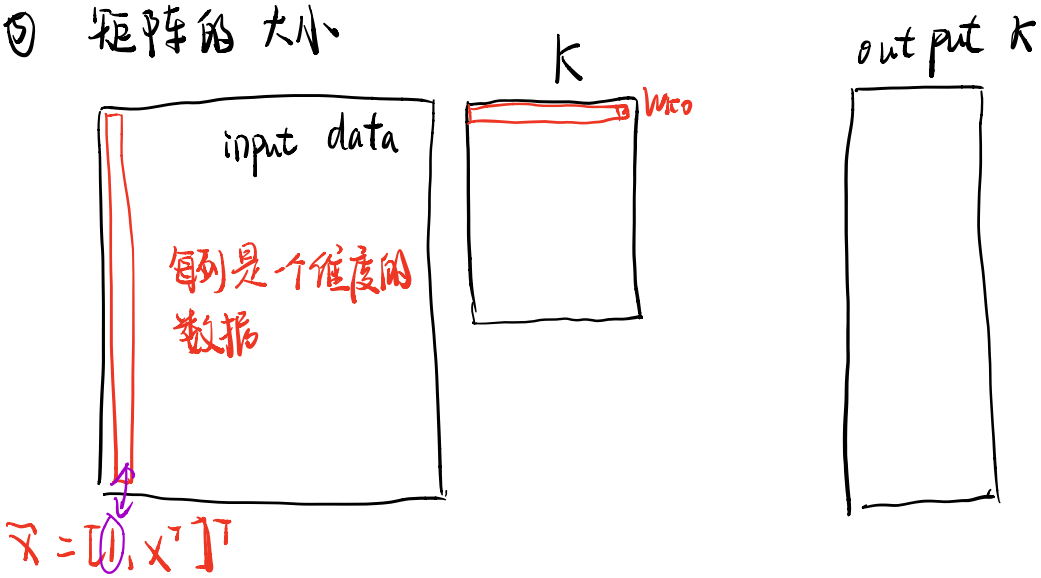
输入数据 Data $=\{x_n, t_n\}$，$n=1,2,\cdots N$

最小化 error：

$$E_d(\tilde{W}) = \frac{1}{2} Tr\{(\tilde{X}\tilde{W}-T)^T(\tilde{X}\tilde{W}-T)\}$$

（$\tilde{X}$：数据 × classes；T：的所有数据）

① 矩阵的大小



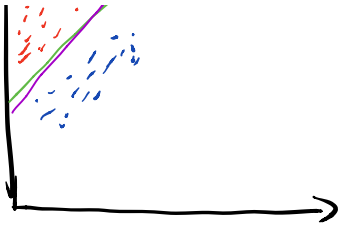
② 结论：

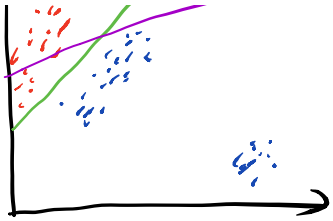
$$\tilde{W} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T T$$

① 不能将分类器问题表述为插值(回归)问题：

I. 原因 1

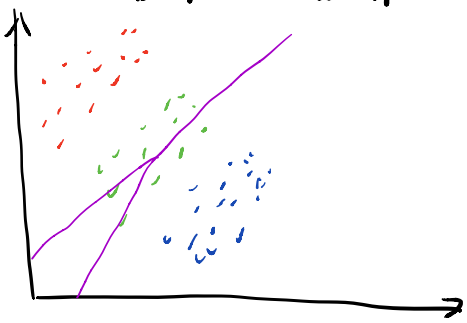
若有一小簇数据远离了则 2 种方法结果不一



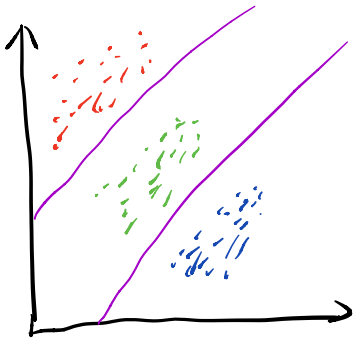


Ⅱ、原因2.

若有3类数据



回归



分类