1 EXTRA: LECTURE (Intro ML) EN.601.475 Lec2 Lec3

Lecture 2

concept ⟶ $\{(x_i, Y_i)\}_{i=1}^{n}$ Data ⟶ / output ⟶ ML ⟶ program $\hat{f}_n$; Data (experience) ⟶ task ⟶ Understanding (performance)

-formance Measures ⟶ X, Y observation
$f(x)$ = y-prediction

1.example loss $(Y, f(x)) = (Y - f(x))^2$ — continuous regression
0, 1 — discrete classification
$x \in \mathcal{X}$, (x,y) any cell

$(x, Y) \sim P_{XY}$

$$\boxed{\text{Risk } R(f) = \mathbb{E}_{XY}[\text{loss}(Y, f(x))]} \longrightarrow (1)$$

yes Optimal Rule ⟶ Goal: construct prediction rule $f^*: \mathcal{X} \to \mathcal{Y}$

Ideal optimal → $f^* = \arg\min_f \mathbb{E}_{XY}[\text{loss}(Y, f(x))] = \{f \mid \min \mathbb{E}_{XY}[\text{loss}(Y, f(x))]\} \to (2)$

large n

ayes Risk ⟶ $R(f^*) \le R(f) \ \forall f$ we have $\hat{f}$ $\left[\frac{1}{n}\sum_{i=1}^{n}[\text{loss}(Y_i, f(x_i))]\right]$

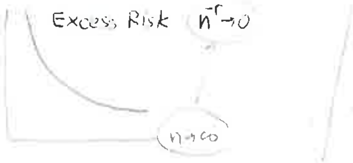
Training Algorithm ⟶ Data = $\{(X_i, Y_i)\}_{i=1}^{n}$, $\hat{f}_n$ = map from $\mathcal{X}$ to $\mathcal{Y}$, not overfit

Supervised Learning ⟶ $\frac{1}{n}\sum_{i=1}^{n}[\text{loss}(Y_i', \hat{f}_n(X_i))]$

Lecture 3

data $D_n = \{(X_i, Y_i)\}_{i=1}^{n}$

Performance Revisit ⟶



Application of ML

Ex1 loan
- credit score → regression
- loan decision → classification

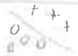
L in Application ⟶

EX2 : chess
- Nature of training sample / Exp
  - Game vs Pro (limited, not much control)
  - Pro's games (nearly unlimited, no control)
  - Self vs Self (unlimited, flexible)

- Task (T)
- Training Sample / Exp (E)
- Type of output (cata, %, number) (O)
- Performance measure / loss Fn (P)
- Input (mace, credit score) (X)
- Hypothesis space (H) function $H: X \to O$
- optimize (P) using $H: X \to O$

light GBM , XGBoost

- supervised →



- unsupervised

Octave ⇒ free open source SW
L Matlab = ✓ good ML invironment

Lecture 3

Linear Regression

Types of supervised Problems → ① regression ② classification

Linear function → $f(x;w) = w_0 + w_1x_1 + \dots w_dx_d = \vec{w}\cdot\vec{x}$
$= \vec{w}^T x_i$

$x \in X \in \mathbb{R}$ line
$x \in \mathbb{R}^2$ plane
$\in \mathbb{R}^d$ hyper Plane

Notation →
- $x_i \overset{[N\times d]}{\in} X \in \mathbb{R}^d$ ↳ $d=0$ mean const ⇨ $w = [w_0, w_1, \dots w_d]$, $x_i = [x_{i0}, x_{i1}, \dots x_{id}]$
- $y_i \in \mathcal{Y}$
- $\bar{X}$: $N \times (d+1)$ data matrix
- $\bar{y}$ = label vector $\bar{y} = [y_1, \dots y_N]^T$

$X =$

$f(x;w) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \mathbf{w}\cdot\mathbf{x}_i$ • loss function $\ell(\hat{y}, y)$

Loss function → from function $y = f(\bar{x};\bar{w})$ x_0, y_0 = new data

$R(w) = E_{(x_0,y_0)\sim p[\bar{x},y]}[\ell(f(x_0;\bar{w}), y_0)]$ goal: minimize to loss $R(\bar{w})$ for new data

$w^* = \arg\min_w \sum_{i=1}^{N} (y_i - \mathbf{w}\cdot x_i)^2$

Least Square → to do this

we should minimize $L(w, X, y) = L(w) = E[\ell(y,w,x)] \approx \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{w}\cdot x_i)^2$

$$\boxed{\frac{\partial L}{\partial w_j} = \nabla_{w_j} L = \emptyset \quad \forall j} = \left[\frac{\partial L}{\partial w_0}, \frac{\partial L}{\partial w_1}, \dots \frac{\partial L}{\partial w_d}\right]$$

$= \frac{1}{N}(y - Xw)^T(y - Xw)$

$= \frac{1}{N}(y^T - X^TW^T)(y - Xw)$

condition 1 $\frac{\partial L}{\partial w_j} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - \mathbf{w}\cdot x_i)x_{ij} = 0$ —— (1)

condition 2 Error are uncorrelated w/ data and linear fn

$\sum_{i=1}^{N}(y_i - w\cdot x_i) = 0$ —— (2)

Derivative of loss

$L(w) = \frac{1}{N}(Y^T - X^TW^T)(y - Xw)$

★ From $\frac{\partial(a^Tb)}{\partial a}$ $\left[\frac{\partial}{\partial a}\right][a^T]\ \square$ ✓ $[b]$ $\left[\frac{\partial}{\partial a}\right]([b^T]\ \square\ [a])$ $\frac{\partial(b^Ta)}{\partial a} = [b]$ ✓

★ matrix $\frac{\partial}{\partial a_i}(a^TBa) = \frac{\partial}{\partial a_i}(a^T(Ba))^T = \frac{\partial}{\partial a_i}((Ba)^T(a^T)^T) = \frac{\partial}{\partial a_i}(a^TB^Ta) \leftarrow \boxed{B = B^T}$ condition

dummy index → $B_{ij}a_i \rightarrow B_{kj}a_k$ | $a_jB_{ih} \rightarrow a_kB_{ik}$

$$\frac{\partial(a^TBa)}{\partial a_k} = \frac{\partial}{\partial a_k}\left(\sum_{i,j} a_jB_{ij}a_j\right) = \sum_{kj} a_jB_{kj}\left(\frac{\partial}{\partial a_k}a_k\right) + \sum_{ik}\left(\frac{\partial}{\partial a_k}a_k\right)B_{ik}a_i$$

$$\frac{\partial(a^TBa)}{\partial a_k} = a_jB_{kj} + B_{ik}a_i \Rightarrow (B^T + B)a$$

ศกาณแนะนำ ลบ ตอวัส
ไหน

$\frac{\partial(a^T(Ba))}{}$ ... $(Ba)^T(a^T)^T$ ... (a^TB^T) ... aBa

Lec3 Lec4

erivative of loss 2 ⤚→ ⇨ $\frac{\partial L(w)}{\partial w} = \frac{1}{N}\frac{\partial}{\partial w}\left[y^Ty - w^TX^Ty - \overset{w^T(y^Tx)^T}{(y^Tx)w} + w^TX^TXw \right]$

$= \frac{1}{N}\left[- X^Ty - (y^TX)^T + (X^TX + (X^TX)^T)w \right]$

$= \frac{1}{N}\left[-2X^Ty + 2X^TXw \right]$

(X^+)

oore- Penrose

udoinverse of X ⤚→

$\frac{\partial L}{\partial w} = -\frac{2}{N}(X^Ty - X^TXw)$ --③ $= 0$

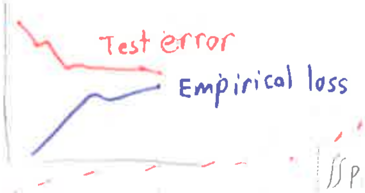
$\therefore \boxed{w^* = (X^TX)^{-1}X^Ty}$ $\boxed{X^+ \triangleq (X^TX)^{-1}X^T}$

prediction : $\overset{\mathbb{R}}{\hat{y}} = w^* \cdot \overset{\mathbb{R}^N}{x_0} = (y^TX^{+T})x_0$

Lecture 4

Generalization

More training data → worse fit
→ better prediction

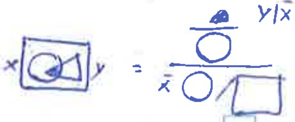


$f^* = \underset{f: X \to \mathbb{R}}{\arg\min}\ \mathbb{E}_{(x_0,y_0) \sim p(x,y)}\left[(f(x_0) - y_0)^2 \right]$ —— (1)

New data

$\iint p(x,y)\,dx\,dy = \int \left(\int p(y|\bar{x})\,dy \right) p(x)\,dx$

rule of probability : $p(\bar{x},y) = \underline{p(y|\bar{x})\,p(\bar{x})}$ —— (2)



(1) → def $\mathbb{E}_{p(y,\bar{x})}[g(y,\bar{x})] = \int_x \left\{ \int_y g(y,\bar{x})\,\underline{p(y|\bar{x})\,p(\bar{x})}\,dy \right\} d\bar{x}$ —— (3)

$\mathbb{E}_{p(y,x)}[g(y,x)] = \int_x \left\{ \int_y g(y,\bar{x})\,p(y|\bar{x})\,dy \right\} p\bar{x}\,d\bar{x}$ —— (4)

notation ⤚→ $(x_0,y_0) \sim P(x,y)$ ⇒ (x_0,y_0) has pb distribution of $P(x,y)$

$\mathbb{E}_{x_0 \sim p(x)}[\quad]$

(4) → $\mathbb{E}_{(x_0,y_0)\sim p(x,y)}\left[(f(x_0) - y_0)^2 \right] = \int_{x_0} \left\{ \mathbb{E}_{y_0 \sim p(y|x)}\left[(f(x_0) - y_0)^2 \,|\, x_0 \right] \right\} p(x_0)\,dx_0$ —— (5)

try to minimize it for each x_0

conditional expectation

vary f(x) given x_0

(5) $\frac{\partial}{\partial f(x)} \mathbb{E}_{p(y|x)}\left[(f(x_0) - y_0)^2 \,|\, x_0 \right] = 2\mathbb{E}_{y_0 \sim p(y|x)}\left[2(f(x_0) - y_0) \,|\, x_0 \right] = 0$

$= 2\left(\int f(x_0)\,p(y|x)\,dy - \mathbb{E}_{y_0 \sim p(y|x)}[y_0|x_0] \right) = 0$

$= 1$

mean function of (x_0)

should reture $\mathbb{E}_p[y_0]$ ← $f^*(x_0) = \mathbb{E}_{p(y|x)}[y_0|x_0] = \mathbb{E}_{y_0 \sim p(y|x)}[y_0|x_0]$

① ไหนะ ส่ม ② รู้ (อาจารย์ ...)

③ sport club - organization/competition/fund raising / com... service

③ ...

④ ...

Rule of Prob

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

Iteration

$w^j = w^j - \alpha \frac{\partial}{\partial \theta_j} \ell(w_0, w_1)$

Lec4

generative vs. discriminate approach

$\hat{y}(x_0) : E_{y \sim p(y|x)}[y|x_0]$

F : observe

$\hat{F}$: estimate

$\overset{*}{F}$: optimal

1) Generative Approach ⟶ Estimate $p(x, y)$ ↓ Normalize → find $p(y|x)$

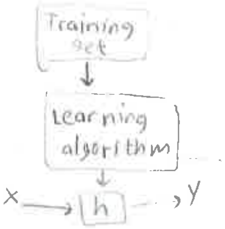
2) discriminate Approach ⟶ Estimate $p(y|x)$ from data

Supervised learning

- training from the right data

Regression Problem

- Predict real-valued output



Decomposition Of error

Expected Loss ⟶ $\hat{w}$ LSQ estimate from training data

$\overset{*}{w}$ optimal regression parameter

} w = parameter vector : $\bar{w} \cdot \bar{x}$

$$y - \hat{w} \cdot x = (y - \overset{*}{w} \cdot x) + (\overset{*}{w} \cdot x - \hat{w} \cdot x)$$

$$E[\overset{\varepsilon}{y - \overset{*}{w} \cdot x}] = 0 \not\propto E[\overset{*}{w} \cdot x - \hat{w} \cdot x] = 0$$

$$E_{p(x,y)}[(y - \hat{w} \cdot x)^2] = E_{p(x,y)}[(y - \overset{*}{w} \cdot x)^2] + 2E_{p(x,y)}[\underbrace{(y - \overset{*}{w} \cdot x)}(\overset{*}{w} \cdot x - \hat{w} \cdot x)] + E_{p(x,y)}[(\overset{*}{w} \cdot x - \hat{w} \cdot x)^2]$$

prediction error $\not\propto$ linear fn $\overset{*}{w} \cdot x - \hat{w} \cdot x$

$$\boxed{E_{p(x,y)}[(y - \hat{w} \cdot x)^2] = E_{p(x,y)}[(y - \overset{*}{w} \cdot x)^2] + E_{p(x,y)}[(\overset{*}{w} \cdot x - \hat{w} \cdot x)^2]} \quad (6)$$

Approximation Error (variance)

Error $\not\propto N$

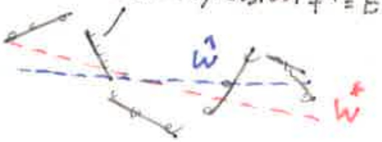
$\propto$ hypothesis

Estimation Error (Bias)

(how close opt $\overset{*}{w}$ from infinit training $\hat{w}$)

$\lim_{N \to \infty} \hat{w} = \overset{*}{w}$

best regression $f^* = E[y|x]$



MSE (Mean Square Error) & bias-variance decomposition

towards datascience

Qustion for underfitting, overfitting, model capacity, MSE for estimator & predictor

$X \sim$ Ddistribution

$X = (x_1, x_2, \dots, x_n)$; $\hat{y} = f(x; W)$;

- estimated variance : $\hat{\sigma}^2 = \frac{1}{n} \sum^{n} (x_i - \bar{x})^2$
- Estimator properties : bias $= E[\hat{y}] - \overset{*}{y}_{real}$

  Variance $=$ Var$(\hat{y})$

Berkeley sheet

Lec 4

Adding Noise → $y = f(x;w) + \nu \rightarrow$ use ful info

$E_{p(y|x)}[f(x;w) + \nu | x] = f(x;w) + E_{p(\nu)}[\nu]$

aussian Noise Model → $y = f(x;w) + \nu$, $\nu \sim \mathcal{N}(\nu; \underset{\text{mean}}{0}, \underset{\text{Variance}}{\sigma^2})$ distribution of

$p(y|x;w) \xrightarrow[\mu=0,\sigma^2]{\text{add}} p(y|x;w,\sigma) = \mathcal{N}(y; f(x;w), \sigma^2)$

$= \frac{1}{(\sqrt{2\pi})\sigma} e^{-\left(\frac{(y-f(x;w))^2}{2\sigma^2}\right)}$ Gausian (7)

ikelihood → likelihood of parameter w given observed data $\overset{\text{matrix}}{X} = [x_1, \dots, x_N]$, $\overset{\text{vector}}{Y} = [y_1, \dots y_N]^T$

$\mathcal{L} = P(Y|X; w, \sigma)$ prob to observe a data y given X under model with parameters ; w , σ

$\boxed{\text{IID}}$ independently , identically, distributed between set data (x_i, y_i)

$$\boxed{p(\vec{Y}|X; \vec{w}, \sigma) = \prod_{i=1}^{N} p(y_i|\vec{x}_i, w, \sigma)}$$ IID (8)

Maximum Likelihood → $\hat{W}_{ML} = \arg\max_{w} P(Y|X; w, \sigma)$ (7),(8)

$\log(\hat{W}_{ML}) = \arg\max_{w} \sum_{i=1}^{N} \log p(y_i|x_i, w, \sigma)$

$= \arg\max_{w} \sum_{i=1}^{N} \left[-\frac{(y_i - f(x_i;w))^2}{2\sigma^2} - \log \sigma\sqrt{2\pi} \right]$

$= \arg\max_{w} \left[-\cancel{\frac{1}{2\sigma}} \sum_{i=1}^{N} (y_i - f(x_i;w))^2 - N \cancel{\log \sigma\sqrt{2\pi}} \right]$ independent to w

max likeli

$$\boxed{L(f(x;w), y) = -\log p(y|x;w,\sigma) = \sum_{i=1}^{N} (y_i - f(x_i;w))^2}$$ (9) for Goussian Noise

General Additive Regression Model → $\hat{y} = f(x;w) = w_0 + w_1 \phi_1(x) + w_2 \phi_2(x) + \dots + w_m \phi_m(x)$ some function (input : vector) x_i = vector

$\hat{W} = \begin{bmatrix} w_0 \\ \vdots \\ w_m \end{bmatrix} = ?$ $X = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_m(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_N) & \cdots & \phi_m(x_N) \end{bmatrix}$

$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_m x^m$

$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m \\ 1 & x_2 & \cdots & x_2^m \\ \vdots & & \ddots & \\ 1 & x_N & \cdots & x_N^m \end{bmatrix}$

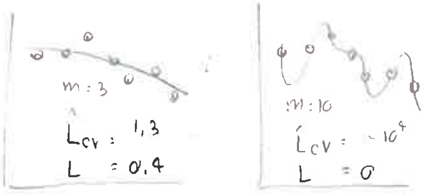
$\hat{w} = (x^T x)^{-1} x^T y$

Overfitted Problem ⟶ สังเกตุ : too sensitive , unstable to each data point

Crossvalidation

Leave-one-out cross-validation

$$\hat{L}_{cv} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i;\hat{w}_{-i}))^2$$ (10) ; $\hat{w}_{-i}$: fit parameter with out i-th data



Lecture 5

Regulation

Road Map ⟶

- ▷ regulation = a tool against over fitting
- ▷ gradient descent

(Penalty)

Penalizing Model complexity ⟶ Itution ; penalize # of bits required to encode the parameter

$$(4.9) \longrightarrow \boxed{w^* = \arg\max_{w}\left\{\frac{1}{2}\sum_{i=1}^{N}\log p(\text{data}_i ; w) - \text{penalty}(w)\right\}} \longrightarrow (5.1)$$

shrinkage Method

→ given $y_i | x_i$

Loss or (E[ℓ( )])

minimum difference $y, \hat{y}$

Ridge Regression ⟶

$$w_{ridge} = \arg\min_{w}\left\{\sum_{i=1}^{N}(y_i - \overset{\hat{y}_i}{w \cdot x_i})^2 + \lambda\sum_{j=1}^{m} w_j^2\right\} \longrightarrow (5.2)$$

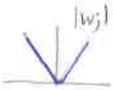
$\vdash f(x_i, w_i)$

(5.4) → $$\hat{w}^*_{ridge} = (\lambda I + X^T X)^{-1} X^T y$$

→ (5.3) proof below

L = E[ℓ( )]

Lasso Regression ⟶ $$E(w^*) = w^*_{lasso} = \arg\min_{w}\left\{\sum_{i=1}^{N}(y_i - w \cdot x_i)^2 + \lambda\sum_{j=1}^{m}|w_j|\right\}$$



Problem of lasso → can't $\frac{\partial L}{\partial w}$ → Need Numerical Opt. tools

eq 5.5

constrain form

$$\bar{w} : \sum_{j=1}^{m} w_j^2 \le t \qquad \bar{w} : \sum_{j=1}^{m} |w_j| \le t$$

Prove $\overset{*}{w}_{ridge}$ (from lecture 3)

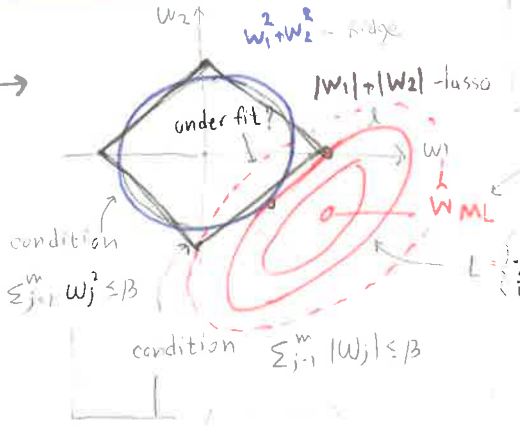
$$L = E[\ell(\,)] \approx \sum (y_i - w \cdot x_i)^2 + \lambda\sum_{j=1}^{m} w_j^2$$

$$= (y^T - x^T w^T)(y - xw) + \lambda w^T w$$

* increase the size of data also reduce over-fitting Pb

w^*

ometry of
rror Surface ⟶


w_2
$w_1^2+w_2^2$ - Ridge
$|w_1|+|w_2|$ -lasso
under fit?
w_1
$\hat{w}_{ML}$
condition
$\Sigma_{j=1}^{m} w_j^2 \leq \beta$
condition $\Sigma_{j=1}^{m} |w_j| \leq \beta$
$L = \{ -\sum_{i=1}^{N}(y_i - \hat{w}\cdot x_i)^2 \}$

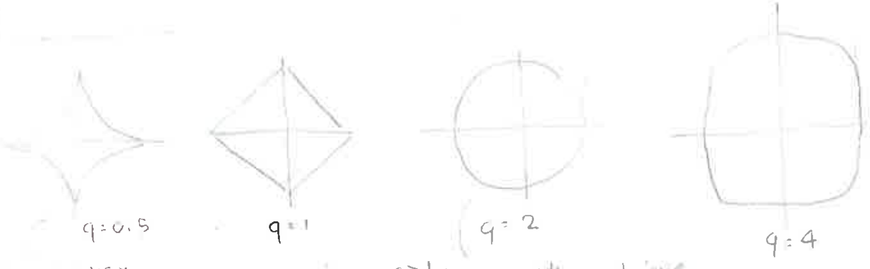
$$\hat{w} = \underset{w:\ \|w\|_q \leq \beta}{\arg\max} \left\{ -\sum_{i=1}^{N}(y_i - w\cdot x_i)^2 \right\}$$

over fit

optimization (5.5)

choice of λ ⟶ high cross validation $\hat{L}_{cv}$

view of L_q ⟶
penalty of
function $\|w\|_q$



q<1 non-convex

q>1 no sparsity achieve

$$\boxed{\|w\|_q = \left(\sum_{j=1}^{m} |w_j|^q \right)^{\frac{1}{q}}} \longrightarrow (5.5)$$

$$\therefore \|w\|_\infty = \max_j |w_j|$$

Lagrange Multiplier 2. <proof (5.c)>

Now we consider 'C' a variable (changable)

$$\left\{ \begin{array}{l} \vec{x}^* \to \vec{x}^*(c) \\ \lambda^* \to \lambda^*(c) \end{array} \right\} \mathcal{L}(\vec{x}^*(c), \lambda^*(c), c) = f(\vec{x}^*(c)) - \lambda^*(c)(g(\vec{x}^*(c)) - c)) \longrightarrow (5.d)$$

$$\mathcal{L}(\vec{x}^*(c), \lambda^*(c), c) = \hat{M}(c) - 0$$ — at optimization $g(\vec{x}^*(c)) = C$ (5.a)

(5.d)

$$\frac{dM^*(c)}{dc} = \frac{d\mathcal{L}}{dc} = \frac{\partial \mathcal{L}}{\partial \vec{x}^*}\frac{d\vec{x}^*}{dc} + \frac{\partial \mathcal{L}}{\partial \lambda^*}\frac{d\lambda^*}{dc} + \frac{\partial \mathcal{L}}{\partial c}\frac{dc}{dc} = \frac{\partial f}{\partial \vec{x}^*}\frac{d\vec{x}^*}{dc} + \dots + \frac{\partial\, c\lambda^*(c)}{\partial c}\frac{dc}{dc} \Rightarrow \boxed{\frac{dL(*)}{dc} = \frac{dM^*}{dc} = \lambda^*} \to (5.c)$$

(5.a bottom)

(Mathematic Optimization)

(wiki) Lagrange Multiplier 1.

General form

Lagrange Multiplier

maximize, constrain function, const

$g(\vec{x}) = x_1 \dots$

$$\boxed{\begin{array}{l} \mathcal{L}(\vec{x}, \lambda) = f(\vec{x}) - \lambda(g(\vec{x}) - c) \\ \nabla\mathcal{L} = 0 \iff \nabla f = \lambda \nabla g \end{array}} \to (5.a)$$

Ex. labor = \$20/h = h
steel = \$2000/ton = s

revenue = $R(h,s) = 100h^{2/3}s^{1/3}$ → what to optimize

g(h,s) = budget = 720000 = 20h + 2000s → constrain

$\boxed{\nabla R = \lambda \nabla g}$ (Lagrange Multiplier)

g(h,s) = ... ∇R

$\nabla\mathcal{L} = 0 \Rightarrow$ we get the maximized sol^n $(\vec{x}^*, \lambda^*)$! → (5.b)

Max = $\hat{M}^* = F(\vec{x}^*) = f(\vec{x}^*(c))$ ⟶ $\vec{x}^*$ is a function of 'C'

$\lambda^* = \frac{d\hat{M}(c)}{dc}$ } change of maximization
by constrain const 'C' → (5.c)

Book Ch 1.2

Probability Theory

pdf, cdf → $\int_{-\infty}^{\infty} p(x)dx = 1$; $\int_{-\infty}^{x} p(x)\,dx = P(x)$; $p(x) \geq 0$ —— → (1)

pdf transform → from x to y given $x = g(y)$ $\Big\}$ $p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$ $= p_x(g(y))\,|g'(y)|$ —— (2)

Expectation → $\mathbb{E}[f] = \overset{\text{discrete}}{\sum_x p(x)\,f(x)} = \overset{\text{continuous}}{\int p(x)\,f(x)\,dx}$

conditional: $\mathbb{E}_x[f|y] = \sum_x p(x|y)\;f(x)$

var $\quad Var[f] = E\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - (\mathbb{E}[f(x)])^2$

cov[x,y] $\quad \mathbb{E}_{x,y}\left[\{\bar{x} - \mathbb{E}[\bar{x}]\}\{\bar{y}^T - \mathbb{E}[\bar{y}^T]\}\right] = \mathbb{E}_{x,y}[x\bar{y}^T] - \mathbb{E}[x]\,\mathbb{E}[\bar{y}^T]$ } matrix

Bayesian Prob → $\bar{w}$ ⇒ polynomial curve fitting
$p(w)$ ⇒ prior probability distribution
D ⇒ $\{t_1, \dots, t_N\}$

$p(D|W)$ conditional probability

posterier prob. $\quad \underbrace{p(\bar{w}|D)}_{}\;$ $\overset{\leftarrow\text{post}\rightarrow}{}$ $= \dfrac{\overset{\leftarrow\text{likelihood}\rightarrow}{p(D|\bar{w})}\,\overset{\leftarrow\text{prior}\rightarrow}{p(\bar{w})}}{p(D)}$ ——→ (3)

Gaussian Distribution → $\mathcal{N}(x|\mu,\sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{1/2}}\, e^{\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}$ ——→ (4)

$E[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\,x^2\,dx = \mu^2 - \sigma^2 \rightarrow Var[x] = E[x^2] - E[x]^2 = \sigma$

For vector $\bar{x}$ [Nx1], Σ [N×N] : covariance matrix, $|\Sigma|$ = det of Σ

$$\boxed{\mathcal{N}(\bar{x}|\bar{\mu},\Sigma) = \frac{1}{(2\pi)^{N/2}\,|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x}-\mu)^T\,\Sigma^{-1}\,(x-\mu)\right\}} \text{ —— (5)}$$

In case of $\bar{X} = \{x_1, x_2, \dots, x_N\}^T$ independent and identically distributed (i.i.d)

↳ $\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \ddots\ \sigma^2 \end{bmatrix}_{N}^{N}$

$$\boxed{p(\bar{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)} \text{ —— (5a)}$$

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

ML : Maximum likelihood $\boxed{\vec{x} = \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n}$ $\boxed{S^2 = \sigma^2_{ML} = \frac{1}{N}\sum^{N}(x_n-\bar{x})^2 = \frac{N-1}{N}\sigma^2 = \mathbb{E}[\sigma^2_{ML}]}$ (6)

non-biased

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

proof ⟶ proof

$$\mathrm{Var}[x_i] = \mathbb{E}[(x_i-\mu)^2] = \mathbb{E}[x_i^2] - \mathbb{E}[\mu]^2 = \sigma^2 = \sigma^2_{model} \longrightarrow (6.a)$$

$$\mathrm{Var}[\bar{x}] = \mathrm{Var}\left[\frac{1}{N}\sum_{n=1}^{N}x_i\right] : \langle \mathrm{Cor}[ax,by] = (ab)\,\mathrm{cor}[X,Y]\rangle = \frac{1}{N^2}\mathrm{Var}\left[\sum_{n=1}^{N}x_i\right]$$

$$= \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}[x_i] = \frac{1}{N}\sigma^2 \longrightarrow (6.b)$$

$$\therefore \mathbb{E}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})^2\right] = \frac{1}{N}\mathbb{E}\left[\sum(x_n-\mu)^2 - 2\sum(x_n-\mu)(\bar{x}-\mu) + \sum(\bar{x}-\mu)^2\right]$$

(6.a) ⟶

$$= \left(\frac{1}{N}\right)\left\{\sum^{N}\mathbb{E}[(x_n-\mu)^2] - \sum^{N}\mathbb{E}[(\bar{x}-\mu)^2]\right\}$$

(6.b) ⟶

$$\mathbb{E}[\sigma^2_{ML}] = \left(\frac{1}{N}\right)\left\{N\sigma^2 - N\left(\frac{1}{N}\sigma^2\right)\right\} = \left(\frac{N-1}{N}\right)\sigma^2 \longrightarrow (6)$$

urve-fitting ⟶ : error minimalization + regulation

Input $\vec{x} = \{x_1, \dots x_N\}^T$, $\bar{Y}_{ob} = \{Y_1, \dots Y_N\}^T$, $y = y(x,\vec{w})$

$$p(t|x,\vec{w},\beta) = \mathcal{N}(Y_{obl}|Y_{fit}(x,\vec{w}),\beta^{-1}) \quad ; \quad \beta^{-1} = \sigma^2$$

$$\mathcal{L}\,(\text{likelihood}) = \prod_{n=1}^{N}\mathcal{N}(Y_{obn}|y(x_n,\vec{w}),\beta^{-1}) \leftarrow (5) \leftarrow (5a)$$

$$\log(\mathcal{L}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\vec{w}) - t_n\}^2 + \frac{N}{2}\cancel{\ln\beta} - \frac{N}{2}\ln(2\pi)$$

1.2.6 Bayesian Curve Fitting ⟶ (3) → post/predict distri $p(Y_+|x_+,\bar{x},\bar{Y}) = \int \underbrace{p(Y_+|x_+,\vec{w})}_{f(\ )}\,\underbrace{p(\vec{w}|\bar{x},\bar{y})}_{pdf}\,d\vec{w}$

Y_+, W_+ = New data , $\bar{x},\bar{Y}$ = trained data

1. Book 1.3

Model Selection

lecture (5.1) (5.2) → get the highest w/ restriction

ross-validation → $\left(\frac{S-1}{S}\right)$ " leave-one-out technique"

① leave 1st point, train the rest 2nd - Nth

② find prediction for 1st point

③ do it to the rest (take out 2nd to Nth)

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

ook 1.5

Decision Theory

- training data $(\vec{x}, \vec{y})$ (regression) \
  $(\vec{x}, C_k)$ (classification) → Inference → $p(\vec{x}, \vec{y})$ \
  $p(\vec{x}, C_k)$ } joint distribution (unknown)

EX: $\vec{x}$ - pixcel of images / C_k = {normal, cancer cells}

$\overset{\parallel}{y} = \{0, 1\}$

5.1 Minimize isclassification rate → **Bayes thm**

Book (3) → $p(C_k|\vec{x}) = \dfrac{p(x|C_k)\,p(C_k)}{p(\vec{x})}$ (7)

$p(\vec{x}, C_k)$ = joint distribution

$p(C_k)$ = prior prob for class $\{C_k\}$

$p(C_k|\vec{x})$ = proterior prob for class $\{C_k\}$

example of

C_0 = normal

C_1 = cancer

→ How to assign class to x?

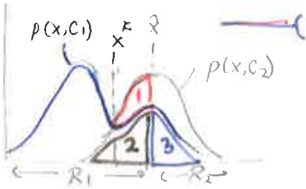
(Input Space) $\vec{x} \in$ (region) $R_1 \xrightarrow{\text{assign}}$ class C_1

Ex $\downarrow p(\text{mistake}) = p(\vec{x} \in R_1, \overset{y}{\overset{\parallel}{C_2}}) + p(\vec{x} \in R_2, C_1)$

$= \int_{R_1} p(\vec{x}, C_2)\,d\vec{x} + \int_{R_2} p(\vec{x}, C_1)\,d\vec{x}$

$\uparrow p(\text{correct}) = \sum_{k=1}^{K} p(x \in R_k, C_k)$

$= \sum_{k=1}^{K} \int_{R_k} \underline{p(\vec{x}, C_k)\,d\vec{x}}$ find C_k that $p(\vec{x}, C_k)$ = largest



(8)

Error

1 : C_2 ✓ C_1 ✗ ● + ●

2 : C_2 ✗ C_1 ✓ ●

Posterior : opt $x = x^*$

where $p(x, C_1)$, $p(x, C_2)$ crosses

k 1.6 → **Information Theory**

- amount of Information → degree of surprise to learn X
- h(x) h(y) → Information content
- p(x,y) = p(x) p(y) "Indep"

$$h(x) = -\log_2 p(x) \xrightarrow{\text{bits}} (8)$$

$$\text{or } = -\ln p(x)$$

ntropy → Entropy $\overset{(8)}{}$ $\boxed{H[x] = E[h(x)] = -\sum_i p(x_i) \log_2 p(x_i)} \longrightarrow (9)$

(uniform h(x) → higher H[x])

w/ regulation $\sum_i p(x_i) = 1$ Lagrange multiplier

$$\nabla \mathcal{L} = 0 : \nabla R - \lambda \nabla G$$

(9) →
$$\boxed{\begin{aligned} \tilde{H} &= -\sum_i p(x_i) \ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right) \\ &= -\int p(x) \ln(p(x)) dx + \lambda\left(\int p(x) dx - 1\right) \end{aligned}} \longrightarrow (10)$$

Example → Goal: find the model of Normal distribution p(x)

ormal distri

Constrains :

$$\left.\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= 1 \\ \int_{-\infty}^{\infty} x p(x) dx &= x \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx &= \sigma^2 \end{aligned}\right\} (11)$$

(11) →
(10) →

$$\hat{H} = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1\left(\int_{-\infty}^{\infty} p(x) dx - 1\right)$$

$$+ \lambda_2\left(\int_{-\infty}^{\infty} x p(x) dx - \mu\right) + \lambda_3\left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx^2 - \sigma^2\right)$$

$$\nabla \tilde{H} = 0 = \overset{\neq 0}{p(x)}\left[\overset{=0}{-\ln p(x) + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2}\right]$$

$$p(x) = \exp\left[\lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right] = \cdots = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

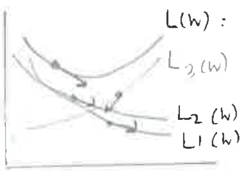
Stirling Approximation $\ln N! \simeq N \ln N - N$

$$\ln N! = \ln 1 + \ln 2 + \ldots + \ln N$$

$$\ln N! - \tfrac{1}{2}(\ln 1 + \ln N) = (\ln 1 + \ldots + \ln N) - \tfrac{1}{2}(\ln 1 + \ln N)$$

Lecture 9 — Decision Trees

Review → Stochastic gradient descent



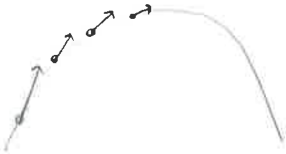
Goal: $\boxed{\min_w L(w) = \sum_{i=1}^{N} L_i(w)}$

Lecture 6 — Gradient descent ; bias-Varian tradeoff

$\underline{(X^TX)^{-1}X} = X^+$ pseudoinverse

Problem w/ X^+ →
- ▹ Least square closed form sol^n (Lee 5.9) → $\boxed{\overset{*}{w} = (\lambda I + X^TX)^{-1} X^T y}$
- ▹ some times 'X^+' too large to compute

Alternative Numerical Opti: 'gradient descent' →



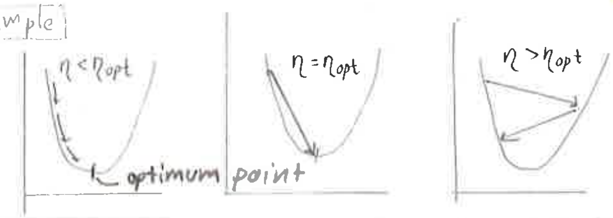
- ▹ Gradient ascent = 'uphill climbing'
- ▹ Gradient descent : 'down hill'

Mech $t = 0$, $w^{(t)}$

$g^{(t)} = \nabla f(\underline{X}, \underline{y} ; w^{(t-1)})$

update model $\boxed{w^{(t)} = w^{(t-1)} - \eta g^{(t)}}$ learning rate } (Lec 6.1)

Example



Bias of an estimator → $\boxed{\text{bias}(\hat{\theta}) \triangleq \mathbb{E}_x [\hat{\theta} - \theta]}$ → (Lec 6.2)

different of predicted from correct

$\mathbb{E}[\hat{\mu}_{ML}] = \frac{1}{N} \Sigma_i^N x_i = \mu$ ✓

$\mathbb{E}[\hat{\sigma}^2_{ML}] = \frac{1}{N} \Sigma_i (x_i - \hat{\mu})^2 = \frac{N-1}{N} \sigma^2$ (book 6)

Consistency of an estimator → Estimator $\hat{\theta}$ is consistent if $\boxed{\lim_{N \to \infty} \hat{\theta}_N = \theta}$ → (Lec 6.3)

$\hat{\sigma}^2_{ML}$ bias , but consistent w/ σ^2

timation &
egression →

True Model $y = F(x) + v$; v: noise with mean: 0

approximate 'F' by $\overset{\text{function}}{f(x;\hat{w})} \in \overset{\text{function space}}{\mathcal{F}}$: estimate $\hat{w}$ from x

- $\hat{f}(\vec{x}) = f(\vec{x};\hat{w})$ estimation based on this $\vec{x}$
- $\bar{f}(\vec{x}) = \mathbb{E}_x[f(\vec{x},\hat{w})]$ avg estimate over training sets $\vec{x}$
- $f^*(\vec{x}) = f(\vec{x};\ \arg\min_x \mathbb{E}_{p(\vec{x},y)}[(y - f(\vec{x};w))^2])$ the best estimate $f \in \mathcal{F}$

→ (Lec 6.4)

ias + Variance →

▷ $\mathbb{E}_x[\text{square loss}] = \mathbb{E}_x[(y_0 - \hat{f}(x_0))^2] = \underbrace{(y_0 - \bar{f}(x_0))^2}_{\square} + \underbrace{\mathbb{E}_x[(\hat{f}(x_0) - \bar{f}(x_0))^2]}_{\text{Variance}}$

→ (Lec 6.5)

□ $(y_0 - \bar{f}(x_0))^2 = \underbrace{(y_0 - \overset{\text{true model}}{F}(x_0))^2}_{\text{noise}} + \underbrace{(F(x_0) - \bar{f}(x_0))^2}_{\text{bias}^2}$

noise: difference btw observe value and true model

bias²: different btw avg estimation and true model

$(y_0 - F(x_0))^2$ **noise** : irreducible (independent to data)

$(F(x_0) - \bar{f}(x_0))^2$ **bias²** = different f & F

$\mathbb{E}_x[(\hat{f}(x_0) - \bar{f}(x_0))^2]$ **Variance** =

} try to minimize

isher
Information 1 →
just information)

measuring information that observe $\vec{x}$ carry about unknown parameter $\vec{\theta}$

$$I[\theta] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\vec{x};\theta)\right)^2 \Big|\ \theta\right] = \int \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx$$

given θ → no ∫dθ

$$= \int \left(\frac{1}{f^2}\left(\frac{\partial}{\partial\theta} f\right)^2\right) f\,dx$$

→ (Lec 6.6)

$$I[\theta] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\Big|\ \theta\right] = \int\left(-\frac{\partial^2}{\partial\theta^2}\{\log f(x;\theta)\}\right) f(x;\theta)\,dx$$

$$= \int\left(-\frac{\partial}{\partial\theta}\left\{\frac{1}{f}\frac{\partial}{\partial\theta}f\right\}\right) f\,dx = \int\left\{\underbrace{\frac{1}{f^2}\left(\frac{\partial}{\partial\theta}f\right)^2}_{\boxed{1}} - \underbrace{\frac{1}{f}\frac{\partial^2}{\partial\theta^2}f}_{\boxed{2}}\right\} f\,dx$$

$$I[\theta] = \underbrace{\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(\vec{x};\theta)\Big|\ \theta\right]}_{\boxed{1}} - \underbrace{\frac{\partial^2}{\partial\theta^2}\int \overset{1}{\cancel{f(x;\theta)\,dx}}}_{0}$$

$$I[\theta] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\Big|\ \theta\right]$$

(lec 6.6)

isher Info

▷ $\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\Big|\ \theta\right] = \int\left(\frac{1}{\cancel{f}}\frac{\partial f}{\partial\theta}\right)\cancel{f}\,dx = \frac{\partial}{\partial\theta}\int f\,dx = \frac{\partial}{\partial\theta}1 = 0$

Fisher Cramér-Rao Bound

3. Prove $\mathrm{Var}[\hat{\theta}] \geq \frac{1}{I(\theta)}$ → (lec 6.7)

Proof

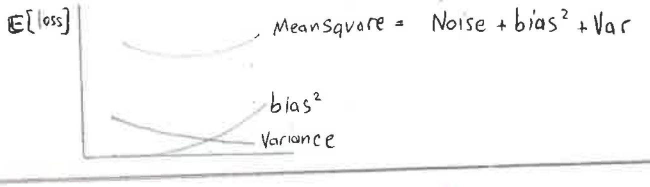
→ unbiased estimator $\hat{\theta}(X) \Rightarrow E[\hat{\theta}(X)-\theta|\theta] = \int (\hat{\theta}(x)-\theta) f(x;\theta)\,dx = 0$

$$0 = \int \frac{\partial}{\partial\theta}(\hat{\theta}(x)-\theta) f(x;\theta)\,d\theta = \int (\hat{\theta}(x)-\theta)\underbrace{\frac{\partial f}{\partial\theta}}_{f\frac{\partial \log f}{\partial\theta}}dx - \underbrace{\int f dx}_{1}$$

$$\int (\hat{\theta}-\theta)\, f\,\frac{\partial \log f}{\partial\theta}\,dx = 1$$

$$1 = \left(\int \left[(\hat{\theta}-\theta)\sqrt{f}\right]\cdot\left[\sqrt{f}\,\frac{\partial \log f}{\partial\theta}\right]dx\right)^2 \leq \overbrace{\left[\int(\hat{\theta}-\theta)^2 f\,dx\right]}^{\mathrm{Var}[\theta]}\overbrace{\left[\int\left(\frac{\partial \log f}{\partial\theta}\right)^2 f\,dx\right]}^{I[\theta]}$$

similar to $|\mathrm{cov}[A,B]|^2 \leq \mathrm{Var}[A]\,\mathrm{Var}[B]$



Lecture 7

Logistic Regression

Classification as regression → $y, \hat{y}$ are classes Ex. $\{-1, 1\}$

$f(x, \hat{w}) = w_0 + \hat{w}\cdot \underline{x}$ = function ≠ classes

∴ decision rule $\hat{y} = 1$ if $f(x;\hat{w}) \geq 0$, otherwise $\hat{y} = -1$ (L 7.1)

→ $\hat{y} = \mathrm{sign}(f(x;\hat{w})) = \mathrm{sign}(w_0 + \hat{w}\cdot x) = h(x)$

→ in here 0 = **decision boundary**

Loss calculation → (L7.1) $L(h(\underline{x}), y) = L(\hat{y}, y) = \begin{cases} 0 & \text{if } h(\underline{x}) = y \\ 1 & \text{if } h(\underline{x}) \neq y \end{cases}$

Risk = Expected Loss → $R(h) = E_{x,y}[L(h(\underline{x}), y)] = \int_{\underline{x}} \sum_{C=1}^{C\,(\text{classes})} \underbrace{L(h(\underline{x}), c)\,p(y=c|x)}_{\text{minimize} = R(h|x)}\overbrace{p(x)}^{p(x,y=c)}d\underline{x}$ → (L 7.2)

$= \int_{\underline{x}} R(h|\underline{x})\,p(x)\,dx$

only count when $L(\hat{y},y)=1$

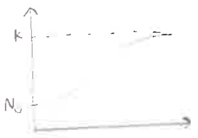
$$R(h|x) = \sum_{c\neq h(x)}^{C} (1)\,p(y=c|\underline{x}) = ?\; 1 - p(y=h(\underline{x})|\underline{x})$$

timum
Rule →→ $h(x) = \underset{c}{\arg\max}\ p(y=c|x)$

$$h(x) = c^* \leftrightarrow \frac{p(y=c^*|x)}{p(y=c|x)} \geq 1 \quad \text{or} \quad \ln\left\{\frac{p(y=c^*|x)}{p(y=c|x)}\right\} \geq 0 \quad \forall c$$

(L7.3)

logistic
lodel →



$$\frac{dN}{dt} = rN\left(1-\frac{N}{K}\right) \quad \text{→logistic}$$

$$r = \left(\frac{1}{N} - \frac{-\frac{1}{K}}{1-\frac{N}{K}}\right)\frac{dN}{dt} = $$

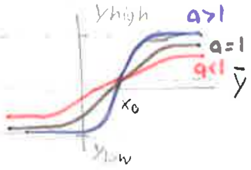
$$r(t)+C = \ln(N) - \ln\left(1-\frac{N}{K}\right) = \ln\left(\frac{N}{1-\frac{N}{K}}\right) \Rightarrow \left(\frac{N}{1-\frac{N}{K}}\right) = $$

$$\boxed{N(t) = \frac{1}{C_3 e^{-rt} + \frac{1}{K}} = \frac{N_0 K}{\underbrace{(K-N_0)e^{-rt}}_{\text{logistic}} + N_0}}$$ → (L7.4)

lationship
σ(x) and p(y|x) → simple logistic (growth) model

simple $\boxed{\sigma(x) = \left(\frac{1}{1+e^{-x}}\right)}$; $\begin{cases}\sigma(-\infty) = 0 \\ \sigma(0) = \frac{1}{2} \\ \sigma(\infty) = 1\end{cases}$

complex $\sigma(\ ) = \frac{(y_{high} - y_{low})}{1+e^{-a(x-x_0)}} + y_{low}$ →(L7.5)



ecision
Boundary →

(L7.3)

Binary

$$\triangleright \ln\left\{\frac{p(y=1|x)}{p(y=0|x)}\right\} \geq \overset{\text{Boundary}}{0} = w_0 + W \cdot X$$

Y = vector	y_i = value
X = Matrix	x_i = vector

$$\triangleright \frac{p(y=1|x)}{1-p(y=1|x)} = e^{(w_0+W\cdot X)} = 1 \quad \Rightarrow$$

$$\therefore \Rightarrow p(y=1|x) = \left(\frac{1}{1+e^{-(w_0+\bar{W}\cdot X)}}\right) = \overset{(L7.5)}{\sigma(w_0 + \bar{W}\cdot\bar{X})} = \frac{1}{2} = \sigma(0)$$

$$p(y=0|x) = \left\{\frac{e^{-(w_0+\bar{W}\cdot\bar{x})}}{1+e^{-(w_0+\bar{W}\cdot\bar{x})}}\right\} = 1-\sigma(w_0+\bar{W}\cdot\bar{X}) = \frac{1}{2}$$ ~~→(L7.6)~~

} (L7.6)

$$p(y_i|\bar{x}_i;W) = \overset{y_i=1}{\sigma(w_0+\bar{W}\cdot x_i)^{y_i}} \left(1-\sigma(\overset{y_i=0}{w_0}+\bar{W}\cdot\bar{x}_i)\right)^{1-y_i}$$

$$\log p(\bar{Y}|\underline{X};\bar{w}) = \sum_{i=1}^{N} \log p(y_i|\bar{x}_i;w)$$

Lecture 8

Regulation in logistic regression; stochastic gradient descent; Softmax

optimal regressor $\hat{y} = E[y|x]$

optimal classifier $\hat{y} = \underset{c}{\mathrm{argmax}}\ p(y=c|x)$

Review: Logistic Regression

- log-odds as a function of $\underline{x}$: Binary $\log \frac{p(y=1|x)}{p(y=0|x)} = f(\phi(\underline{x});\vec{w}) = 0$

$$p(y=0|x) \to (1-p(y=1|x)) \qquad \frac{1}{1+e^{-x}}$$

- $$p(y=1|\underline{x}) = \frac{1}{1+e^{(-f(\phi(x);\underline{w}))}} = \frac{1}{1+e^{(-w_0-\underline{w}\cdot\underline{x})}}$$

can be non linear $\phi[x] = [1, x_1, x_2, x_1x_2]$

Gradient Descent

= 1st order iterative optimization algorithm

(L7.7) → $\log p(\bar{Y}|\underline{x};\bar{w}) = \sum_{i=1}^{N}[y_i \log \sigma(w_0+\bar{w}\cdot\bar{x}_i)\ (1-y_i)\log(1-\sigma(w_0+\bar{w}\cdot\bar{x}_i)]$

(L7.8)

$$\frac{\partial}{\partial w_0}\log p(\bar{Y}|\underline{x};\bar{w}) = \sum_{i=1}^{N}\left[\frac{y_i\ \cancel{\sigma}(1-\sigma)}{\cancel{\sigma}}\left(\cancel{\frac{\partial w_0}{\partial w_0}}\right) + \frac{(1-y_i)(-\sigma)\cancel{(1-\sigma)}}{\cancel{(1-\sigma)}}\cancel{\frac{\partial w_0}{\partial w_0}}\right]$$

$$= \sum_{i=1}^{N}[y_i - \cancel{y_i\sigma} - \sigma + \cancel{y\sigma_i}] = \sum_{i=1}^{N}[y_i - \sigma(w_0+\bar{w}\cdot\bar{x}_i)] = 0 \text{ — (L7.9A)}$$

$$\frac{\partial}{\partial w_j}\log p(\bar{y}|\bar{x};\bar{w}) = \sum_{i=1}^{N}\left[\frac{y_i\ \sigma(1-\sigma)}{\cancel{\sigma}}\frac{\partial\bar{w}\cdot\bar{x}_i}{\partial w_j} + \frac{(1-y_i)(-\sigma)\cancel{(1-\sigma)}}{\cancel{(1-\sigma)}}\left(\frac{\partial\bar{w}\cdot\bar{x}_i}{\partial w_j}\right)\right]$$

$$= \sum_{i=1}^{N}[(y_i - \sigma(w_0+\bar{w}\cdot\bar{x}_i))x_{ij}] \text{ — (L7.9B)}$$

Updated W

$$W_{new}^{(t+1)} = \bar{w}^{(t)} + \eta\frac{\partial}{\partial\bar{w}}\log p(\underline{X};\bar{w})$$

$$= \bar{w}^{(t)} + \eta\sum_{i=1}^{N}(y_i - \sigma(w_0+\bar{w}\cdot\bar{x}_i))\begin{bmatrix}1\\x_i\end{bmatrix} \text{ — (L7.9)}$$

(L7.9A)

(L7.9B)

tochastic adient scent: tuition

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial w}L(y_i,\bar{x}_i;\bar{w}) \approx \frac{\partial}{\partial w}L(y_t,\bar{x}_t;\bar{w})$$

$L=\Sigma_i L_i$
L_1
L_2

$\nabla L(w) \approx N \nabla L_1(w)$

$w := w + \eta \frac{\partial}{\partial w}\log p(y_i|\bar{x}_i;\bar{w})$ ↙ L7.9 ; $\log p(\bar{Y}|\underline{X};\bar{w})$ for full update

$w := w + \eta\,(y_i - \sigma(\bar{w}'\bar{x}_i))\,x_i$

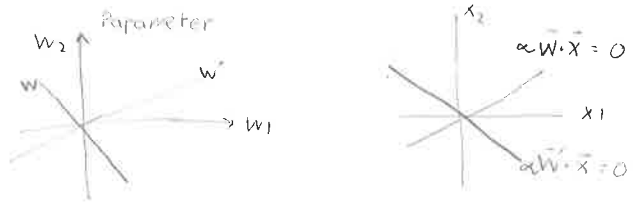
$w := w + \eta(\underline{y} - \sigma(\bar{w}^T\underline{x}))\,\underline{x}$ for full update

unlizing he log-likelihood surfa

Example 2D, $w_0 = 0$

$\therefore \hat{p}(y=1|x) = \sigma(\cancel{w_0} + \bar{w}\cdot\underline{x}) = \sigma(w_1x_1 + w_2x_2)$

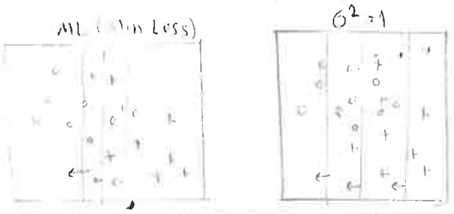
Mapping b.c. to parameter



$\log p(Y|X,\bar{w};\sigma) = \log p(\bar{Y}|\underline{X},\bar{w}) + \log p(\bar{w};\sigma)$ (lec 5)

$$= \sum_{i=1}^{N}\log p(y_i|\bar{x}_i,\bar{w}) - \left(\frac{1}{2\sigma^2}\sum_{j=1}^{d}w_j^2\right) + \text{const}(w)$$

similar to penalty $-\frac{1}{2\sigma^2}\|w\|_2^2$ λ



Soft Max function

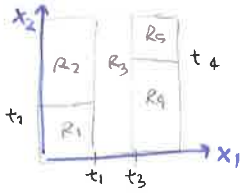
$$\sigma(\bar{z})_i = \frac{e^{z_i}}{\Sigma_{j=1}^{k}e^{z_j}} \quad \text{for } i=1,\dots,k \text{ and } \bar{z}=(z_1,\dots,z_k)\in\mathbb{R}^k$$

$$p(y=c|x) = \frac{e^{\bar{w}_c\cdot\bar{\phi}(x)-a}}{\sum_{k=1}^{c}e^{\bar{w}_k\cdot\bar{\phi}(x)-a}} \;;\; a = \max_c \bar{w}_c\cdot\bar{\phi}(x) \quad \therefore \frac{e^{\bar{w}_c\cdot\bar{\phi}(x)-a}}{\square} = \frac{e^{\le 0}}{\square} = \frac{\le 1}{\square}$$

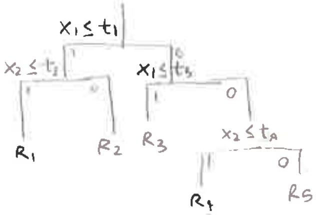
Lecture 9

Decision Tree

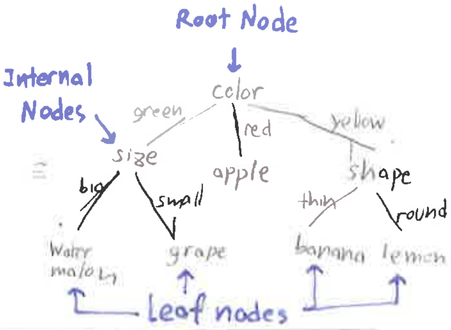
Space Partition ↣ Ex 2D



≡



≡



Regression Tree ↣ Algorithm → CART (Classification and Regression Trees)

Usages 1) To calculate Prob that a given data belong to each class

2) To classify the new data to the most likely class