

## ML\_5

Singular value decomposition  $\rightarrow$  ماتریس را به ۳ ماتریسِ خارکن.

$$X \in \mathbb{R}^{n \times d} \rightarrow \boxed{X = UDV^T} \rightarrow \begin{cases} U \in \mathbb{R}^{n \times d} \\ D \in \mathbb{R}^{d \times d} \quad (\text{diagonal matrix}) \\ \text{(singular values)} \end{cases}$$

$$U^T U = I \rightarrow \text{orthonormal matrix}$$

$$V^T V = V V^T = I \rightarrow \text{orthogonal matrix}$$

ridge regression

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

ridge

effective degree of freedom:

$df(\lambda) \leftarrow LR$  مقدار اعتماد

$$df(\lambda) = \text{tr}(H) = \text{tr}(X(X^T X + \lambda I)^{-1} X^T) = \sum_{i=1}^d \frac{d_{ii}^{-2}}{d_{ii}^{-2} + \lambda}$$

~~intercept~~  $\downarrow$  خوفن

$df(0) = d$   $\rightarrow$  ridge مینیمیز کرزا -

$df(\infty) = 0$  -> ridge و linear regression

مین اسے کیا اگر  $\lambda$  مینیمیز کر جاؤ.



+ کاربردی در جی آزادی در تابعی طبیعی

AIC

BIC

GCV

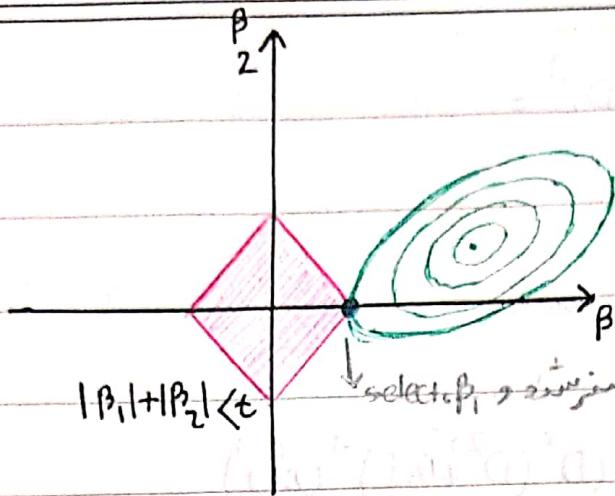
## LASSO

آن بحاجی اینکه نرم ۲ باشد، نرم ۱ است. ridge penalty خطای مربعات مینیموم

$$L(\beta) = (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_1$$

$$\sum_{i=1}^d |\beta_i|$$

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} L(\beta)$$



\* چون مرتبه مسقی نزدیک است، پس از LASSO، ridge عنوان مسقی کرن و  
کتابی صفر حراردار.

+ no analytical solution

+ convex optimization  $\rightarrow$  نیز ترکیبی Gradient descent ؟

است و صفحه خارجی آن  $\text{convex}$  است. (الورقة تلقياً خارجية)

[ لغة خواص ]

$$\lambda \text{ selection : } \hat{\lambda}^* = \arg \min_{\lambda} \hat{\Sigma}^{CV}(\lambda) \rightarrow \text{lasso, ridge}$$

## 1 Maximum Likelihood Estimation

موسیٰ بن موسیٰ بن ابی حمّان سارا احمد

+ Given data  $\rightarrow D = \{x[1], \dots, x[n]\}$ ,  $x \in \mathbb{R}^d$

(دریک کھرکام data in)

توزیع  $P(x; \theta)$  کے سرکی پا را ہر دو ایسے ← Sampling اس کے توزیع پر ماحضن ہی نہیں۔

A hand-drawn circle with a horizontal line through its center, representing a population.

point:

Condition

$x; \theta$

$x|\theta \rightarrow$  conditional

$P_{\theta}(x)$

Wish

ex: goal: to estimate  $\theta$

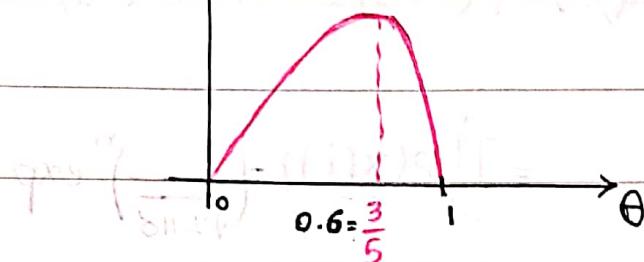
The coin example  $\rightarrow$  + outcome H/T  $\rightarrow \theta = P(X=H)$

+ Data  $\rightarrow D = \{H, T, T, H, H\}$

$$P(D|\theta) = P(H)P(T)P(T)P(H)P(H)$$

$$= \theta(1-\theta)(1-\theta)\theta\theta = \boxed{\theta^3(1-\theta)^2}$$

$P(D|\theta)$



$\rightarrow$  - Data  $\tilde{\theta}$   $\approx$   $\theta$  j.w.  
in explain  $\tilde{\theta}$

Def: Likelihood function  $\rightarrow L(\theta)$

$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(\theta)$$

ex: Cont n tosses, m heads

$$L(\theta) = \theta^m (1-\theta)^{n-m}$$

$$\boxed{l(\theta) = m \log \theta + (n-m) \log (1-\theta)}$$

↓  
log likelihood

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{m}{\theta} - \frac{n-m}{1-\theta} = 0 \rightarrow \hat{\theta} = \frac{m}{n}$$

↙

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = \frac{m}{\theta^2} - \frac{n-m}{(1-\theta)^2} < 0$$

## MLE for Univariate Gaussian

$$x \sim N(\mu, \sigma^2) \quad \underline{\sigma^2}: \text{Known}$$

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \leftarrow \text{Gaussian (جودج زفال)$$

$$D = \{x[1], \dots, x[n]\} \xrightarrow{\text{likelihood}} L(\mu) = P(x[1]) \times \dots \times P(x[n])$$

$$= \prod_{i=1}^n P(x[i])$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x[i]-\mu)^2\right)$$

maximize  $L(\mu)$  (أقصى  $L(\mu)$ )  
 - count  $\log$   $\Rightarrow l(x) = \log L(\mu)$

$$= -n \log \sqrt{2\pi} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x[i]-\mu)^2$$

$$\rightarrow \frac{\partial L(\mu)}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^n -2(x[i]-\mu) = 0 \Rightarrow n\mu = \sum_{i=1}^n x[i]$$

$$\hat{\mu}_{ML} = \bar{x}$$

$$\frac{\partial^2 l(\mu)}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0 \Rightarrow \text{max point}$$

HW: MLE for multinomial dist  
 (lagrange multiplier)

## 2 MAP estimate (Maximum A posteriori)

جز آمار کلاسیک محبوب می شود. جای تجربی درین از data بپارافر، دادهای خود اصلی وجود  
likelihood based (1)

Random ← درین ربطه پراور حاصلت bayesian based (2)  
دارند.

هری از وسیع Bayesian هست و خوب نسبت به MLE (ارد)  
من است که میتوان prior knowledge را هم با آن اختلاف نکرد.

برای اگر  $n$  (سازه data) کم باشد، میتواند باعث باشد.  
متاثر شدن از این اتفاق،  $\hat{\theta} = \text{Head}$  آن است، اما این -  
این ایجاد میشود.

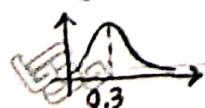
و میتواند باعث باشد.

وسیع کسی ساخته شده برای MAP هم وجود ندارد، بدین جایی خواهد شد.

+ MAP estimate.

- prior knowledge

$$D = \{x[1], \dots, x[n]\} \quad \theta: \text{random variable}$$



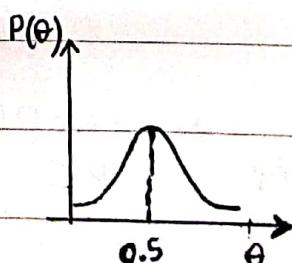
prior dist.  $\leftarrow P(\theta)$

bayesian پیشی از  $\rightarrow$  posterior dist.  $\leftarrow P(\theta|D)$

likelihood  $\leftarrow P(D|\theta)$

$$P(D)$$

$$P(D, \theta)$$



کی از تقدیرات اسی کہ اس دستگاه باریخاد متبی (MLE) دلیل این است که چون  $\theta$  نباید تغیر پذیر باشی  
است، ھر مقداری سین و اس تو اندازبند دین بر این نک توزیع چور دارد.  
اسن حالت را لازم و فواید داشتم. point estimate ، MLE  $\rightarrow$   
توزیع را به دست آوریم.

$$\text{Goal: to estimate } \theta \rightarrow \hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(\theta | D)$$

$$\text{Baye's rule} \rightarrow P(\theta | D) = \frac{P(\theta) P(D|\theta)}{P(D)} \times P(\theta) P(D|\theta)$$

ex: MAP estimate for a univariate Gaussian

$$x | \mu \sim N(\mu, \sigma^2) \quad , \quad D = \{x[1], \dots, x[n]\}$$

fixed

$$\text{* likelihood} \rightarrow P(D|\mu) \propto \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x[i] - \mu)^2 \right)$$

$$\text{* prior: } \mu \sim N(\mu_0, 1) \rightarrow P(\mu) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} (\mu - \mu_0)^2 \right)$$

$$P(\mu | D) \propto P(\mu) P(D|\mu)$$

$$\hat{\mu}_{\text{MAP}} = \arg \max_{\mu} P(\mu | D) = \arg \max_{\mu} [\log P(\mu) + \log P(D|\mu)]$$

$$\frac{\partial P(\mu | D)}{\partial \mu} = 0 \rightarrow \frac{1}{\sigma^2} \sum_{i=1}^n (x[i] - \mu) - (\mu - \mu_0) = 0$$

$$\hat{\mu}_{MAP} = \frac{n}{n + \sigma^2} \bar{x} + \frac{\sigma^2}{n + \sigma^2} \mu_0$$

prior knowledge

Convex combination of  $\bar{x}$  and  $\mu_0$

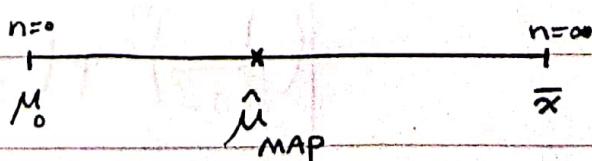
حقیقت خود ری نہ دریافت بے دست آور خواهد بود اما جی شود و  
ووچے  $n \downarrow$  باشد، میں خوب جواب عن دهد، بنابرین اترانک تحریک شوند و  
برای  $n$   $\hat{\mu}_{MAP}$   $\rightarrow$  prior knowledge

اگر  $n$  وجود ندارد راستاً میں دهد و اتر کے زمانہ باشد، پس از این  $\hat{\mu}_{MAP}$   $\rightarrow$   $\sigma^2$   
بستر است و مترجع ہوں ہن ان احتمال کردن.

+ convex Combination of  $\bar{x}$  and  $\mu_0$ .

$$\Rightarrow \alpha \bar{x} + (1-\alpha) \mu_0 \quad \text{if } \alpha \in [0, 1]$$

in this case:



3 + True Bayesian approach (latter) ✓  $\rightarrow P(\theta | D)$

## Linear regression - Probabilistic View

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d + \varepsilon$$

white noise

$$\varepsilon \sim N(0, \sigma^2)$$

(ε) خونتی noisy و بودادار در دروغ نمایان است.  
حالا جواب می‌باشد که عین پارامتر، عین بگیر.

$$D = \{(x[1], y[1]), \dots, (x[n], y[n])\}$$

x: fixed

y: random

$\beta$ : fixed  $\rightarrow$  true parameter

$$\hat{\beta}: \text{random} \rightarrow Y | X \sim N(\beta^T x, \sigma^2)$$

Likelihood function  $\rightarrow L(\beta) = P(y[1], \dots, y[n] | x[1], \dots, x[n], \beta)$

$$= \prod_{i=1}^n p(y[i] | x[i], \beta)$$

$$= \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y[i] - \beta^T x[i])^2 \right)$$

Maximization  $\rightarrow \ell(\beta) = \log L(\beta)$

$$= C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y[i] - \beta^T X[i])^2$$

least squares error

+ maximizing  $\ell(\beta)$  is equal to minimizing least squares loss.

## ML-6

### Lagrangian

$$\begin{array}{l} \min / \max f(x) \\ \text{s.t. } g(x)=0 \end{array} \quad \left\{ \Rightarrow L(x, y, \lambda) = x + y + \lambda(g(x)) \right. \\ \quad \quad \quad \left. \frac{\partial L}{\partial x} = 0, \frac{\partial L}{\partial y} = 0, \frac{\partial L}{\partial \lambda} = 0 \right. \checkmark$$

ex. 1  $\max \rightarrow x+y$        $\left\{ \begin{array}{l} \text{s.t. } x^2+y^2=1 \\ g(x)=x^2+y^2-1=0 \end{array} \right.$        $\left. \begin{array}{l} (1) \\ \Rightarrow L(x, y, \lambda) = x+y+\lambda(x^2+y^2-1) \\ = x+y+\lambda x^2+\lambda y^2-\lambda \end{array} \right\}$

مکرر (2)  $\frac{\partial L}{\partial x} = 1+2\lambda x = 0 \rightarrow x = \frac{-1}{2\lambda}$

$$\frac{\partial L}{\partial y} = 1+2\lambda y = 0 \rightarrow y = \frac{-1}{2\lambda}$$

$$\frac{\partial L}{\partial \lambda} = x^2+y^2-1=0 \rightarrow \left(\frac{-1}{2\lambda}\right)^2 + \left(\frac{-1}{2\lambda}\right)^2 - 1 = 0$$

$$\rightarrow \frac{1}{4\lambda^2} + \frac{1}{4\lambda^2} - 1 = 0 \rightarrow \lambda \uparrow = +\frac{\sqrt{2}}{2} \\ \downarrow = -\frac{\sqrt{2}}{2}$$

جاوده (3) Case 1  $\lambda = \frac{+\sqrt{2}}{2} \rightarrow \left\{ \begin{array}{l} x = \frac{-1}{\sqrt{2}} \\ y = \frac{-1}{\sqrt{2}} \end{array} \right.$

پارای این  $x$  و  $y$  و  $\lambda$  جواب  
Case 2  $\lambda = -\frac{\sqrt{2}}{2} \rightarrow \left\{ \begin{array}{l} x = \frac{1}{\sqrt{2}} \\ y = \frac{1}{\sqrt{2}} \end{array} \right. \rightarrow \checkmark \text{ سو} \max$

$$\text{ex.2} \quad \max \rightarrow -x^2 - y^2 + 4 \quad \left. \begin{array}{l} \\ \text{s.t. } x+y=1 \\ g(x)=x+y-1=0 \end{array} \right\} \rightarrow L(x, y, \lambda) = -x^2 - y^2 + 4 + \lambda(x+y-1)$$

$$= -x^2 - y^2 + 4 + \lambda x + \lambda y - \lambda$$

$$\frac{\partial L}{\partial x} = 0 \rightarrow -2x + \lambda = 0 \rightarrow x = \frac{\lambda}{2}$$

$$\frac{\partial L}{\partial y} = 0 \rightarrow -2y + \lambda = 0 \rightarrow y = \frac{\lambda}{2}$$

$$\frac{\partial L}{\partial \lambda} = 0 \rightarrow x+y-1=0 \rightarrow \boxed{\lambda=1} \rightarrow x=y=\frac{1}{2} \rightarrow \boxed{\max = \frac{1}{2} + 4 = 3.5}$$

+ point: If there are several  $g(x)$  functions, we add  $\lambda$  for each  $g(x)$

$$\text{ex: } -x^2 - y + 2$$

$$\text{s.t. } \left. \begin{array}{l} x+y=1 \quad g_1(x)=x+y-1=0 \\ x-y=1 \quad g_2(x)=x-y-1=0 \end{array} \right\} \rightarrow L(x, y, \lambda_1, \lambda_2) =$$

$$= -x^2 - y + 2 + \lambda_1(x+y-1) + \lambda_2(x-y-1)$$

## Bayesian interpretation of linear regression:

Review:

$$\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \sum_{i=1}^n (y[i] - \beta^T x[i])^2 + \lambda \sum_{j=1}^d \beta_j^2$$

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d + \epsilon$$

$$Y|x; \beta \sim N(\beta^T, \sigma^2) \Rightarrow \text{likelihood}$$

$$\beta_i \sim N(0, T^2) \text{ iid prior dist.}$$

$$P(\beta) = P(\beta_1) \cdot P(\beta_2) \cdots P(\beta_d)$$

$$\text{likelihood: } p(y[1], \dots, y[n] | x, \beta) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y[i] - \beta^T x[i])^2\right)$$

MAP estimate:

$$\text{prior} \rightarrow P(\beta) \propto \exp\left(-\frac{1}{2T^2} \sum_{i=1}^d (\beta_i - 0)^2\right)$$

$$\text{posterior} \rightarrow P(\beta|x, y) \propto \text{likelihood} \times \text{prior}$$

$$\log P(\beta|x, y) \propto \log P(y|x, \beta) + \log P(\beta)$$

$$\frac{-1}{2\sigma^2} \sum_{i=1}^n (y[i] - \beta^T x[i])^2 + \frac{1}{2T^2} \sum_{i=1}^d \beta_i^2$$

MAP  $\rightarrow \arg \max \log P(\beta | x, y) =$

$$= \arg \min \sum_{i=1}^n (y[i] - \beta^T x[i])^2 + \boxed{\frac{\sigma^2}{T^2} \sum_{i=1}^d \beta_i^2}$$

$\hat{\alpha} + \hat{\beta}$  ridge  $\lambda$  در صفر رفته شود دو واقع این عبارت هان