18.650 Spring 2024

LECTURE 1 2/5/24 (Missed)

distributions $\underset{\text{stats}}{\overset{\text{probability}}{\rightleftarrows}}$ data

LLN: $\bar{X}_n \to \mu$ as $n \to \infty$

CLT: $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \rightsquigarrow N(0,1)$

$\uparrow$

X_i's are i.i.d mean μ, variance σ^2

LECTURE 2 2/7/24 (Missed)

- histograms
- summary statistics
  - → mean, stder, median, quantiles, IQR, outliers

100(1-α) percentile → quantile of order 1-α $= q_\alpha$ = value that α of data exceeds

- → boxplots, scatterplots.

LECTURE 3 2/9/24 (Missed)

$X_n \xrightarrow{\mathbb{P}} X$ as $n \to \infty$ if $\forall \varepsilon > 0$,

"convergence in probability" $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.

ex if $X \sim \text{Ber}(1/2)$ $\forall n$. Then $X_n \xrightarrow{\mathbb{P}} \text{Ber}(1/2)$

because $\mathbb{P}(|X_n - X| > \varepsilon)$ for $\varepsilon \in (0,1)$

$= \mathbb{P}(X_n = 0 \cap X = 1) + \mathbb{P}(X_n = 1 \cap X = 0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \not\to 0$

but $X_n \rightsquigarrow X$ as $n \to \infty$ if $\forall x$ for which $x \mapsto \mathbb{P}(X \le x)$ cdf is continuous

"convergence in distribution" $\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x)$ as $n \to \infty$.

$$X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \rightsquigarrow X$$

Lemma If $X_n \rightsquigarrow c$ for constant c, then $X_n \xrightarrow{\mathbb{P}} c$.

Pf $\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}(X_n \ge \varepsilon + c) + \mathbb{P}(X_n \le c - \varepsilon)$

$\to 0 + 0 = 0$ as $n \to \infty$.

if $X_n \xrightarrow{\mathbb{P}} X$, $Y_n \xrightarrow{\mathbb{P}} Y$ $\Rightarrow$ $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$, $X_n Y_n \xrightarrow{\mathbb{P}} XY$

if $X_n \xrightarrow{\mathbb{P}} X \Rightarrow g(X_n) \xrightarrow{\mathbb{P}} g(X)$

if $X_n \rightsquigarrow X \Rightarrow g(X_n) \rightsquigarrow g(X)$

Slutsky's Theorem

if $X_n \rightsquigarrow X$, $Y_n \xrightarrow{\mathbb{P}} c$ $\Rightarrow$ $X_n + Y_n \rightsquigarrow X + c$, $X_n Y_n \rightsquigarrow Xc$

Delta Method

$\frac{\sqrt{n}}{\sigma}(Y_n - \mu) \rightsquigarrow Y \sim N(0,1)$

implies for any diff'able g

$\frac{\sqrt{n}}{\sigma}(g(Y_n) - g(\mu)) \rightsquigarrow N(0, g'(\mu)^2)$

LECTURE 4 2/12/24 (Missed)

Gaussian $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 68-95-99.7 rule

$$X \mapsto Z = \frac{X-\mu}{\sigma}$$ z-score "standardization"

standard normal: $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

LECTURE 5 2/14/24 1PM

random matrix $X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \in \mathbb{R}^k$

$$\mathbb{P}(X \in A) = \int_A f(x_1, x_2, \ldots x_k)\, dx_1 dx_2 \ldots dx_k$$

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_k] \end{pmatrix} = \mu \qquad Cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$$

covariance matrix of X

$$\mathbb{V}[X] = \Sigma = (Cov(X_i, X_j))_{i,j}$$

$$\Rightarrow \Sigma_{ii} = Cov(X_i, X_i) = Var(X_i)$$

note $X^T X$ = inner product = scalar

XX^T = matrix $\in \mathbb{R}^{k \times k}$ = outer product

$$\Sigma := \mathbb{V}[X] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T = \mathbb{E}[XX^T] - \mu\mu^T$$

Σ^{-1} = precision matrix

Linear Transformations

Let $X \in \mathbb{R}^k$ be random vector, $\mathbb{E}[X] = \mu$, $\mathbb{V}[X] = \Sigma$.

Let $a \in \mathbb{R}^k$ be deterministic vector.

$$\Rightarrow \mathbb{E}[a^T X] = a^T \mu \quad \text{(trivial)}$$

$$\mathbb{V}[a^T X] = a^T \Sigma a$$

$$\mathbb{E}[a^T X X^T a] - a^T \mu \mu^T a$$

$$= a^T \mathbb{E}[XX^T] a - a^T \mu \mu^T a$$

$$= a^T \Sigma a.$$

Let $A = \mathbb{R}^{k \times \ell}$ matrix (deterministic), $b \in \mathbb{R}^\ell$ deterministic vector.

Then $A^T X + b \in \mathbb{R}^\ell$ s.t.

$$\mathbb{E}[A^T X + b] = A^T \mu + b$$

$$\mathbb{V}[A^T X + b] = A^T \Sigma A$$

Multivariate Gaussian

$$X \sim \mathcal{N}_k(\mu, \Sigma)$$

$$\Rightarrow f(x) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

note

(1) $A^T X + b \sim N_\ell(A^T\mu + b,\ A^T \Sigma A)$

(2) $Z = \Sigma^{-1/2}(X-\mu) \sim N_k(0, I_k)$, $X = \Sigma^{1/2} Z + \mu$

Multivariate CLT

$X_1, X_2, \dots, X_n \in \mathbb{R}^k$ random , $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \Sigma$

Then $\sqrt{n}(\bar{X}_n - \mu) \sim N(0, \Sigma)$

Delta Method

$X_1, \dots, X_n \in \mathbb{R}^k$ random, $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \Sigma$.

$g : \mathbb{R}^k \to \mathbb{R}$

Then $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \rightsquigarrow N(0, \nabla g(\mu)^T \Sigma \nabla g(\mu))$

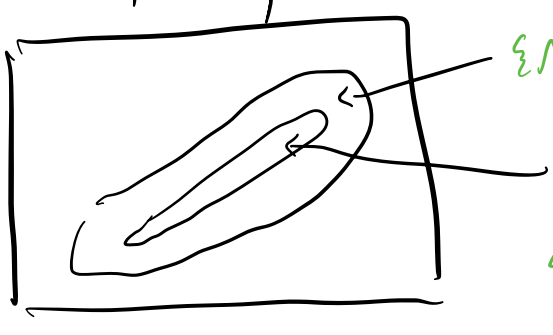
LECTURE 6 2/16/24 1PM

estimation

$X_i \overset{iid}{\sim} \mathbb{P} \to$ [statistical method] $\to \hat{\mathbb{P}} \simeq \mathbb{P} \to$ estimation / confidence intervals / test a hypothesis

statistical model = subset of all probability distributions

ex) all pdfs/pmfs



$\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$

$\{N(1, \sigma^2), \sigma^2 > 0\}$

$\{Ber(p) : p \in (0,1)\}$, $\{Exp(\lambda), \lambda \in (0, 15)\}$.

we can all specify models based on

pmf/pdf or cdf

$\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} , \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$ $\{F(x) : F \text{ is a continuous cdf}\}$

generally

$\{ \mathbb{P}_\theta, \theta \in \Theta \}$

$\mathbb{P}_\theta$: parameter(s); Θ: parameter space

ex) $\theta = p$

$\mathbb{P}_p = \text{Ber}(p)$

dimension of statistical model θ

$\text{Exp}(\lambda)$ ← Θ finite dimension → parametric model

all polynomials as pdfs ← Θ infinite dimension → nonparametric model

Notation for Estimation

$\mathbb{P}_\theta(X \geq 1)$, $\mathbb{E}_\theta[X]$, $\mathbb{V}_\theta[X]$

↑

$\mathbb{P}(X \geq 1)$ for $X \sim \mathbb{P}_\theta$

predicted by

$\theta \rightsquigarrow \hat{\theta}$

goal of estimation: provide single best guess $\hat{\theta}$ for θ based on data

An estimation $\hat{\theta}$ is a function of the data

$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ ← data given

① $\text{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$

"unbiased" if 0 ; asymptotically unbiased if $\to 0$ as $n \to \infty$

ex $\{N(\mu, 1), \mu \in \mathbb{R}\}$

$\hat{\mu}_1 = \bar{X}_n$ $\quad \text{bias}(\hat{\mu}_1) = 0$

$\hat{\mu}_2 = X_1$ $\quad \text{bias}(\hat{\mu}_2) = 0$

$\hat{\mu}_3 = 0$ $\quad \text{bias}(\hat{\mu}_3) = -\mu$

② standard error

$$\text{se}(\hat{\theta}) = \sqrt{\mathbb{V}_\theta[\hat{\theta}]} = \text{stder}(\hat{\theta})$$

$\text{se}(\hat{\mu}_1) = \frac{1}{\sqrt{n}}$

$\text{se}(\hat{\mu}_2) = 1$

$\text{se}(\hat{\mu}_3) = 0$

③ mean squared error

$$\text{MSE}(\hat{\theta}) = \text{bias}^2(\hat{\theta}) + \mathbb{V}_\theta[\hat{\theta}] = \text{bias}^2(\hat{\theta}) + \text{se}^2(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$$

$\text{MSE}(\hat{\mu}_1) = \frac{1}{n}$ $\quad \text{MSE}(\hat{\mu}_2) = 1$ $\quad \text{MSE}(\hat{\mu}_3) = \mu^2$

consistency

Def $\hat{\theta}_n$ is consistent estimator of θ if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$

← continuous mapping theorem often used

ex) $\hat{\mu}_1 = \bar{X}_n \xrightarrow{\mathbb{P}} \mu$ ✓ $\quad \hat{\mu}_2 = X_1 \xrightarrow{\mathbb{P}} \mu$ ✗ $\quad \hat{\mu}_3 = 0 \xrightarrow{\mathbb{P}} \mu$ ✗

Def

$\hat{\theta}_n$ asymptotically normal if $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2)$ for some $\sigma > 0$

σ^2 is the asymptotic variance

ex) $\bar{X}_n \Rightarrow \sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow N(0, \sigma^2)$ by CLT

$g(\bar{X}_n) \Rightarrow \sqrt{n}(g(\bar{X}_n) - g(\mu)) \rightsquigarrow N(0, g'(\mu)^2 \sigma^2)$ by Delta Method

$\to \hat{\theta}_n \approx N(\theta, \frac{\sigma^2}{n}) \approx \theta + \frac{\sigma}{\sqrt{n}} Z$ ← $N(0,1)$

$\Rightarrow \text{se}(\hat{\theta}_n) \approx \frac{\sigma}{\sqrt{n}}$

another definition

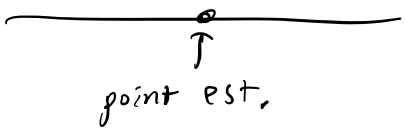
$\hat{\theta}_n$ is asymptotically normal if

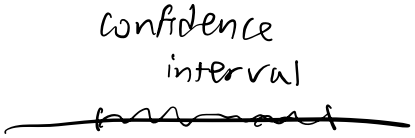
$$\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \rightsquigarrow N(0,1)$$

e.g. $\hat{\theta}_n = \bar{X}_n$ is by CLT

LECTURE 7 2/20/24 1PM

Confidence Intervals





Def A $(1-\alpha)$ confidence interval is a random interval

$C_n = (A_n, B_n)$ s.t. $\mathbb{P}_\theta(\theta \in (A_n, B_n)) = \mathbb{P}_\theta(A_n < \theta < B_n) \geq 1-\alpha \quad \forall \theta$

functions $A_n(X_1, \ldots, X_n)$ $B_n(X_1, \ldots, X_n)$

usually $\bar{X}_n - \frac{\sigma}{\sqrt{n}}$ $\bar{X}_n + \frac{\sigma}{\sqrt{n}}$

when computed, either $\theta \in (A_n, B_n)$ or NOT.

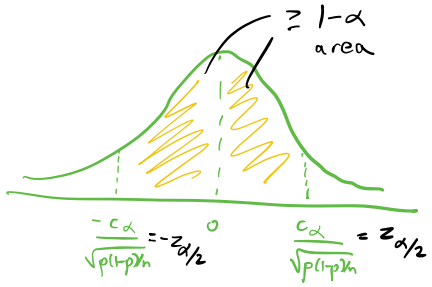
over many trials we get $\sim 1-\alpha$ captured.

constructing a CI

Kiss Example $X_1, \ldots, X_n \sim Ber(p)$

$$\hat{p} = \bar{X}_n \approx N\left(p, \frac{p(1-p)}{n}\right)$$

usually use estimator



want

$$1-\alpha = \mathbb{P}_p(\bar{X}_n - c_\alpha < p < \bar{X}_n + c_\alpha) = \mathbb{P}_p(p - c_\alpha \leq \bar{X}_n \leq p + c_\alpha)$$

$$= \mathbb{P}_p\left(-\frac{c_\alpha}{\sqrt{p(1-p)/n}} \leq \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \leq \frac{c_\alpha}{\sqrt{p(1-p)/n}}\right)$$

$$\left(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} z_{\alpha/2}, \bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} z_{\alpha/2}\right)$$

LECTURE 8 2/26/24 1PM

parameter of interest vs. nuisance parameter

might still need to be estimated

ex) $N(\mu, \sigma^2)$

- (μ, σ^2)
- μ but not σ^2 $\longrightarrow$ $\mu \in \bar{x}_n \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}$
- σ^2 but not μ
- $\frac{\mu}{\sigma}$ but not μ or σ alone

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ using $\hat{\mu}$

given $X_1, \ldots, X_n \sim \mathbb{P}_\theta$, can we even hope to estimate θ?

identifiability – θ identifiable if $\mathbb{P}_\theta = \mathbb{P}_{\theta'} \rightarrow \theta = \theta'$

i.e. $\theta \mapsto \mathbb{P}_\theta$ is injective

ex) $\{N(\mu, \sigma^2), \mu, \sigma \in \mathbb{R}\}$ is not because $\underset{-\sigma}{\sigma} \Rightarrow \sigma^2$ not injective

ex) dose-response

$$N\left(E_0 + \frac{D \times E_{max}}{D + ED_{50}}, \sigma^2\right)$$

E_0 = base effect (0 dose) E_{max} = max possible effect

D = dose of drug (known) ED_{50} = dose producing half effect of E_{max}

is not because varying E_{max}, ED_{50} can give same normal distribution

asymptotically normal / consistent due to LLN and Continuous Mapping Theorem + CLT / Delta

Plug-in Method

$\mathbb{E} \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n}$ substitution (only for expectations)

ex) $N(\mu, \sigma^2)$

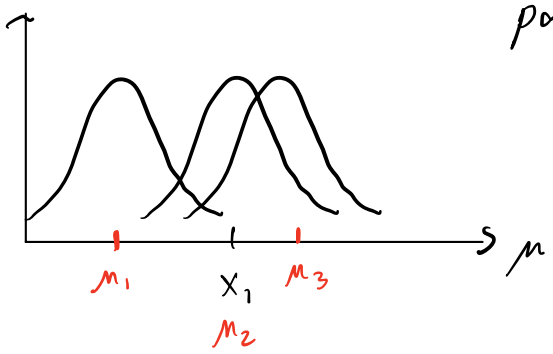
$\mu = \mathbb{E}[X] \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n} X_i$

$\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Method of Maximum Likelihood

the parameter value most likely to have generated the data

ex) $X_1 \sim N(\mu^*, 1)$



pdf $f_\mu(X_1)$ is highest when $\mu = X_1$

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\, f_\mu(X_1)$$

for $X_1, X_2, \dots, X_n$ iid $\sim f_{\theta^*}$

Def The likelihood is $L_n(\theta) = f_\theta(X_1) f_\theta(X_2) \cdots f_\theta(X_n)$

maximum likelihood estimator (MLE)

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_n(\theta)$$

$$\rightarrow \hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, f_\theta(X_1) f_\theta(X_2) \cdots f_\theta(X_n)$$

Def Log likelihood $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i)$

increasing, so $\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell_n(\theta)$ too.

also called sample log likelihood

LECTURE 9 2/28/24 (Did not attend)

MLE

① Consistent: $\hat{\theta}^{MLE} \xrightarrow{\mathbb{P}} \theta^*$

② Asymptotically normal: $\sqrt{n}(\hat{\theta}^{MLE} - \theta^*) \rightsquigarrow \mathcal{N}(0, \sigma^2_{MLE})$

③ Asymptotic efficiency: for any other asymptotically normal $\hat{\theta}$ with $\sqrt{n}(\hat{\theta} - \theta^*) \rightsquigarrow \mathcal{N}(0, \sigma^2)$, $\sigma^2 \geq \sigma^2_{MLE}$

if log likelihood is diff'ble

why is MLE good? notion of distance $dist(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$

① computable from samples (approximation)

② minimized only at θ^*

Kullback-Leibler (KL) divergence

not a distance!

$$D_{KL}(\mathbb{P}_{\theta^*} \| \mathbb{P}_\theta) = \int \log\left(\frac{f_{\theta^*}(x)}{f_\theta(x)}\right) f_{\theta^*}(x)\, dx$$

Jensen

≥ 0 always

$= 0$ if $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$ $\Rightarrow$ $\theta = \theta^*$ if identifiable

$$= \int \log(f_{\theta^*}(x))\, f_{\theta^*}(x)\, dx - \int \log(f_\theta(x))\, f_{\theta^*}(x)\, dx$$

- $\theta^* = \arg\min_\theta D_{KL}(\mathbb{P}_{\theta^*} \| \mathbb{P}_\theta)$ ← 0 only at $\theta = \theta^*$

$= \arg\max_\theta \int \log(f_\theta(x))\, f_{\theta^*}(x)\, dx$ ← since θ only appears here

$= \arg\max_\theta \mathbb{E}_{\theta^*}[\log f_\theta(X)], \quad X \sim \mathbb{P}_{\theta^*}$

$$\approx \operatorname*{argmax}_\theta \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i) = \hat{\theta}^{MLE}$$

sample log likelihood used to approximate

$\ell(\theta) = \mathbb{E}_{\theta^*}[\log f_\theta(X)]$ is the population log likelihood

$\theta^* = \operatorname*{argmax}_\theta \ell(\theta) \Rightarrow \ell'(\theta^*) = 0.$

note $\frac{1}{n} \ell_n(\theta) \xrightarrow{P} \ell(\theta)$ via LLN.

so the maximizer $\hat{\theta}^{MLE}$ of $\frac{1}{n}\ell_n(\theta)$ converges to the maximizer θ^* of $\ell(\theta)$

Asymptotic normality

Thm $\sqrt{n}(\hat{\theta}^{MLE} - \theta^*) \rightsquigarrow N(0, I(\theta^*)^{-1})$

$I(\theta^*)$ is Fisher Information

$$I(\theta^*) = \mathbb{E}_\theta\left[-\frac{d^2}{d\theta^2} \log f_\theta(X)\right] = \mathbb{V}_\theta\left[\frac{d}{d\theta} \log f_\theta(X)\right]$$

random variable X evaluated @ θ

Hessian for multidimensional → ∇_θ^2

gradient for multidimensional → ∇

ex $P_\theta = N(\theta, 1)$

$$f_\theta(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(X-\theta)^2}{2}}$$

$$\log f_\theta(X) = \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(X-\theta)^2}{2}$$

$$\frac{d}{d\theta} \log f_\theta(X) = X - \theta \rightarrow \mathbb{V}_\theta[X - \theta] = 1$$

$X \sim N(\theta, 1)$

PF Let $\hat{\theta} = \hat{\theta}^{MLE}$. WTS $\sqrt{n}(\hat{\theta} - \theta^*) \rightsquigarrow N_k(0, I(\theta^*)^{-1})$

1D: Taylor Expansion $0 = \ell_n'(\hat{\theta}) \approx \ell_n'(\theta^*) + \ell_n''(\theta^*)(\hat{\theta} - \theta^*)$

($0 = \ell_n'(\hat{\theta})$ by definition)

$$\sqrt{n}(\hat{\theta} - \theta^*) \approx -\sqrt{n}\frac{\ell_n'(\theta^*)}{\ell_n''(\theta^*)} = -\sqrt{n}\frac{\sum_{i=1}^n \frac{d}{d\theta}\log f_\theta(X_i)\big|_{\theta=\theta^*}}{\sum_{i=1}^n \frac{d^2}{d\theta^2}\log f_\theta(X_i)\big|_{\theta=\theta^*}}$$

$Y_i = \frac{d}{d\theta}\log f_\theta(X_i)\big|_{\theta=\theta^*}$

$W_i = \frac{d^2}{d\theta^2}\log f_\theta(X_i)\big|_{\theta=\theta^*}$

$$= -\sqrt{n}\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n W_i} = -\sqrt{n}\frac{\bar{Y}_n}{\bar{W}_n}$$

$$\mathbb{E}[Y_i] = \mathbb{E}_{\theta^*}\left[\frac{d}{d\theta}\log f_\theta(X_i)\big|_{\theta=\theta^*}\right] = \frac{d}{d\theta}\mathbb{E}_{\theta^*}\left[\log f_\theta(X_i)\big|_{\theta=\theta^*}\right] = \ell'(\theta^*) = 0$$

$$\mathbb{V}[Y_i] = \mathbb{V}_{\theta^*}\left[\frac{d}{d\theta}\log f_\theta(X_i)\big|_{\theta=\theta^*}\right] = I(\theta^*)$$

$$\bar{W}_n \xrightarrow{\mathbb{P}} \mathbb{E}[W_i] = \mathbb{E}_{\theta^*}\left[\frac{d^2}{d\theta^2}\log f_\theta(X_i)\big|_{\theta=\theta^*}\right] = -I(\theta^*) \quad \text{by LLN}$$

$$\sqrt{n}\bar{Y}_n = \sqrt{n}(\bar{Y}_n - 0) \rightsquigarrow N(0, I(\theta^*))$$

Slutsky's and CLT: $\sqrt{n}(\hat{\theta} - \theta^*) \approx -\frac{\sqrt{n}\bar{Y}_n}{\bar{W}_n} \rightsquigarrow N(0, I^{-1}(\theta^*))$

LECTURE 10 3/1/24 1PM

recall θ^* = true parameter

$\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i) = s\ell\ell$ (max at $\theta = \hat{\theta}^{MLE}$)

$\frac{1}{n}\ell_n \xrightarrow{\mathbb{P}} \ell$

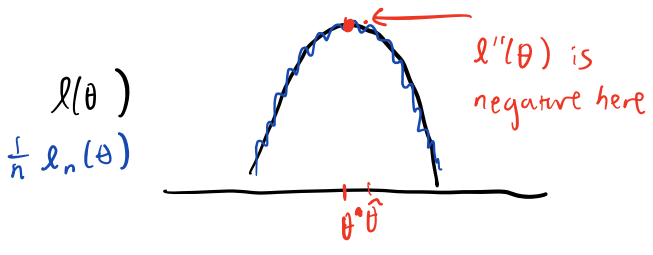
$\ell(\theta) = \mathbb{E}_{\theta^*}[f_\theta(X)] = p\ell\ell$ (max at $\theta = \theta^*$)

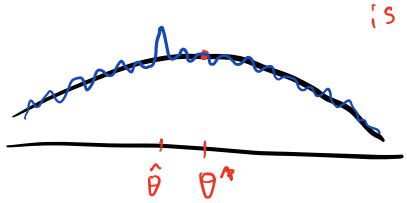
if model is regular.

$$\sqrt{n}(\hat{\theta} - \theta^*) \rightsquigarrow N(0, I(\theta^*)^{-1})$$

$$\hat{\theta} \approx N\left(\theta^*, \frac{1}{n I(\theta^*)}\right)$$

intuition for $I(\theta^*) = -\ell''(\theta^*)$





$\mathrm{Var}(\hat{\theta})$ is larger when ℓ is flatter $\leftrightarrow -\ell''(\theta^*)$ is smaller

building CI's. 95% CI would be $\theta^* \in \hat{\theta} \pm \frac{1.96}{\sqrt{nI(\hat{\theta})}}$

ex) Ber(p) $\rightarrow$ $\theta^* \in \hat{\theta} \pm \frac{1.96}{\sqrt{n}} \cdot \sqrt{\hat{\theta}(1-\hat{\theta})}$

- if MLE is output of algorithm, we have no closed form.

But we can still reason about it.

Algorithms to compute MLE

$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell_n(\theta)$ $\Rightarrow$ optimization problem

gradient ascent

$$\theta_{j+1} = \theta_j + \eta \nabla \ell_n(\theta_j)$$

η: unique parameter

$$\rightarrow \theta_{j+1} = \theta_j - \frac{f'(\theta_j)}{f''(\theta_j)}$$

↳ gradient ascent with $\eta = -\frac{1}{f''(\theta_j)}$

$$\theta_{j+1} = \theta_j - \nabla^2 \ell_n(\theta_j)^{-1} \nabla \ell_n(\theta_j)$$

↳ much more computationally expensive than gradient ascent.

Newton-Raphson

$$0 = f'(\theta) \approx f'(\theta_0) + f''(\theta_0)(\theta - \theta_0)$$

(θ: new, θ_0: old)

Expectation Maximization (EM) Algorithm

usually to find MLE for a mixture: $f(x) = (1-p)f_0(x) + p f_1(x)$ (diff distributions)

can write $Z = \mathrm{Ber}(p)$, $X_0 \sim f_0$, $X_1 \sim f_1$ $\quad Y = \begin{cases} X_0 & \text{if } Z=0 \\ X_1 & \text{if } Z=1 \end{cases}$

ex) $f_0 = N(\mu_0, \sigma_0^2)$ $f_1 = N(\mu_1, \sigma_1^2)$

5 unknown parameters $\theta = (p, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$

hopeless to find log of sum of exponentials!

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\left((1-p)\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} + (1-p)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\right)$$

EM! assume $p = \frac{1}{2}$, $\sigma_0^2 = \sigma_1^2 = 1$ $\rightarrow$ $f_0 = N(\mu_0, 1)$

$f_1 = N(\mu_1, 1)$

$$\ell_n^{full}(\theta) = \sum_{i=1}^n \log\left(f_0(X_i)^{1-\hat{z}_i} f_1(X_i)^{\hat{z}_i} \right)$$

$$= \sum_{i=1}^n \underbrace{(1-\hat{z}_i)\log(f_0(X_i))}_{\mu_0 \text{ param}} + \underbrace{\hat{z}_i \log(f_1(X_i))}_{\mu_1 \text{ param}}$$

LECTURE 11 3/4/24 1PM

$$\hat{\mu}_0 = \frac{\sum_{i; z_i=0} Y_i}{\#\{i; z_i=0\}} = \frac{\sum_{i=1}^n (1-z_i) Y_i}{\sum_{i=1}^n (1-z_i)} \qquad \hat{\mu}_1 = \frac{\sum_{i; z_i=1} Y_i}{\#\{i; z_i=1\}} = \frac{\sum_{i=1}^n z_i Y_i}{\sum_{i=1}^n z_i}$$

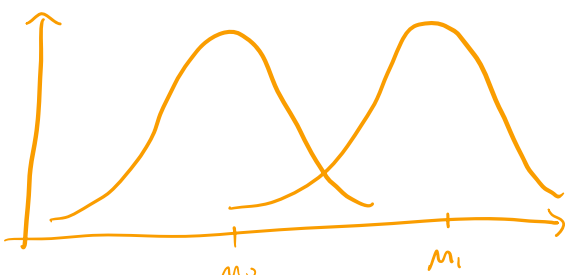
Main idea of EM: estimate z_i with $\hat{z}_i \in [0,1]$ which tells how strongly we believe $z_i = 0$ or 1.

then, maximize ℓ_n^{full} over μ_0, μ_1

$$\hookrightarrow \tilde{\ell}_n^{full}(\theta) = \sum_i (1-\hat{z}_i)\left(\log\frac{1}{\sqrt{2\pi}} - \frac{(Y_i-\mu_0)^2}{2} \right) + \hat{z}_i \left(\log\frac{1}{\sqrt{2\pi}} - \frac{(Y_i-\mu_1)^2}{2} \right)$$

$$= -\sum_i (1-\hat{z}_i)\tfrac{1}{2}(Y_i-\mu_0)^2 - \sum_i (\hat{z}_i)\tfrac{1}{2}(Y_i-\mu_1)^2 + \text{const.}$$

how to get $\hat{z}_i$



$$\hat{z}_i = \mathbb{E}[z_i \mid Y_i]$$

$$= \mathbb{P}[z_i = 1 \mid Y_i]$$

$$= \frac{f(Y_i \mid z_i=1)\,\mathbb{P}(z_i=1)}{f(Y_i \mid z_i=0)\,\mathbb{P}(z_i=0) + f(Y_i \mid z_i=1)\,\mathbb{P}(z_i=1)} = \frac{f(Y_i \mid z_i=1)}{f(Y_i \mid z_i=0) + f(Y_i \mid z_i=1)}$$

($\mathbb{P}(z_i=1) \leftarrow \frac{1}{2}$; $\mathbb{P}(z_i=0) \uparrow \frac{1}{2}$; $\mathbb{P}(z_i=1) \uparrow \frac{1}{2}$)

rename

$$= \frac{f_1(Y_i)}{f_0(Y_i) + f_1(Y_i)} = \frac{e^{-(Y_i-\mu_1)^2/2}}{e^{-(Y_i-\mu_0)^2/2} + e^{-(Y_i-\mu_1)^2/2}} \approx \frac{1}{0+1} = 1 \quad \text{when } Y_i \approx 1$$

$\hat{\mu}_0^{old}, \hat{\mu}_1^{old} \longrightarrow \hat{z}_1, \dots, \hat{z}_n \longrightarrow \hat{\mu}_0^{new}, \hat{\mu}_1^{new}$

$\hat{z} = \mathbb{E}[z_i | Y_i]$

E Step

$= \underset{\mu_0, \mu_1}{\text{argmax}}\ \hat{\ell}_n^{full}(\mu_0, \mu_1)$

M step

EM Algorithm

$p = \frac{1}{2}$, two normals with $\sigma_0^2 = \sigma_1^2 = 1$.

initialize $\hat{\mu}_0^{(0)}, \hat{\mu}_1^{(0)}$

for $j = 0, 1, 2, \dots$

E step. $\hat{z}_i^{(j+1)} = \mathbb{E}_{\hat{\mu}_0^{(j)}, \hat{\mu}_1^{(j)}}[z_i | Y_i] = \dfrac{e^{-\frac{1}{2}(Y_i - \hat{\mu}_1^{(j)})^2}}{e^{-\frac{1}{2}(Y_i - \hat{\mu}_1^{(j)})^2} + e^{-\frac{1}{2}(Y_i - \hat{\mu}_0^{(j)})^2}}$

M step. $\hat{\mu}_0^{(j+1)}, \hat{\mu}_1^{(j+1)} = \underset{\mu_0, \mu_1}{\text{argmax}}\ \hat{\ell}_n^{full}(\mu_0, \mu_1)$

↳ $\hat{\mu}_0^{(j+1)} = \dfrac{\sum_{i=1}^n (1 - \hat{z}_i^{(j+1)}) Y_i}{\sum_{i=1}^n (1 - \hat{z}_i^{(j+1)})}$ $\quad \hat{\mu}_1^{(j+1)} = \dfrac{\sum_{i=1}^n \hat{z}_i^{(j+1)} Y_i}{\sum_{i=1}^n \hat{z}_i^{(j+1)}}$

Generalized EM

unknown p, $f(y) = (1-p) f_{\theta_0}(y) + p\, g_{\theta_1}(y)$

need not be from same family

initialize $\hat{\theta}_0^{(0)}, \hat{\theta}_1^{(0)}, p^{(0)}$

for $j = 0, 1, 2, \dots$

E step. $\hat{z}_i^{(j+1)} = \dfrac{g_{\theta_1^{(j)}}(Y_i)\, p^{(j)}}{g_{\theta_1^{(j)}}(Y_i)\, p^{(j)} + f_{\theta_0^{(j)}}(Y_i)(1 - p^{(j)})}$

$p^{(j+1)} = \frac{1}{n} \sum_{i=1}^n \hat{z}_i^{(j+1)}$

M step.

$\theta_0^{(j+1)} = \underset{\theta_0}{\text{argmax}} \prod_{i=1}^n f_{\theta_0}(Y_i)^{1 - \hat{z}_i^{(j+1)}}$ $\quad$ if log 0, try these $\quad$ else

$\sum_{i=1}^n (1 - \hat{z}_i^{(j+1)}) \log f_{\theta_0} Y_i$

$\theta_1^{(j+1)} = \underset{\theta_1}{\text{argmax}} \prod_{i=1}^n g_{\theta_1}(Y_i)^{\hat{z}_i^{(j+1)}}$

$\sum_{i=1}^n z_i^{(j+1)} \log g_{\theta_1} Y_i$

Method of Moments

$X_1, X_2, \dots, X_n \sim \mathbb{P}_\theta$. Want to recover θ from estimated moments

Def $\alpha_k(\theta) = \mathbb{E}_\theta(X^k)$ is the k^{th} moment

(1) Work out the moments as a function of θ
→ find a function for $\alpha_k(\theta)$ for $k = 1, 2, \dots$

(2) Estimate moments via plug-in rule
→ $\alpha_k = \mathbb{E}[X^k] \rightsquigarrow \frac{1}{n}\sum_{i=1}^n X_i^k = \hat{\alpha}_k$

(3) Solve systems for θ

$$\rightarrow \begin{cases} \alpha_1(\theta) = \hat{\alpha}_1 \\ \vdots \\ \alpha_k(\theta) = \hat{\alpha}_k \end{cases}$$

ex) $\mathbb{P}_\theta(\mu, \sigma^2)$

(1) $\alpha_1(\theta) = \mu$ $\quad \alpha_2(\theta) = \mu^2 + \sigma^2$

(2) $\hat{\alpha}_1 = \bar{X}_n$ $\quad \hat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$

(3) solve $\begin{cases} \mu = \bar{X}_n \\ \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \end{cases}$

Method of Moments also gives

$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$ $\qquad \sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$

LECTURE 12 3/6/24 1PM

Bootstrap Brad Efron 1979

→ simulation-based method to approximate the variance of an estimator $\hat{\theta}(X_1, \dots, X_n)$ and ultimately build confidence intervals

so far $\theta \in \hat{\theta} \pm 1.96\, \hat{se}$ ← based on asymp. normality, $\hat{se} = \frac{\hat{\sigma}}{\sqrt{n}}$

$\hat{se}$ uses
- CLT (if $\hat{\theta} = \bar{X}_n$)
- Delta Method (if $\hat{\theta} = g(\bar{X}_n)$)
- Fisher info (if $\hat{\theta} = \hat{\theta}^{MLE}$)

but what if we lack asymp. normality??

Ex $\mathbb{P}_\lambda = Pois(\lambda)$

goal: estimate $\theta = Median(\mathbb{P}_\lambda)$

i.e. θ = integer k s.t. $\sum_{j=0}^{k-1} f_\lambda(j) < \frac{1}{2}$, $\sum_{j=0}^{k} f_\lambda(j) \geq \frac{1}{2}$

can try $\hat{\theta} = g(\bar{X}_n)$, but can't get $\hat{se}(\hat{\theta})$ because g not continuous

$\hat{\theta} = Med(X_1, \dots, X_n)$ isn't any one of the three forms above

> **we want:** general purpose method of estimating $se(\hat{\theta})$ without relying on special structure of $\hat{\theta}$, using only X_i's

Recall $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ say $Y = \hat{\theta}(X_1, \dots, X_n)$, $Y \sim Q$

we want to get more samples of $\hat{\theta}$, but we've used up all the X_i's for one $\hat{\theta}$.

Thought Experiment (too expensive...)

if we could get as many X_i as we wanted, we can make multiple (B) sets of $\hat{\theta}$ estimators.

$$X_{1:n}^{(1)} = \{X_1^{(1)}, \ldots, X_n^{(1)}\} \to \hat{\theta}^{(1)}$$
$$X_{1:n}^{(2)} = \{X_1^{(2)}, \ldots, X_n^{(2)}\} \to \hat{\theta}^{(2)}$$
$$\vdots$$
$$X_{1:n}^{(B)} = \{X_1^{(B)}, \ldots, X_n^{(B)}\} \to \hat{\theta}^{(B)}$$

histogram of the $\hat{\theta}^{(b)}$ for $b = 1, \ldots, B$ could be revealing.

we can also estimate $\widehat{\mathbb{V}[\hat{\theta}]} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{(c)} \right)^2$

Better Alternative

sample from $\hat{\mathbb{P}}_n = \text{Uniform}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

"empirical distribution"

$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le x)$ → cdf.

LLN says $\hat{F}_n(x) \xrightarrow{\mathbb{P}} F(x)$ as $n \to \infty$.

Bootstrap Sample

Given a sample $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$, a <u>bootstrap sample</u> is a collection of m random variables $X_1^*, \ldots, X_m^*$ s.t.

$$X_i^* \overset{iid}{\sim} \text{Unif}(X_1, \ldots, X_n), \quad i = 1, 2, \ldots, m$$

obtain

$$\{X_1^{*(1)}, \ldots, X_n^{*(1)}\} \to \hat{\theta}_1^*$$
$$\{X_1^{*(2)}, \ldots, X_n^{*(2)}\} \to \hat{\theta}_2^*$$
$$\vdots$$
$$\{X_1^{*(B)}, \ldots, X_n^{*(B)}\} \to \hat{\theta}_B^*$$

as before but now uniform sample.

we can use

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \frac{1}{B}\sum_{c=1}^{B}\hat{\theta}_c^*\right)^2 \approx \mathbb{V}_{\hat{\mathbb{P}}_n}[\hat{\theta}] \approx \mathbb{V}_{\mathbb{P}}[\hat{\theta}]$$

ok if B large

need large n, harder

LECTURE 13 3/8/24 1PM

Bootstrap samples: $X_1^*, \ldots, X_m^* \sim \text{Unif}(X_1, X_2, \ldots, X_n)$

$$\{X_1^{*(1)}, \ldots, X_n^{*(1)}\} \to \hat{\theta}_1^*$$
$$\{X_1^{*(2)}, \ldots, X_n^{*(2)}\} \to \hat{\theta}_2^*$$
$$\vdots$$
$$\{X_1^{*(B)}, \ldots, X_n^{*(B)}\} \to \hat{\theta}_B^*$$

If $\hat{\mathbb{P}} = \text{Unif}(X_1, \ldots, X_n)$ is close to $\mathbb{P}$ then $\hat{\theta}^*$ is close to $\hat{\theta}$ distribution

mean/variance of $\hat{\mathbb{P}}$

Suppose $\mathbb{E}[X_1] = \mu$, $\mathbb{V}[X_1] = \sigma^2$

$X_i^* \overset{iid}{\sim} \text{Unif}(X_1, \ldots, X_n)$ conditionally on $X_1, \ldots, X_n$, meaning

$\mathbb{P}(X_i^* = X_j \mid X_1, \ldots, X_n) = \frac{1}{n}$ for all j

unconditional distribution of X_i^* is $\mathbb{P}$ itself

conditional mean

$$\mathbb{E}[X_i^* \mid X_1, \dots, X_n] = \sum_{j=1}^{n} X_j \, \mathbb{P}(X_i^* = X_j \mid X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{j=1}^{n} X_j = \bar{X}_n$$

sample mean

unconditional mean

$$\mathbb{E}[X_i^*] = \mathbb{E}[\mathbb{E}[X_i^* \mid X_1, \dots, X_n]] = \mathbb{E}[\bar{X}_n] = \mu$$

tower property

true mean

conditional variance

$$\mathbb{V}[X_i^* \mid X_1, \dots, X_n] = \mathbb{E}[X_i^{*2} \mid X_1, \dots, X_n] - \mathbb{E}[X_i^* \mid X_1, \dots, X_n]^2$$

$$= \frac{1}{n} \sum_{j=1}^{n} X_j^2 - \left(\frac{1}{n} \sum_{j=1}^{n} X_j\right)^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2 = \hat{\sigma}^2$$

Sample variance

unconditional variance

$$\mathbb{V}[X_i^* \mid X_1, \dots, X_n] = \mathbb{E}[\mathbb{V}[X_i^* \mid X_1, \dots, X_n]] + \mathbb{V}[\mathbb{E}[X_i^* \mid X_1, \dots, X_n]]$$

law of total variance

$$= \mathbb{E}[\hat{\sigma}^2] + \mathbb{V}[\bar{X}_n] = \frac{n-1}{n} \sigma^2 + \frac{\sigma^2}{n} = \sigma^2$$

true variance

$$\mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2\right] = \sigma^2$$

Bootstrap CIs

(1) normal interval $\hat{\theta} \pm 1.96 \sqrt{v_{boot}}$ ← defeats purpose of bootstrap :(

(2) pivotal interval

→ let θ = ground truth (number)

→ let $\hat{\theta}$ = random variable (estimator of θ)

→ let $\hat{\theta}_0$ = single sample of $\hat{\theta}$ (number)

our CI will look like $\theta \in (\hat{\theta} - r_1, \hat{\theta} + r_2)$

use $\hat{\theta}_0$ to estimate $\hat{\theta}$ ⇒ $\theta \in (\hat{\theta}_0 - r_1, \hat{\theta}_0 + r_2)$

(1) uses $r_2 = r_1 = z_{\alpha/2}\, se(\hat\theta)$

we want

$$1-\alpha = \mathbb{P}\left(\hat\theta_0 - r_1 \le \theta \le \hat\theta + r_2\right) = \mathbb{P}\left(\underbrace{\theta - r_2 \le \hat\theta}_{\approx \alpha/2 \text{ mass}} \le \underbrace{\theta + r_1}_{\approx \alpha/2 \text{ mass}}\right)$$

recall $\mathbb{P}(\hat\theta \le q_{1-\alpha/2}) = \alpha/2$

$\mathbb{P}(\hat\theta \ge q_{\alpha/2}) = \alpha/2$

$q_x \to x$ of data is greater

take $r_2 = \theta - q_{1-\alpha/2} \rightsquigarrow \hat r_2 = \hat\theta_0 - q^*_{1-\alpha/2}$ ← quantiles of bootstrap samples

$r_1 = q_{\alpha/2} - \theta \rightsquigarrow \hat r_1 = q^*_{\alpha/2} - \hat\theta_0$ ←

↳ $\theta \in \left(\hat\theta_0 - (q^*_{\alpha/2} - \hat\theta_0),\ \hat\theta_0 + (\hat\theta_0 - q^*_{1-\alpha/2})\right)$

$= \left(2\hat\theta_0 - q^*_{\alpha/2},\ 2\hat\theta_0 - q^*_{1-\alpha/2}\right)$ → pivotal CI

(3) percentile interval

$$CI = \left(q^*_{\alpha/2},\ q^*_{1-\alpha/2}\right)$$

↑ ↑ quantiles of bootstrap samples

LECTURE 14 3/11/24 1PM

Hypothesis Testing lets us answer binary questions.

A test is a function Ψ : data $\to \{0,1\}$. In particular, Ψ is an estimator.

rejection region ⇒ $R = \{$ datasets for which $\Psi(\text{data}) = 1\}$

↳ $\Psi(\text{data}) = \mathbb{1}(\text{data} \in R)$

A test statistic is a function that summarizes the data and is sufficient to compute a test Ψ.

ex) $\bar X_n$, $\bar X_n^3$, $\bar X_n - 30$ for a test $\Psi(\text{data}) = \mathbb{1}(\bar X_n > 30)$

all sufficient for test, but $\bar X_n$ most natural

A hypothesis test takes the form

$$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$ = full parameter space.

↳ H_0 = innocent OR status quo

↳ H_1 = guilty OR discovery

	conclude H_0 ($\Psi = 0$)	conclude H_1 ($\Psi = 1$)
H_0 true	✓	Type I Error
H_1 true	Type II Error	✓

false positive

false negative

The size of a test Ψ is

$$\text{size}(\Psi) = \max_{\theta \in \Theta_0} P_\theta(\Psi = 1)$$

← θ should be on boundary between Θ_0, Θ_1

→ max probability of type I error.

The test has level $\alpha \in (0,1)$ if $\text{size}(\Psi) \leq \alpha$.

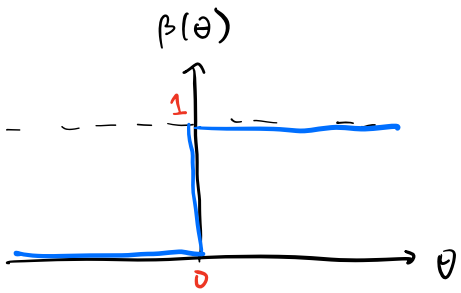
→ typically use $\alpha = 0.01$ or $\alpha = 0.05$.

we can get $\text{size}(\Psi) = 0$ if $\Psi = 0$ ∀ datasets but this is not helpful.

The power function is defined as

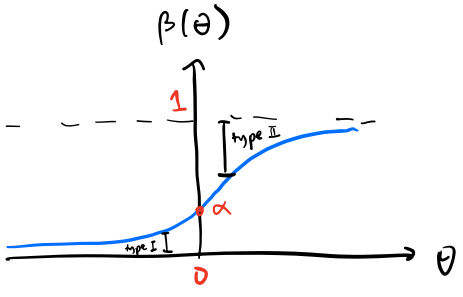
$$\beta(\theta) = P_\theta(\Psi = 1)$$

defined for specific parameter set θ.



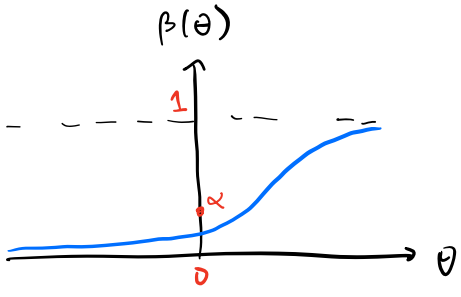
ideal

(but not easy with finite data)



size α = level α

good



size $\leq$ level α

bad

LECTURE 15 3/13/24 1PM

Today: construct a test Ψ. compute power function

ex) $X_1, \dots, X_n$ iid, $\mathbb{E}[X_i] = \mu$

$H_0 : \mu \le 30 \quad H_1 : \mu > 30.$

find c_α

$$\Psi = \mathbb{1}\{\bar{X}_n - 30 > c_\alpha\}$$

$$\alpha = \text{size}(\Psi) = \max_{\mu \le 30} \mathbb{P}_\mu(\bar{X}_n - 30 > c_\alpha)$$

but $\sqrt{n}(\bar{X}_n - 30) \rightsquigarrow N(0, \sigma^2)$ by CLT so $\bar{X}_n - 30 \approx N(0, \frac{\sigma^2}{n})$

$$\Rightarrow \alpha = \mathbb{P}\left(\frac{\sigma}{\sqrt{n}} Z \ge c_\alpha\right) \quad \text{so} \quad c_\alpha = \frac{\sigma}{\sqrt{n}} z_\alpha \approx \boxed{\frac{\hat{\sigma}}{\sqrt{n}} z_\alpha}$$

↑ max at boundary

$$\alpha = \mathbb{P}\left(Z \ge \frac{\sqrt{n}}{\sigma} c_\alpha = z_\alpha\right)$$

So our test is $\Psi = \mathbb{1}\{\bar{X}_n - 30 > \frac{\hat{\sigma}}{\sqrt{n}} z_\alpha\}$.

for $\alpha = 0.05$, $z_\alpha = 1.65$, $z_{\alpha/2} = 1.96$

→ if $n = 164$, $\hat{\sigma} = 12$, $\bar{X}_n = 33.4$

then $\bar{X}_n - 30 = 3.4 > \frac{12}{\sqrt{164}} \cdot 1.65 = 1.54$ so reject

Power Function

$\bar{X}_n \sim N(\mu, \frac{\hat{\sigma}^2}{n})$

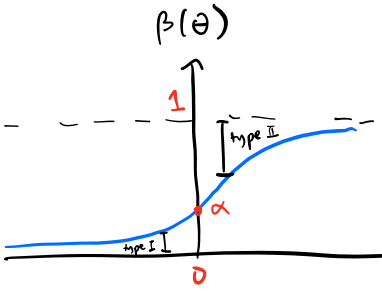
$$\beta(\mu) = \mathbb{P}_\mu\left(\bar{X}_n - 30 > \frac{\hat{\sigma}}{\sqrt{n}} z_\alpha\right)$$

$$\approx \mathbb{P}_\mu\left(\mu + \frac{\hat{\sigma}}{\sqrt{n}} Z - 30 > \frac{\hat{\sigma}}{\sqrt{n}} z_\alpha\right) = \mathbb{P}\left(Z > z_\alpha + \frac{\sqrt{n}}{\hat{\sigma}} 30 - \frac{\sqrt{n}}{\hat{\sigma}} \mu\right)$$

higher μ ⇒ higher $\beta(\mu)$ →

$$= \boxed{1 - \Phi\left(z_\alpha + \frac{\sqrt{n}}{\hat{\sigma}}(30 - \mu)\right)}$$

Recall our power function from before.



We want to maximize slope at 0 to get a sharp transition

in our example,

$$\beta'(\mu) \approx \frac{\sqrt{n}}{\hat{\sigma}} \phi\left(z_\alpha + \frac{\sqrt{n}}{\hat{\sigma}}(30-\mu)\right)$$

at $\mu = 30$ (our boundary),

$$\beta'(\mu) = \frac{\sqrt{n}}{\hat{\sigma}} \phi(z_\alpha)$$ is maximized when

→ n up ← more data

→ $\hat{\sigma}$ down ← less variance

Wald's Test

Suppose $\frac{\hat{\theta}-\theta}{\hat{se}} \rightsquigarrow N(0,1)$ as $n \to \infty$

Case 1. $\theta = \theta_0$ or $\theta \neq \theta_0$

$$\psi = \mathbb{1}\left(\left|\frac{\hat{\theta}-\theta}{\hat{se}}\right| > z_{\alpha/2}\right)$$

Case 2. $\theta \leq \theta_0$ or $\theta > \theta_0$

$$\psi = \mathbb{1}\left(\frac{\hat{\theta}-\theta}{\hat{se}} > z_\alpha\right)$$

Case 3. $\theta \geq \theta_0$ or $\theta < \theta_0$

$$\psi = \mathbb{1}\left(\frac{\hat{\theta}-\theta}{\hat{se}} < -z_\alpha\right)$$

ex $\sqrt{n}\left(\hat{\theta}^{MLE} - \theta\right) \rightsquigarrow N(0, I^{-1}(\theta))$

so $se(\hat{\theta}^{MLE}) \approx \frac{1}{\sqrt{nI(\theta)}}$ or can use $\frac{1}{\sqrt{nI^{-1}(\theta_0)}}$

$\theta = \theta_0$ or $\theta \neq \theta_0$

$$\psi = \mathbb{1}\left(\sqrt{nI(\hat{\theta}^{MLE})}\left|\hat{\theta}^{MLE} - 0\right| > z_{\alpha/2}\right)$$

↑ or $I(\theta_0)$

LECTURE 16 3/16/24 (Did not attend)

p-values answer "How close was the call between H_0 and H_1?"

small p-value gives evidence against H_0.

evidence against H_0 =
- very strong: $p < 0.01$
- strong: $0.01 < p < 0.05$
- weak: $0.05 < p < 0.1$
- little or no: $0.1 < p$

The p-value is

→ smallest level α for which we can reject H_0

→ if T_n is test statistic and T_n^{obs} is observed value of T_n, rejection regions are

$$R = \{T_n > c_\alpha\} \Rightarrow p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T_n > T_n^{obs})$$

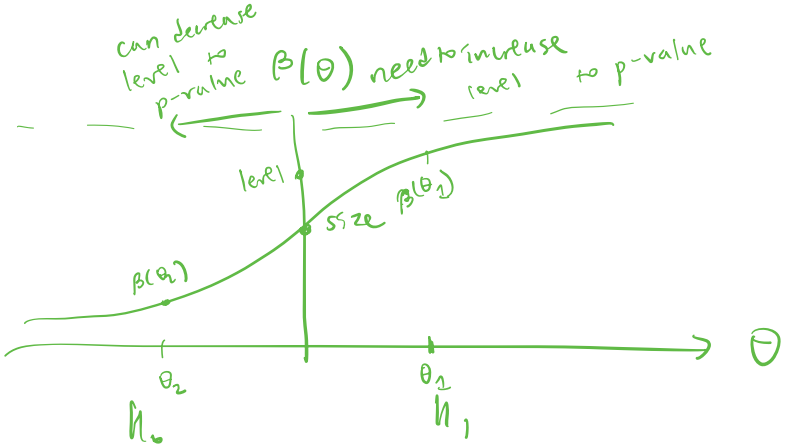
$$R = \{T_n < c_\alpha\} \Rightarrow p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T_n < T_n^{obs})$$

$$R = \{|T_n| > c_\alpha\} \Rightarrow p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(|T_n| > T_n^{obs})$$

if T_n is standardized as $N(0,1)$ then (for first kind)

$$p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(N(0,1) > T_n^{obs})$$

"largest probability of observing a more extreme value than T_n^{obs} over all $\theta \in \Theta_0$"



example $H_0: p = \frac{1}{2}$ $H_1: p \neq \frac{1}{2}$ $n = 124$ $\bar{X}_n = 0.645$.

Ber(p).

$$\frac{\bar{X}_n - p}{\sigma} \sim N(0,1) \qquad \bar{X}_n \sim N\left(p, \frac{\sigma^2}{n}\right)$$

$$\to N\left(p, \frac{p(1-p)}{n}\right)$$

$$\overset{\text{under } \bar{X}_n \to p}{\approx} N\left(\frac{1}{2}, \frac{\bar{X}_n(1-\bar{X}_n)}{n}\right) \text{ (a)}$$

$$\overset{\text{under } \theta \in \Theta_0}{\approx} N\left(\frac{1}{2}, \frac{1}{4n}\right) \text{ (b)}$$

A test of level α is

(a) $$\psi = \mathbb{1}\left(\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}\,|\bar{X}_n - 0.5| > z_{\alpha/2}\right)$$

$$= \mathbb{1}\left(3.37 > z_{\alpha/2}\right)$$

can have $\alpha/2 \geq 0.0004$

$\alpha \geq \underline{0.0008}$ ← p-value.

(b) $$\psi = \mathbb{1}\left(\sqrt{4n}\,|\bar{X}_n - 0.5| > z_{\alpha/2}\right)$$

$$= \mathbb{1}\left(3.23 > z_{\alpha/2}\right)$$

can have $\alpha/2 \geq 0.0006$

$\to$ p-value $= \underline{0.0012}$

} either works!

example $H_0: \mu \leq 30$ $H_1: \mu > 30$ $n = 164$ $\bar{X}_n = 33.4$ $\hat{\sigma} = 12$

$$\bar{X}_n \rightsquigarrow N\left(30, \frac{12}{n}\right) \quad \text{under } H_0$$

$$\psi = \mathbb{1}\left(\frac{\sqrt{n}}{12}(\bar{X}_n - 30) > z_\alpha\right)$$

$$= \mathbb{1}(3.63 > z_\alpha) \to \text{p-value} = \underline{0.0002}$$

LECTURE 17 3/18/24 1PM

$X \sim \chi^2_k$ is a chi-squared distribution with k deg. of freedom

$$X := Z_1^2 + Z_2^2 + \cdots + Z_k^2, \quad Z_i \sim \mathcal{N}(0,1).$$

$$\mathbb{E}[X] = \mathbb{E}[Z_1^2 + Z_2^2 + \cdots + Z_k^2] = k$$

$$\mathbb{V}[X] = \sum_{i=1}^{k} \mathbb{V}[Z_i^2] = \sum_{i=1}^{k} \mathbb{E}[Z_i^4] - \mathbb{E}[Z_i^2]^2 = \sum_{i=1}^{k} 2 = 2k$$

Let $\chi^2_{k,\alpha}$ be the α^{th} quantile of χ^2_k.

↳ $\mathbb{P}(X > \chi^2_{k,\alpha}) = \alpha$, $X \sim \chi^2_k$

Suppose we have a pmf f_0 and want to test if a discrete r.v. has pmf f_0.

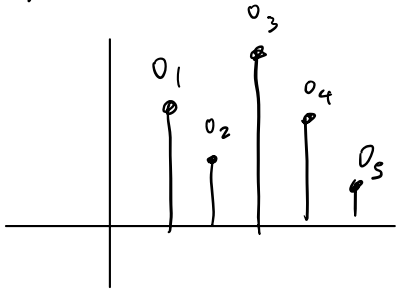
χ^2 Goodness-of-Fit Test

X discrete, takes k values $\{1, 2, \ldots, k\}$

$H_0 : f = f_0$ $\quad$ $H_1 : f \neq f_0$

> Also read on χ^2 independence test!

for $X_1, X_2, \ldots, X_n \sim f$



$$O_j = \#\{i : X_i = j\} = \sum_{i=1}^{n} \mathbb{1}(X_i = j)$$

expect $O_j \approx n f_0(j)$

↳ under null, $n f_0(j) = E_j$

O_j = observed # of j's

E_j = expected # of j's under null

$$T = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

Reject if $T > c_\alpha = \chi^2_{k-1,\alpha}$

p-value $= \mathbb{P}_{H_0}(T > T_{obs}) = \mathbb{P}(\chi^2_{k-1} > T_{obs})$

Theorem $T \rightsquigarrow \chi^2_{k-1}$ as $n \to \infty$

↳ $k-1$ degrees of freedom since $O_1 + O_2 + \cdots + O_k = n$.

Ex 100 surveys

	exp.	obs.
cliff	50	48
happy	30	40
sad	20	12

$$T = \frac{2^2}{50} + \frac{10^2}{30} + \frac{8^2}{20} = 6.61$$

$$\chi^2_{k-1,\alpha} = \chi^2_{2,0.05} = 5.991 \rightarrow \text{reject}$$

Ex Bernoulli, $H_0: p = \frac{1}{2}$, $H_1: p \neq \frac{1}{2}$ → a pmf between $\{1,2\}$.

$n = 124$

	exp.	obs.
right	62	80
left	62	44

$$T = \frac{18^2 + 18^2}{62} = 10.46$$

$$\text{p-value} = \mathbb{P}(\chi^2_1 = Z^2 > 10.46)$$

$$= \mathbb{P}(|Z| > 3.23)$$

same as Wald's Test?

LECTURE 18 4/1/24 (Did not attend)

Nonparametric Tests

$X \sim N(0,1)$ vs. $X \sim$ any other distribution

$X \sim$ Normal vs. $X \sim$ not normal

$X \sim F_0$ vs. $X \neq F_0$

} much more than just one "family" of distributions

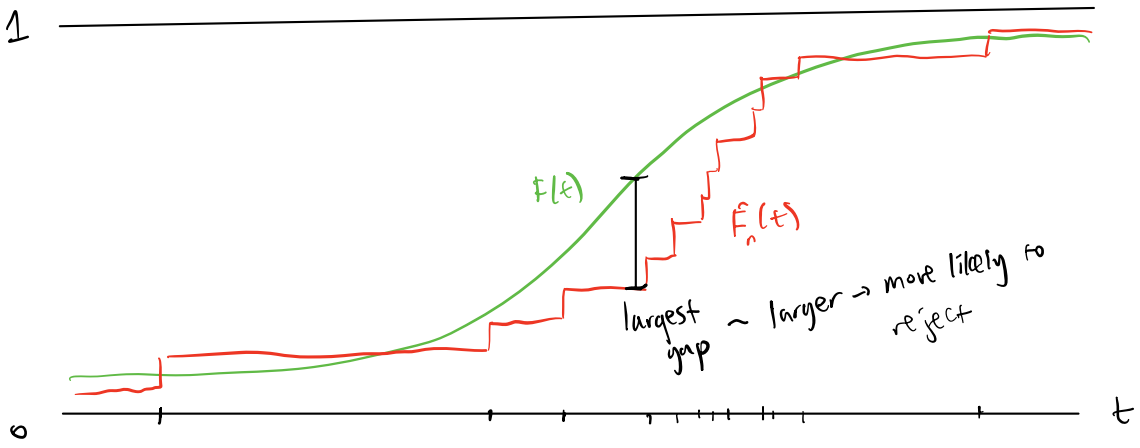
Komolgorov-Smirnov

given $X_1, \ldots, X_n \overset{iid}{\sim} F$, test $H_0: F = F_0$ vs. $H_1: F \neq F_0$

empirical cdf $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t)$

- $\mathbb{E}[\hat{F}_n(t)] = F(t)$ → unbiased estimator
- $n\hat{F}_n(t) \sim \text{Bin}(n, F(t))$ → average of Bernoulli r.v.s

↳ $\sqrt{n}(\hat{F}_n(t) - F(t)) \rightsquigarrow N(0, F(t)(1-F(t)))$



Komolgorov - Smirnov Test

$$\psi = \mathbb{1}\left(\underbrace{\sup_t |\hat{F}_n(t) - F(t)|}_{T} > c_\alpha \right)$$

→ c_α doesn't depend on F_0, only on n !!

Komolgorov - Lillie fors

for H_0: F is Gaussian vs. H_1: F is not Gaussian,

using $T = \sup_t |\hat{F}_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)|$

requires us to use KL distribution, NOT KS.

KL: ↓ H_0 = any Gaussian

KS: ↓ H_0 = specific Gaussian

Note $c_\alpha^{KL} < c_\alpha^{KS}$

Permutation Test "2-sample test"

Given $X_1, \dots, X_n \sim F_X$ $Y_1, \dots, Y_m \sim F_Y$

Test $H_0: F_X = F_Y$ vs. $H_1: F_X \neq F_Y$

can't use $T = \sup_t |\hat{F}_X(t) - \hat{F}_Y(t)|$ and $\psi = \mathbb{1}(T > c_\alpha)$

since F_0 and c_α are unknown

instead use $T = |\bar{X}_n - \bar{Y}_n|$, $Z_i = X_i$, $Z_{i+n} = Y_i$ for $1 \le i \le n$

let m=n for simplicity

compute $T^1, T^2, \dots, T^B$ by picking random permutation σ_j of $\{1, 2, \dots, 2n\}$, then

$$Z^{\sigma_j}_{left} = \frac{1}{n}\sum_{i=1}^{n} Z_{\sigma_j(i)} \qquad Z^{\sigma_j}_{right} = \frac{1}{n}\sum_{i=n+1}^{2n} Z_{\sigma_j(i)} \qquad T^j = |Z^{\sigma_j}_{left} - Z^{\sigma_j}_{right}|$$

reject if $T > t_\alpha$ where $\frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(T^i > t_\alpha) = \alpha$

↑ "only α of the permutations gave higher differences"

$$\text{p-value} = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(T^i > T)$$

note that type II error can be large, especially if $\mathbb{E}[X] = \mathbb{E}[Y]$

LECTURE 19 4/3/24 (Did not attend)

Multiple Hypothesis Testing

↳ you will inevitably falsely reject some tests if you perform many.

① Bonferroni Correction → very conservative

m tests, reject each at level α/m

↳ $\mathbb{P}_{H_0}(\exists i, H_{0,i} \text{ is rejected}) = \mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m}\{p_i \le \alpha/m\}\right)$

$$\le \sum_{i=1}^{m} \mathbb{P}_{H_0}\{p_i \le \alpha/m\} = m \cdot \alpha/m = \alpha.$$

② Benjamini-Hochberg (BH) Method

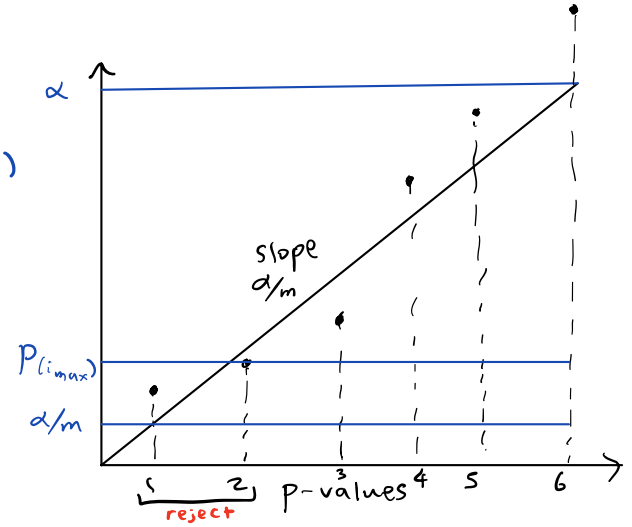
$$FDP = \frac{\#\text{ false positives}}{\#\text{ positives}} \qquad FDR = \mathbb{E}_{H_0}[FDP]$$

instead of ∃ false positive ↑ can approximate/bound

a) order your m p-values $p_{(1)} \le \cdots \le p_{(m)}$

b) Let $i_{max} = \max_{i \in [m]} : p_{(i)} \le \frac{\alpha}{m} i$

c) Reject all tests s.t. $p_{(i)} \le p_{(i_{max})}$



Summary

		# rejected on picture
usually reject for	$\le \alpha$	(5)
② BH	$\le p_{(i_{max})}$	(2)
① Bonferroni	$\le \alpha/m$	(0)

T-test

Wald Test had asymptotically normal data.

T-test $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ so $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is exactly $N(0,1)$

↳ but NOT $\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}$, especially for small n.

$$\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma} \sim \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Student t-distribution with $\nu = n-1$ degrees of freedom

Student's t-test

Given $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ (μ, σ^2 unknown), $\psi = \mathbb{1}\left(\sqrt{n}\frac{|\bar{X}_n - \mu_0|}{\hat{\sigma}} \geq t_{n-1,\alpha/2}\right)$

$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0$ p-value $= \mathbb{P}\left(|t_{n-1}| \geq \sqrt{n}\frac{|\bar{X}_n - \mu_0|}{\hat{\sigma}}\right)$

LECTURE 20 4/5/24 IPM

Bayesian vs. Frequentist

Bayesian: alternative way to produce estimators, build CIs, do hypothesis testing, based on "degree of belief"

Frequentist: repeat experiment many times → based on frequency of things happening

Bayesian

Use weight likelihood $L_n(\theta)$ via prior pdf $f(\theta)$

"belief before seeing data $X_1, \dots, X_n$"

Bayesian Method

→ start with prior belief about θ^*

→ observe data $X_1, \dots, X_n$

→ update prior belief into a posterior belief

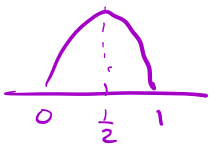
Prior Belief (probability distribution on unknown θ^*)

ex) assume $X_1, \dots, X_n \overset{iid}{\sim} Ber(p^*)$

↳ convenient prior is $Beta(a,b) = \frac{1}{k}\,p^{a-1}(1-p)^{b-1}$ (k: proportional constant)

for $a \ll b$

to determine if $p^a = \frac{1}{2}$ we use $\text{Beta}(2,2) = 6p(1-p)$



Prior to Posterior

The posterior is the probability of θ given the data

Recall $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$

↳ $\mathbb{P}(\theta | \text{data}) = \frac{\mathbb{P}(\text{data}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\text{data})}$ ($\mathbb{P}(\text{data})$ ← proportionalizing constant; $\mathbb{P}(\text{data}|\theta)$ ← likelihood; $\mathbb{P}(\theta)$ ← prior)

↳ $f(\theta | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|\theta) f(\theta)}{f(x_1, \dots, x_n)} = \frac{f(x_1, \dots, x_n|\theta) f(\theta)}{c_n}$

$f(x_1, \dots, x_n)$: no dependence on θ

c_n: to make whole thing integrate to 1

just write $f(\theta | x_1, \dots, x_n) \propto L_n(\theta) f(\theta)$

posterior ∝ likelihood · prior

back to example,

$L_n(p) = p^{\sum x_i}(1-p)^{n - \sum x_i}$ $\quad f(p) = 6p(1-p)$

$\Rightarrow f(p | x_1, \dots, x_n) \propto p^{\sum x_i + 1}(1-p)^{n - \sum x_i + 1}$

↳ $\text{Beta}(\sum x_i + 2, n - \sum x_i + 2)$

if prior/posterior are from same family of distributions, we have a conjugate prior (e.g. Beta from above)

Bayes estimator := $\mu_{\text{posterior}} = \mathbb{E}[f(\theta | x_1, \dots, x_n)]$

Maximum A Posteriori = MAP := $\text{mode}_{\text{posterior}} = \text{argmax}_\theta L_n(\theta) f(\theta)$

LECTURE 21 4/8/24 (Did not attend)

Markov Chain Monte Carlo to estimate $\hat{\theta}^{Bayes}_{posterior} = \int \theta f(\theta | X_1, \ldots, X_n) d\theta$

① draw $\theta_1, \ldots, \theta_T$ iid from $f(\theta | X_1, \ldots, X_n)$ by simulating a Markov Chain

② compute $\hat{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_t$ (Monte Carlo)

MAP = mode$_{posterior}$ $\rightarrow$ $\hat{\theta}^{MAP}$.

Find usually using $\arg\max_{\theta} (\log [f(\theta | X_1, \ldots, X_n)])$

$\log (f(\theta | X_1, \ldots, X_n)) = \log (c_n \mathcal{L}_n(\theta) f(\theta))$ ← normalizing

$= \log c_n + \ell_n(\theta) + \log f(\theta)$

set gradient $0 = \nabla \ell_n(\theta) + \nabla \log f(\theta)$, solve for θ.

↳ gradient ascent if needed: $\theta^{(k+1)} = \theta^{(k)} + \eta_k (\nabla \ell_n(\theta^{(k)}) + \nabla \log f(\theta^{(k)}))$

Def $(1-\alpha)$ Posterior interval

ANY $[a, b]$ s.t. $\int_a^b f(\theta | X_1, \ldots, X_n) d\theta = 1 - \alpha$.

Choosing Priors

① True prior knowledge (e.g. ≥ 0)

② convenient calculation (e.g. Beta conjugate for Ber(p), normal conjugate for normal)

③ without prior knowledge, uniform prior

↓

$\hat{\theta}^{MAP} = \hat{\theta}^{MLE}$ since $f(\theta)$ uniform

↓

if you want uniform over $\mathbb{R}$, can take $f(\theta) = 1$

even if $\int_{-\infty}^{\infty} 1 \, d\theta$ diverges $\rightsquigarrow$ improper prior

$\hat{\theta}^{MAP}$ might not be equal to $\hat{\theta}^{Bayes}$

ex) prior $f(p) = \text{Unif}([0,1])$

given $X_1, \dots, X_n$

$$f(p \mid X_1, \dots, X_n) \propto p^{\sum X_n} (1-p)^{n - \sum X_n} \propto \text{Beta}(\overbrace{\sum X_n + 1}^{a}, \overbrace{n - \sum X_n + 1}^{b})$$

biased $\hat{\theta}^{MAP} = \frac{a}{a+b} = \frac{\sum X_n + 1}{n+2} \neq \hat{\theta}^{MLE} = \frac{\sum X_n}{n}$

Normal prior is a conjugate prior for normal data

if $\theta \sim N(0, \sigma^2)$ $\quad f(\theta) \propto e^{-\frac{x^2}{2\sigma^2}}$

given data $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$,

we have $f(\theta \mid X_1, \dots, X_n) \propto e^{-\frac{\theta^2}{2\sigma^2}} e^{-\sum_i \frac{(X_i - \theta)^2}{2}}$

$$= \exp\left(-\frac{1}{2}\left[\sum_i X_i^2 - 2\sum_i X_i \theta + n\theta^2 + \frac{\theta^2}{\sigma^2}\right]\right)$$

$e^{-\frac{\sum X_i^2}{2}}$ is constant ↘

$$\propto \exp\left(-\frac{1}{2}\left[\left(n + \frac{1}{\sigma^2}\right)\theta^2 - 2\sum X_i \theta\right]\right)$$

$\frac{1}{2\left(\frac{1}{n+1/\sigma^2}\right)}\left(\frac{\sum X_i}{n+1/\sigma^2}\right)^2$ is constant ⟶

$$\propto \exp\left(-\frac{1}{2\left(\frac{1}{n+1/\sigma^2}\right)}\left[\theta - \frac{\sum X_i}{n+1/\sigma^2}\right]^2\right)$$

$$\text{posterior} \sim N\left(\frac{\sum X_i}{n+1/\sigma^2}, \frac{1}{n+1/\sigma^2}\right)$$

$$\hat{\theta}^{MAP} = \hat{\theta}^{Bayes} = \frac{\sum X_i}{n+1/\sigma^2}$$

LECTURE 22 4/10/24 1PM

Robust statistics -

How to summarize data in presence of outliers?

Def A statistic/estimator is robust if it does not change the data even when you modify some of the data a lot

Def Let m = max amount of observations we can make arbitrarily large or small while leaving estimator bounded.

Then $\frac{m}{n}$ is the breakdown point

ex)

$\hat{\theta}(X_1, \dots, X_5)$ = sample median

→ $|\hat{\theta}(X_1, X_2', X_3, X_4', X_5) - \hat{\theta}(X_1, X_2, X_3, X_4, X_5)| \leq B < \infty$

→ $|\hat{\theta}(X_1, X_2', X_3', X_4', X_5) - \hat{\theta}(X_1, X_2, X_3, X_4, X_5)| \longrightarrow \infty$

} breakdown point $3/5$.

breakdown point for median for odd n is $\frac{(n+1)/2}{n} = \frac{n+1}{2n} \longrightarrow \frac{1}{2}$ as $n \to \infty$,

note something like $\hat{\theta}(X_1, \dots, X_n) = 4$ has breakdown point 1 but isn't really useful. Best we can do is $1/2$.

Median as MLE

MLE is median in Laplace distribution $f_{\mu,b}(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$

↳ simplify to $b=1$ → $f_{\mu,1}(x) = \frac{1}{2} e^{-|x-\mu|}$

then $\ell_n(\theta) = \sum_{i=1}^{n} \ln\left(\frac{1}{2}\right) - |X_i - \mu| \Rightarrow - \sum_{i=1}^{n} |X_i - \mu|$

$\hat{\mu}^{MLE} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} |X_i - \mu| \rightarrow \text{median}(X_1, \dots, X_n)$

Huber's Contamination Model

For data "contaminated" with outliers.

$$Z \sim \text{Ber}(\epsilon) \qquad X_{true} \sim P_{\theta^*} \qquad X_{out} \sim Q \qquad \epsilon \text{ small}$$

Let $Y = (1-Z) X_{true} + Z X_{out}$

ex) $P_{\theta^*} = N(\theta^*, 1)$ and Q has pdf q.

$$f(y) = (1-\epsilon) \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\theta^*)^2}{2}} + \epsilon q(y)$$

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\left((1-\epsilon) \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i-\theta^*)^2}{2}} + \epsilon q(Y_i) \right)$$

↳ but we need to know q!

to fix this, we can maximize over all $q \in$ family $\mathcal{Q}$.

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \max_{q \in \mathcal{Q}} \ell_n(\theta)$$

quasi MLE

(uses iid assumption for both P_{θ^*}, Q)

Arbitrary Contamination

if we only know there are m outliers

↳ set $C = \{i \text{ s.t. } X_i \text{ is outlier}\}$, $|C| = m$.

let $\ell_n(\theta, C) = -\sum_{i \in C^c} (X_i - \theta)^2$ ← only define on non-outliers

T normal dist.

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \max_{|C|=m} \ell_n(\theta, C)$$

"trimmed mean"

- takes mean of middle $n-m$ values

LECTURE 23 4/12/24 (Did not attend)

Regression - use features $X \in \mathbb{R}^k$ to predict response variable $Y \in \mathbb{R}$

goal: predict Y given X / understand how Y changes with X

challenge: for EACH fixed $X=x$, there is a whole distribution $Y|X=x$
↳ might not have exact $f(x)$ where $X=x \Rightarrow Y=f(x)$

best prediction property: find f and try to minimize $\mathbb{E}[(Y-f(X))^2]$

$$\mathbb{E}[(Y-f(X))^2] = \mathbb{E}[\mathbb{E}[(Y-f(X))^2 \mid X]]$$

↳ $\min_{f(x)} \mathbb{E}[(Y-f(x))^2 \mid X=x]$ for each x

$f(x)=a$ where $h(a)=\mathbb{E}[(Y-a)^2 \mid X=x]$ is minimized. (error)

make $h'(a)=0 \rightarrow \mathbb{E}[-2Y+2a \mid X=x] \Rightarrow 2\mathbb{E}[Y-a \mid X=x]=0$

$$a = f(x) = \mathbb{E}[Y \mid X=x]$$

works for multivariate Y → regression function of Y onto X.

→ it's hard to perfectly compute $f(x)$ for all x since we are given only a few (X_i, Y_i) to observe.

→ $\mathbb{E}[Y|X=x]$ is only the mean, which doesn't fully capture $Y|X=x$.
↳ quantile regression uses Confidence bands around $\mathbb{E}[Y|X=x]$ instead

Linear Regression

$$f(x) = \mathbb{E}[Y|X=x] = x^T\beta$$

for some $\beta = \beta^* \in \mathbb{R}^k$ fixed.

use MLE to find β, but assume the following:

→ $Y|X=x \sim N(f(x), \sigma^2(x))$ [gaussian]

→ $f(x)$ linear → $f(x)=x^T\beta^*$

→ $\sigma^2(x)=\sigma^2$ $\forall x$ (constant)

then

$$\ell_n(\beta) = \sum_{i=1}^n \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - X_i^T\beta)^2}{2\sigma^2}\right)\right) = -\sum_{i=1}^n \frac{(Y_i - X_i^T\beta)^2}{2\sigma^2} + \text{const.}$$

minimize: $\hat{\beta}^{MLE} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$

for least squares

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n \qquad X = \begin{pmatrix} — X_1^T — \\ — X_2^T — \\ \vdots \\ — X_n^T — \end{pmatrix} \in \mathbb{R}^{n \times k}$$

want $\underset{\beta}{\operatorname{argmin}} \sum_{i=1} (Y_i - X_i^T \beta)^2 = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2$

gradient: $\nabla_\beta \|Y - X\beta\|^2 = 2X^T(Y - X\beta) = 0$

↳ $X^T Y = X^T X \beta \rightarrow \hat{\beta}^{LS} = (X^T X)^{-1} X^T Y$ (also $\hat{\beta}^{MLE}$)

$X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{P} Y$

"projection onto" span (X) $\qquad P^2 = P$

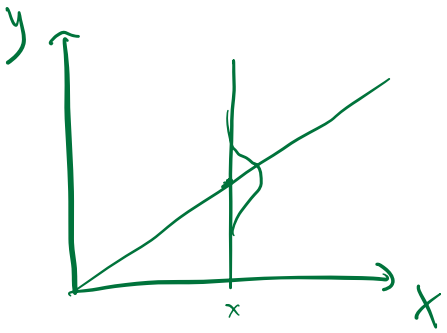
LECTURE 24 4/17/24 1PM

Lin. reg.

$X \in \mathbb{R}^k$, $Y \in \mathbb{R}$ predict Y from X

Given (X_i, Y_i) $i \in [n]$

Assume



$\mathbb{E}[Y | X = x] = x^T \beta^*$ with Gaussian errors

$Y_i = X_i^T \beta^* + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n \qquad X = \begin{pmatrix} — X_1^T — \\ — X_2^T — \\ \vdots \\ — X_n^T — \end{pmatrix} \in \mathbb{R}^{n \times k} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\underset{n \times 1}{Y} = \underset{n \times k}{X} \underset{k \times 1}{\beta} + \underset{n \times 1}{\vec{\varepsilon}} \qquad \vec{\varepsilon} \sim N(0, \sigma^2 I_n)$$

note $\hat{\beta}^{MLE} = \hat{\beta}^{LS} = (X^T X)^{-1} X^T (X^T \beta^* + \varepsilon) = (X^T X)^{-1} (X^T X) \beta^* + \underbrace{(X^T X)^{-1} X^T}_{A} \varepsilon$

$= \beta^* + A\varepsilon$

$\varepsilon \sim N(0, \sigma^2 I_n) \Rightarrow A\varepsilon \sim N(0, A \sigma^2 I_n A^T) \rightarrow N(0, \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1T})$

$\rightarrow N(0, \sigma^2 (X^T X)^{-1})$

ex) blood pressure $= Y$

$X = \begin{pmatrix} \text{age} \\ \text{weight} \\ \text{shoe size} \end{pmatrix}$ $\quad bp = \beta_1^* \text{age} + \beta_2^* \text{weight} + \beta_3^* \text{shoe size} = X^T \beta^*$

if $\beta_1^* > 0$ then $bp \uparrow$ as age $\uparrow$ if $\beta_1^* = 0$, no.

$\hat{\beta}_j^{LS} \sim N(\beta_j^*, \sigma^2 (X^T X)^{-1}_{jj})$ so CI: $\beta_j^* \in [\hat{\beta}_j^{LS} \pm \hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}\, z_{\alpha/2}]$

$X^T X = X_1 X_1^T + X_2 X_2^T + \cdots + X_n X_n^T$ so it plays role of $\frac{1}{\sqrt{n}}$

Hypothesis Testing

$H_0: \beta_j^* = 0$ vs $\beta_j^* \neq 0$ under null, $\hat{\beta}_j^{LS} \sim N(0, \sigma^2 (X^T X)^{-1}_{jj})$

Reject (Wald) if $\left| \frac{\hat{\beta}_j^{LS}}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}} \right| > z_{\alpha/2}$ $\quad$ p-value $= \mathbb{P}\left(|Z| > \frac{\hat{\beta}_j^{LS}}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}}\right)$

how to get $\hat{\sigma}$? use residuals $\hat{\varepsilon}_i = Y_i - X_i^T \hat{\beta}^{LS}$ column space of X has dimension k

so $\hat{\varepsilon} = Y - X^T \hat{\beta}$ dimension $n-k$

we can use $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-k} = \frac{1}{n-k} \sum_{i=1}^n \varepsilon_i^2$.

LECTURE 25 4/19/24 (Did not attend)

What if the Y_i's are binary?

↳ $Y | X = x \sim \text{Ber}(f(x)) \rightarrow \mathbb{E}[Y | X = x] = f(x)$

Now $x^T \beta^*$ is bad since this is unbounded.

Need $f(x) = \sigma(x^T \beta^*)$ where

- σ increasing
- $\sigma(t) \to 0$ as $t \to -\infty$, $\sigma(t) \to 1$ as $t \to \infty$
- $\sigma(0) = 1/2$

Alternative 1 [Logistic Regression Model]

$Y \mid X = x \sim \text{Ber}(\sigma(x^T \beta^*))$ where $\sigma(t) = \frac{e^t}{1+e^t}$ (sigmoid), (logistic).

Alternative 2 [Probit Regression Model]

$Y \mid X = x \sim \text{Ber}(\sigma(x^T \beta^*))$ where $\sigma(t) = \Phi(t)$

can also have $Y = \mathbb{1}(x^T \beta^* + Z > 0)$ where $Z \sim N(0,1)$

↑ any other $\tilde{Z}$ w/ symmetric distribution and cdf $F = \sigma$ can also be an alternative:

$Y \mid X = x \sim \text{Ber}(F(x^T \beta^*))$

$Y = \mathbb{1}(x^T \beta^* + \tilde{Z} > 0)$

for sigmoid

$$\mathbb{P}(Y_i \mid X_i) = \sigma(X^T\beta)^{Y_i} (1 - \sigma(X^T\beta))^{1-Y_i}$$

$$\frac{\sigma}{1-\sigma} = \frac{\frac{e^t}{1+e^t}}{\frac{1}{1+e^t}} = e^t$$

$$\ell_n(\beta) = \sum_{i=1}^n Y_i \log \sigma(X_i^T\beta) + (1-Y_i) \log(1 - \sigma(X_i^T\beta))$$

$$= \sum_{i=1}^n \left[Y_i \log \frac{\sigma(X_i^T\beta)}{1-\sigma(X_i^T\beta)} + \log(1 - \sigma(X_i^T\beta)) \right]$$

$\frac{1}{1+e^{X_i^T\beta}}$

concave so use gradient ascent → $= \sum_{i=1}^n [Y_i X_i^T\beta - \log(1 + e^{X_i^T\beta})]$

$\beta^{(0)} \in \mathbb{R}^K \qquad \beta^{(j+1)} = \beta^{(j)} + \gamma \nabla \ell_n(\beta)$

Multiclass Classification → $Y \in \{0, 1, \ldots, M\}$ instead of $\{0,1\}$

want $\mathbb{P}(Y = \ell \mid X = x) = p_\ell(x) \qquad \sum_{\ell=0}^M p_\ell(x) = 1 \quad \forall x.$

M degrees of freedom so we shouldn't use

$$p_j(x) = \frac{e^{x^T \beta_j^*}}{\sum_{\ell=0}^M e^{x^T \beta_\ell^*}}$$ (note $M+1 \neq M$ different β_j^*'s)

for $M=1$ recall

"logit" function $= \sigma^{-1}$

$$f(x) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}} \Rightarrow x^T\beta = \log\left(\frac{f(x)}{1-f(x)}\right) = \log\left(\frac{p_1(x)}{p_0(x)}\right)$$

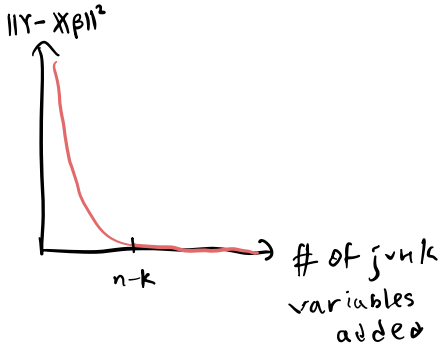
analogously let $\log\left(\frac{p_j(x)}{p_0(x)}\right) = x^T\beta_j$ for $j = 1, 2, \ldots, M$

$$\hookrightarrow p_j(x) = \frac{e^{x^T\beta_j}}{1+\sum_{\ell=1}^{M} e^{x^T\beta_\ell}}, \quad p_0(x) = \frac{1}{1+\sum_{\ell=1}^{M} e^{x^T\beta_\ell}}$$

LECTURE 26 4/22/24 1PM

Avoiding Overfitting → but how do we pick which features (variables) to include in β?

$\|Y - X\beta\|^2$ gets better the more features we use



"more junk → better predictor" BAD

Goal Given k features, pick subset S and get model

$$Y = \sum_{i \in S} \beta^{(i)} X^{(i)} + \varepsilon$$

Hypothesis Testing Fails : e.g. Wald's for each $\beta^{(j)}$

$$H_{0_j} : \beta^{(j)} = 0 \quad \text{vs.} \quad H_{1_j} : \beta^{(j)} \neq 0$$

↳ Bonferroni is too conservative.

↳ BH requires independent test statistics

Moreover,

↳ we want to test $H_{0_j} : \beta^{(j)} \neq 0$ vs. $H_{1_j} : \beta^{(j)} = 0$

↳ need $n-k$, not n to be large for asymptotics

New Framework, R^2

given $S \subset \{1, 2, \dots, k\}$

$\hat{\beta}$ using only subset S of variables

$$R^2(S) := 1 - \frac{\|Y - X\hat{\beta}(S)\|^2}{\|Y - \bar{Y}_n \mathbb{1}\|^2}$$

($\mathbb{1}$: all 1's vector; denominator: constant fit)

($R^2(S)$: coefficient of determination)

- better linear fit $\Rightarrow R^2$ closer to 1
- if one feature is always 1, then $0 \le R^2(S)$
- larger $|S|$ $\hookrightarrow$ larger $R^2(S)$

forward/greedy model selection

→ start with $S = \emptyset$

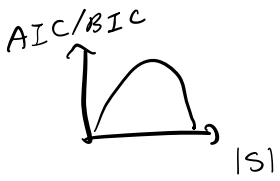
→ add $i \in \{1, 2, \dots, k\} \setminus S$ to S to maximize new $R^2(S)$.

→ stop when $R^2(S)$ plateaus (or exceeds desired R^2 threshold)

other metrics

AIC (Akaike info criterion) $= \ell_n(\hat{\beta}(S)) - |S|$

BIC (Bayesian info criterion) $= \ell_n(\hat{\beta}(S)) - \frac{\log n}{2}|S|$



Issues / Solutions

→ $\binom{k}{|S|}$ models for given $|S|$.

$\hookrightarrow$ use greedy to avoid cost

→ backward model selection (start with S and remove elements decreasing R^2 the least)

→ stepwise model selection (pick multiple features at a time)

→ AIC/BIC can be maximized without worrying about threshold due to self-penalization

LECTURE 27 4/24/24 IPM

Survival Analysis - estimate distribution of time until event of interest.

ex) # months since subs

T_i	cancelled δ_i	C_i
5	1	≥ 5
12	1	≥ 12
7	1	≥ 7
35	1	≥ 35
48	0 (censored)	48
6	1	≥ 6
48	0 (censored)	48
48	0 (censored)	48

random variable T, observe $T_1, \dots, T_n \overset{iid}{\sim} T$

Def Let F be the cdf of T. Survival function is $S(t) = 1 - F(t)$.

we want to estimate $S(t)$

If we know all the T_i, we can just say $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_i > t)$

unbiased, consistent, asymp. normal

$\sqrt{n}[\hat{S}(t) - S(t)) \rightsquigarrow N(0, S(t)(1-S(t)))$

but some T_i may be censored due to dropout / ongoing.

so each T_i has a censoring time $C_i \overset{iid}{\sim} C$ and we instead see

$$\tilde{T}_i = \min(T_i, C_i), \quad \delta_i = \mathbb{1}(C_i \geq T_i) = \mathbb{1}(\tilde{T}_i = T_i)$$

throw out censored data?

$$\hat{S}^{naive}(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(T_i > t, C_i = \infty)}{\sum_{i=1}^{n} \mathbb{1}[C_i = \infty)}$$

include it?

$$\hat{S}^{better}(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(\tilde{T}_i > t)}{\sum_{i=1}^{n} \mathbb{1}(C_i \geq t)} \approx \mathbb{P}(T > t \mid C \geq T)$$

Kaplan-Meier Estimator

$$S(t) = \mathbb{P}(T > t) = \frac{\mathbb{P}(T>t)}{\mathbb{P}(T>t-1)} \cdot \frac{\mathbb{P}(T>t-1)}{\mathbb{P}(T>t-2)} \cdot \dots \cdot \frac{\mathbb{P}(T>1)}{\mathbb{P}(T>0)} \leftarrow 1$$

$q(t)$, $q(t-1)$, $q(1)$

Hazard Rate

$h(t) = 1 - q(t)$ = probability that you die by t given you survived to $t-1$.

$$\hat{h}(s) = \frac{\sum_{i=1}^{n} \mathbb{1}(\tilde{T}_i = s, C_i \geq s)}{\sum_{i=1}^{n} \mathbb{1}(\tilde{T}_i \geq s)} \xrightarrow{\times \frac{1}{n}} \frac{\mathbb{P}(\tilde{T} = s, C \geq s)}{\mathbb{P}(\tilde{T} \geq s)} = \frac{\mathbb{P}(\tilde{T} = s, C \geq s)}{\mathbb{P}(\tilde{T} \geq s, C \geq s)} = \frac{\mathbb{P}(\tilde{T} = s)}{\mathbb{P}(\tilde{T} \geq s)} = h(s)$$

$\min(T, C)$; indep.; indep.

$$S(t) = \prod_{s=1}^{t}(1 - h(s)) = \prod_{s=1}^{t}\left(1 - \frac{\#\{i : \tilde{T}_i = s, C_i \geq s\}}{\#\{i : \tilde{T}_i \geq s\}}\right)$$

variations - e.g. if it depends on features $x \in \mathbb{R}^k$

So survival function can become $S(t,x)$ depending on selected features x.

Cox proportional hazard regression model

$h(t,x) = h_0(t) \exp(\beta^T x)$ → decouples effect of t vs. x

also, continuous version of Kaplan-Meier.

$$h(t) = \frac{\mathbb{P}(T=t)}{\mathbb{P}(T>t-1)} \quad \text{vs.} \quad h(t) = \frac{\mathbb{P}(t \le T \le t+dt)}{\mathbb{P}(T \ge t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log(S(t))$$

LECTURE 28 4/26/24 (Did not attend)

causation ≠ correlation

↳ Counterfactual Model $X \in \{0,1\}$ no vs. yes treatment

$Y \in \mathbb{R}$ response

potential outcomes: $Y = \begin{cases} C_0 & \text{if } X=0 \\ C_1 & \text{if } X=1 \end{cases}$

X	C_0	C_1
0	obs.	counterfactual
1	counterfactual	obs.

average treatment effect (ATE) $\theta := \mathbb{E}[C_1] - \mathbb{E}[C_0]$

association $\alpha := \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$

not necessarily the same; e.g. if $Z \sim \text{Unif}([-1,1])$, $X = \mathbb{1}(Z>0)$

outcomes $C_0 = C_1 = Z$.

Then $\theta = 0$, $\alpha = 1$. → Z is confounding variable.

We want to find θ, so we try to estimate it using α.

If $\alpha = \theta$, just use plug-in estimator

$$\hat{\alpha} = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i=1)}{\sum_{i=1}^n \mathbb{1}(X_i=1)} - \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i=0)}{\sum_{i=1}^n \mathbb{1}(X_i=0)}.$$

To guarantee, we can use randomized control trials - flip coin with prob p to assign $X_i = 1$ and $1-p$ to assign $X_i = 0$.

↳ $\hat{\alpha} \approx \alpha = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] = \mathbb{E}[C_1|X=1] - \mathbb{E}[C_0|X=0] = \mathbb{E}[C_1] - \mathbb{E}[C_0] = \theta$.

↳ we have $(C_0, C_1) \perp\!\!\!\perp X$ → "strong unconfoundedness"
(independent)

↳ can be unethical to deny treatment by coin flip.

Sometimes, we can instead find a feature Z to condition on s.t. (e.g. age, salary, etc.)

$$C_i \perp\!\!\!\perp X | Z \quad \text{for} \quad i = 0, 1.$$

but if Z is high dimensional, we can't use $\hat{\theta}_1 = \hat{\alpha}(z_1)$, $\hat{\theta}_2 = \hat{\alpha}(z_2)$.

instead we use propensity score $p(z) = \mathbb{P}(X=1 \mid Z=z)$

↳ likelihood of assigning treatment based on $Z=z$.

Thm If $C_i \perp\!\!\!\perp X|Z$, then $C_i \perp\!\!\!\perp X | p(Z)$.

so we can just divide X into bins based on $p(z)$ and compute $\hat{\theta}_j$ for each.

↳ calculate $p(z)$ using logistic regression: $\text{logit}(p(z)) = z^T \beta$.

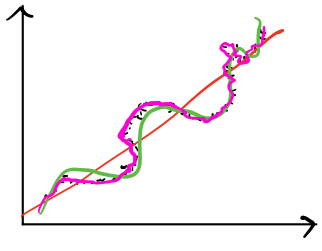
LECTURE 29 4/29/24 1PM

Nonparametric Curve Estimation

1: density estimation — estimate density $p(x)$ of iid $X_1, \dots, X_n$

2: regression — estimate $f(x) = \mathbb{E}[Y|X=x]$ given iid $(X_1, Y_1), \dots, (X_n, Y_n)$

we don't want too smooth nor too choppy.



want a tradeoff after a smoothness assumption.

↑ Smoothness ↓ | ↑ bias ↓ | ↓ variance ↑

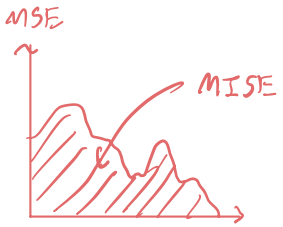
smoothness means close to Taylor approx of some order

Suppose we estimate g by $\hat{g} \Rightarrow \hat{g}(x)$ for $g(x)$ for each x.

Bias: $b(x) = \mathbb{E}[\hat{g}(x)] - g(x)$

Variance: $v(x) = \mathbb{V}[\hat{g}(x)]$

MSE: $b(x)^2 + v(x)$



MISE: $\int \text{MSE}(\hat{g}(x))\,dx = \mathbb{E}\left[\int (\hat{g}(x) - g(x))^2 dx\right]$

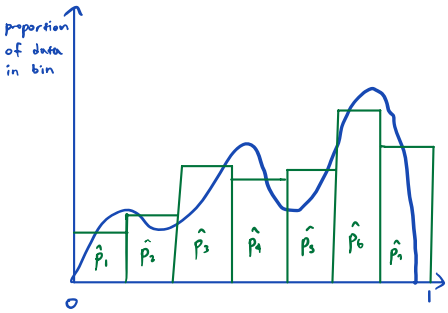
MISE is also called risk

$\int b(x)^2 dx + \int v(x)\,dx.$

Density Estimation

probability density function

Suppose $X_1, \dots, X_n \overset{iid}{\sim} p(x)$



① histogram estimator $\hat{p}(x)$

m bins $B_1, \dots, B_m$ each width $|B_j| = \frac{1}{m} = h$

Let $n_j = \#\{i : X_i \in B_j\} = \sum_{i=1}^{n} \mathbb{1}(X_i \in B_j)$

Then let $\hat{p}_j = \frac{n_j}{n}$. Define $\hat{p}(x) := \frac{\hat{p}_j}{h} = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}(x \in B_j)$

↑ normalization

note $\hat{p}(x) \geq 0 \quad \forall x \geq 0$ and

$$\int \hat{p}(x)\,dx = \sum_{j=1}^{m} \int \frac{\hat{p}_j}{h} \mathbb{1}(x \in B_j)\,dx = \sum_{j=1}^{m} \int_{B_j} \frac{\hat{p}_j}{h}\,dx = \sum_{j=1}^{m} h \cdot \frac{\hat{p}_j}{h} = \sum_{j=1}^{m} \hat{p}_j = 1.$$

bias for given $x \in B_j$

$$b(x) = \mathbb{E}[\hat{p}(x)] - p(x) = \mathbb{E}\left[\frac{\hat{p}_j}{h}\right] - p(x)$$

but $\hat{p}_j = \frac{n_j}{n}$, $n_j \sim \text{Bin}\left(n, \int_{B_j} p(y)\,dy\right)$

so $$\mathbb{E}\left[\frac{\hat{p}_j}{h}\right] - p(x) = \frac{1}{h}\int_{B_j} p(y)\,dy - p(x) \to 0 \text{ as } h \to 0.$$

Also true that $\int b(x)^2 dx \to 0$ as $h \to 0$.

$\Rightarrow$ smaller h means smaller bias

variance

$$v(x) = \mathbb{V}\left[\frac{\hat{p}_j}{h}\right] = \frac{1}{h^2}\mathbb{V}[\hat{p}_j] = \frac{1}{h_n^2}\left[\int_{B_j} p(y)\,dy\right]\left[1 - \int_{B_j} p(y)\,dy\right]$$

$$= \frac{1}{hn} \cdot \frac{\int_{B_j} p(y)\,dy}{h} \cdot 1 - \int_{B_j} p(y)\,dy$$

$\Rightarrow$ smaller h means bigger variance.

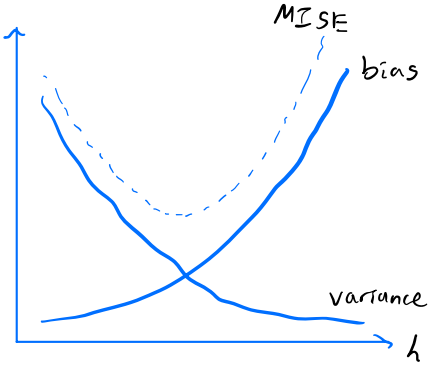
$$\to \frac{1}{hn} \cdot p(x) \cdot 1 = \frac{p(x)}{hn}$$

as $h \to 0$.

so we have a tradeoff!

→ to find optimal h without actually knowing p,

we use cross-validation



LECTURE 30 5/6/24 1PM

Cross-validation: note $\text{MISE} \Leftarrow \int (\hat{p}(x) - p(x))^2\,dx$ ← unbiased estimator

$$= \underbrace{\int \hat{p}(x)^2\,dx}_{\text{we know}} - 2\int \hat{p}(x)p(x)\,dx + \underbrace{\int p(x)^2\,dx}_{\text{constant, no need to optimize}}$$

note $\int \hat{p}(x)p(x)\,dx = \mathbb{E}_{X\sim p}[\hat{p}(X)]$

$\hat{p}(x)$ depends on X_i so write as $\hat{p}(x; X_1, X_2, \ldots, X_n)$

$$\int \hat{p}(x)p(x)\,dx = \mathbb{E}_{X\sim p}[\hat{p}(X; X_1, \ldots, X_n) \mid X_1, \ldots, X_n]$$

suppose we have some $X_1', \ldots, X_n'$ independent of $X_1, \ldots, X_n$ ⟶

$$= \frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i'; X_1, \ldots, X_n)$$

↳ $$\int \hat{p}(x)p(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i; X_{(-i)})$$ ← $\{X_1, \ldots, X_n\} \setminus \{X_i\}$

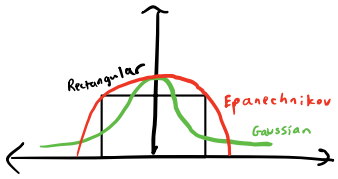
so minimize $$\widehat{\text{MISE}} = \int_0^1 \hat{p}(x)^2\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{p}(X_i; X_{(-i)}) + \text{const.}$$

over bin width h.

② Kernel density estimator – smooth estimator of density.

↳ Kernel is any function k s.t. $k(x) \geq 0$, $\int k(x)\,dx = 1$, $\int x\,k(x)\,dx = 0$.

↑ any even pdf works.

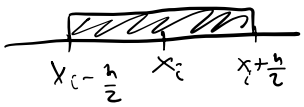


KDE is $\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$ h is bandwidth here.

(how wide each kernel stretches)

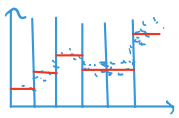
ex) rectangular kernel $k(x) = \mathbb{1}(|x| \leq \frac{1}{2})$ "sliding bin histogram"

$$k\left(\frac{x - X_i}{h}\right) = \mathbb{1}\left(\left|\frac{x - X_i}{h}\right| \leq \frac{1}{2}\right) = \mathbb{1}\left(x \in \left[X_i - \frac{h}{2}, X_i + \frac{h}{2}\right]\right)$$



Nonparametric Regression Given (X_i, Y_i), predict $E[Y|X=x]$ if nonlinear.

① Regressogram – break into various bins B_j along x-axis



$$E[Y|X=x] \approx \hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i\,\mathbb{1}(X_i \in B_j)}{\sum_{i=1}^{n} \mathbb{1}(X_i \in B_j)}$$ → estimate as average of each bin.

② Nadaraya–Watson Estimator – sort of like KDEs

↳ estimate both $p(x,y)$ and $p(y)$ using KDEs

then $E[Y|X=x] = \int y\,p(y|x)\,dx = \frac{\int y\,p(x,y)\,dx}{p(x)} \approx \frac{\int y\,\hat{p}(x,y)\,dx}{\hat{p}(x)}$

$$\Rightarrow \hat{f}(x) = \frac{\frac{1}{nh^2}\int y \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) k\left(\frac{y - Y_i}{h}\right) dy}{\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)} = \frac{\sum_{i=1}^{n} Y_i\,k\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)}.$$

curse of dimensionality: $MISE \sim \left(\frac{1}{h}\right)^k / n$ with k-dimensional X_i

↳ need LARGE n to make MISE small

LECTURE 31 5/8/24 1PM

Survey Sampling - pick a representative set of people for survey.

ex) MIT undergrads who prefer online > in-person. $N = 4657$.

Let $y_i = \begin{cases} 1, & \text{if } i \text{ prefers online} \\ 0, & \text{if } i \text{ prefers in-person} \end{cases}$ We want $T = \sum_{i=1}^{N} y_i$

we instead find a representative sample $S \subset \{1, 2, \ldots, N\}$, use

$$\hat{T} = \sum_{i \in S} \lambda_i y_i$$

$\lambda_i \geq 1$ compensation weights

Now, how do we choose S? The probability distribution of the random subset S is called the design.

common design choices

↳ without replacement

one S from all $S \subset \{1, \ldots, N\}$ with $|S| = n$ $\rightarrow$ $\mathbb{P}(S = s) = \frac{1}{\binom{N}{n}}$

↳ random

each $i \in \{1, \ldots, N\}$ is included in S with $p = \frac{n}{N}$, size of S is $|S| \sim \text{Bin}(N, \frac{n}{N})$, expected size n.

↳ stratified

break up into groups (e.g. major) then sample from each group. S union of samples

constructing an estimator

let p = pmf of random subset S. $p(s) = \mathbb{P}(S = s)$ where $s \in \{1, \ldots, N\}$.

let $\pi_i = \mathbb{P}(i \in S) = \sum_{\substack{s \in \{1, \ldots, N\} \\ i \in s}} p(s)$ be the inclusion probability.

let $\pi_{ij} = \mathbb{P}(i, j \in S)$ be similarly defined.

Horvitz-Thompson Estimator

$$\hat{T}^{HT} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \mathbb{1}(i \in S)$$

$\lambda_i = \frac{1}{\pi_i}$ here.

for sampling w.o. replacement

$$\hat{T}^{HT} = \frac{N}{n} \sum_{i \in S} y_i$$

$$\mathbb{E}[\hat{T}^{HT}] = \sum_{i=1}^{N} \mathbb{E}\left[\frac{y_i}{\pi_i} \mathbb{1}(i \in S)\right] = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \mathbb{P}(i \in S) = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \cdot \pi_i = T \rightarrow \text{unbiased.}$$

$$\mathbb{V}[\hat{T}^{HT}] = \mathbb{V}\left[\sum_{i=1}^{N} \frac{y_i}{\pi_i} \mathbb{1}(i \in S)\right] = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} \operatorname{Cov}(\mathbb{1}(i \in S), \mathbb{1}(j \in S))$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$ ← in random sampling $\mathbb{1}(i \in S) \perp\!\!\!\perp \mathbb{1}(j \in S)$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} (\pi_i - \pi_i^2) = \sum_{i=1}^{N} y_i^2 \left(\frac{1 - \pi_i}{\pi_i}\right)$$

but then this can be large if π_i small

estimate variance using

$$\hat{\mathbb{V}}(\hat{T}^{HT}) = \sum_{i \in S} \sum_{j \in S} \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} \cdot \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$$

then (in some design choices)

$$\frac{\hat{T}^{HT} - T}{\hat{\mathbb{V}}(\hat{T}^{HT})} \rightsquigarrow N(0, 1)$$

Methods for sampling w.o. replacement

1. one at a time

   N choices for 1st, $N-1$ for 2nd, ..., $N-n+1$ for n^{th} $\Rightarrow O(Nn)$.

2. give each person a uniform random variable $\in [0, 1]$

   sort random variables and take n people with smallest variables.

   $O(N \log N)$

3. reservoir

   → set $S = \{1, 2, \ldots, n\}$

   → for $k = n+1, \ldots, N$, replace a random element of S with k with $p = \frac{n}{k}$. Else, do nothing.

   $O(n^2 \log N)$ expected

LECTURE 32 5/10/24 (Did not attend)

classifier $h: X \to Y \in \{0,1\}$ (binary) → defines a classification region

ex) linear classifier $h(x) = \mathbb{1}(w^T x + b \geq 0)$

$k+1$ parameters $(w_1, \ldots, w_k, b)$

Bayes Classifier h^*

minimize probability of making a mistake.

true error $\mathbb{P}(Y \neq h(X))$ vs. empirical error $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq h(X_i))$

$h^* := \operatorname{argmin}_h \mathbb{P}(Y \neq h(X)) = \mathbb{1}(r(x) \geq 1/2)$

same since both minimize error → $\mathbb{P}(Y=1 \mid X=x)$

↳ alternative way to arrive at this.

let proportion of 0's, 1's in Y be $\frac{1}{2}$ each.

$f_0(x) = f(x \mid Y=0)$ dist. of 0's

$f_1(x) = f(x \mid Y=1)$ dist of 1's

then

$$r(x) = \mathbb{P}(Y=1 \mid X=x) = \frac{f_1(x) \cdot \frac{1}{2}}{f_1(x) \cdot \frac{1}{2} + f_0(x) \cdot \frac{1}{2}} = \frac{f_1(x)}{f_1(x) + f_0(x)} \geq \frac{1}{2} \text{ iff } f_1(x) \geq f_0(x)$$

3 estimators of Bayes Classifier h^*

① $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq h(X_i))$ "empirical risk minimizer" (ERM)

↳ brute force, perceptron

② $\hat{h} = \mathbb{1}(\hat{r}(x) \geq \frac{1}{2})$, using $\hat{r}(x) = \sigma(\hat{\beta}^T x)$ for example

③ $\hat{h} = \mathbb{1}(\hat{f}_1(x) \geq \hat{f}_0(x))$, using density estimators (eg. KDEs) of choice

→ ex) Fisher's discriminant analysis

LECTURE 33 5/13/24 (did not attend)

More on (3)

can use $\hat{f}_0(x) = \frac{1}{n_0 h^k} \sum_{i : Y_i = 0} K\left(\frac{\|X_i - x\|}{h}\right)$ (# of $Y_i = 0$; k = dimension of X_i)

$$\hat{f}_1(x) = \frac{1}{n_1 h^k} \sum_{i : Y_i = 1} K\left(\frac{\|X_i - x\|}{h}\right)$$

but error $|f_1(x) - \hat{f}_1(x)| \propto n^{-1/(2+k)}$ so big k is problematic without a LOT of data

Solutions

(1) Naive Bayes. → assume all features independent.

$$\hat{f}_0(x) = \prod_{j=1}^{k} \hat{f}_0^{(j)}(x^{(j)}) \qquad \hat{f}_1(x) = \prod_{j=1}^{k} \hat{f}_1^{(j)}(x^{(j)})$$

(2) Discriminant Analysis → assume f_0, f_1 normal.

then

$$\hat{f}_0(x) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma_0)}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right)$$

$$\hat{f}_1(x) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma_1)}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)\right)$$

if $\Sigma_0 = \Sigma_1 = \Sigma$,

$$\hat{f}_1(x) \geq \hat{f}_0(x) \Leftrightarrow (x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0) \geq (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)$$

$$\Leftrightarrow x^T (\Sigma^{-1}(\mu_1 - \mu_0)) \leq \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0)$$

we can estimate μ_1, μ_0, Σ to get a classifier using sample mean, covariance.

if $\Sigma_0 \neq \Sigma_1$, we get a quadratic boundary,

→ "quadratic discriminant analysis"

③ k - nearest neighbor classifier

$\hat{h}(x) = 1$ if $\geq \frac{k}{2}$ nearest neighbors are $Y_i = 1$, else 0.

④ logistic regression

0,1

$$\hat{p}(x) = \frac{e^{\hat{\beta}^{MLE\,T} x}}{1 + e^{\hat{\beta}^{MLE\,T} x}} \qquad \hat{h}(x) = \mathbb{1}(\hat{p}(x) \geq \tfrac{1}{2})$$

multiclass

$$\hat{p}_j(x) = \frac{e^{\hat{\beta}_j^{MLE\,T} x}}{1 + \sum_{\ell=1}^{M} e^{\hat{\beta}_\ell^{MLE\,T} x}} = \mathbb{P}(Y = j \mid X = x) \quad \text{for } j = 0, 1, \ldots, M$$

(1 if $j=0$, M different $\hat{\beta}_j^{MLE}$'s)

$$\hat{h}(x) = \operatorname{argmax}_{j=0,\ldots,M} \hat{r}_j(x) \qquad \log\left(\frac{p_j(x)}{p_0(x)}\right) = x^T \beta_j$$

⑤ neural networks

$$\log\left(\frac{p_j(x)}{p_0(x)}\right) = f_{\theta_j}(x)$$

more complex function parametrized by θ

last layer does multiclass logistic regression