STAT 391

5/4/23

Lecture 12

Part I ML Estimation ✓

Quiz 2 upcoming

- Mixture Models ← Estimation

  Model selection ← between Normal and KDE

  Advanced estimation

  — ML estimator as RV

  — Bayesian estimation

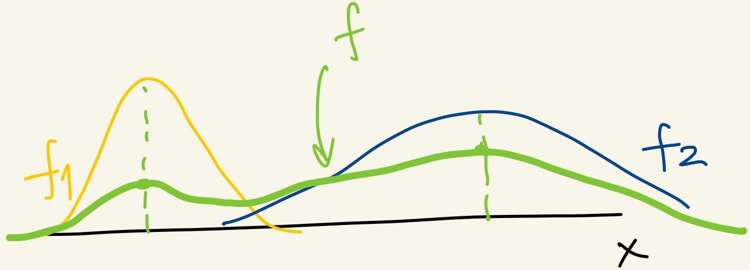
2 or more RV → prediction
- regression $X \to Y$ $P_{Y|X} = ?$
- classification
- ranking
- structured prediction
- link prediction

↳ reasoning $X, Y, Z, \ldots$ (causal)

$X \to Y$, $Z \to Y$

Mixtures of Gaussians

$$\underbrace{f(x)}_{\text{mixture density}} = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{mixture proportions}} \underbrace{f_k}_{\text{components}}(x; \underbrace{\mu_k, \sigma_k^2}_{\theta_k})$$

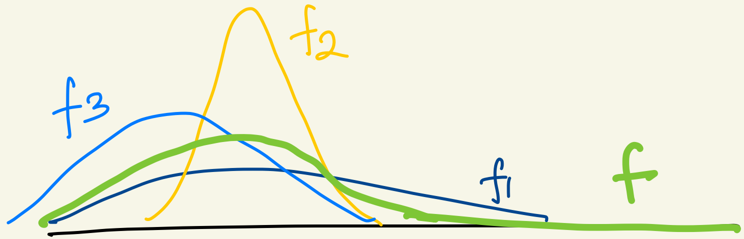


$\pi_1 = \frac{1}{3}$ $\qquad$ $\pi_2 = \frac{2}{3}$

need

$f(x) \geq 0$ for all x $\Leftrightarrow$ $\pi_{1:k} \geq 0$

$\int_{\mathbb{R}} f(x)\,dx = 1$ $\qquad$ $\sum_{k=1}^{K} \pi_k = 1$



$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

Gaussian
- single Normal(μ, σ^2)
- KDE $n \times$
- Mixture $K \times$

K → fixed, not growing with n

KDE for $x \in \mathbb{R}^d$

$K_h(z) = N(0, h^2 I_d)$

Skipped but not forgotten

- ML estimation for $N(\mu, \Sigma)$ for $x \in \mathbb{R}^d$ exact
- ML —u— for Mixtures EM Algorithm

Model Selection in Statistics

Given $\mathcal{D}$

Model ~~family~~ families

(try them all)

$\mathcal{F}_1$ (gaussian) $\xrightarrow{\text{estimation}}$ $f_1 \in \mathcal{F}_1$ best in family

$\mathcal{F}_2$ (logistic) $\rightarrow$ f_2

$\mathcal{F}_3$ (KDE h_1) $\rightarrow$

$\mathcal{F}_4$ (-"- h_2) $\rightarrow$ f_4

$\dots$ $\dots$

mixture $\rightarrow$ f_M

Select "best" $f_1, f_2, \dots f_M$

$$f^* = \underset{k=1:M}{\text{argmax}} \ \text{score}(f_k)$$

CV

AIC, BIC

Lecture Notes VI – Model selection by AIC and BIC

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2021

AIC Akaike's Information criterion (max)

BIC Bayesian —"— —"— (max)

↳ restricted to simpler parametric families

↳ only for parametric uses

only for ML $\rightarrow \theta^{ML}$.

AIC and BIC

The number of parameters d

Ex: $N(\mu, \sigma^2) \Rightarrow d = 2$

$N(\mu, \Sigma) \Rightarrow d = \underset{\substack{\uparrow \\ \text{for } \mu}}{m} + \underset{\substack{\uparrow \\ \text{for } \Sigma}}{\frac{m(m+1)}{2}}$

$x, \mu \in \mathbb{R}^m$

$\Sigma \in \mathbb{R}^{m \times m}$

multivariate normal

AIC and BIC

Logistic $(a,b) \Rightarrow d=2$

Exponential $(\lambda) \Rightarrow d=1$

Mixture of 3 Gaussians $\Rightarrow d = \overbrace{3\times 2}^{K} + \overbrace{(3-1)}^{K-1} = 8$ for π

$$f(x) = \sum_{k=1}^{3} \pi_k f_k(x; \underbrace{\mu_k, \sigma_k^2}_{2})$$

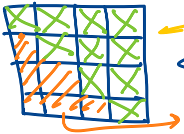
$x \in \mathbb{R}$

Discrete categorical

$S = \{1, \dots m\} \to \theta_{1:m} \Longrightarrow d = m-1$

(# free parameters

Σ $m \times m$ covariance matrix



$\frac{m(m+1)}{2}$

symmetric

known by symmetry

Assume $\Sigma = \sigma^2 I_m$, $\mu \in \mathbb{R}^m$ $\Big\} \Rightarrow d = \underbrace{m}_{\mu} + \underbrace{1}_{\sigma^2}$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{bmatrix} \Rightarrow d = \underbrace{m}_{\mu} + \underbrace{m}_{\sigma^2_{1:m}}$$

AIC and BIC

Multivariate Normal (μ, Σ)

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{\frac{m}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$$-(x-\mu)^2 \cdot \frac{1}{2\sigma^2}$$

$x, \mu \in \mathbb{R}^{\underline{\underline{m}}}$

$\Sigma = \mathbb{R}^{m \times m}$

$$\boxed{(x-\mu)^T} \, \boxed{\Sigma^{-1}} \, \underbrace{\boxed{}}_{x-\mu} \in \mathbb{R}$$

AIC and BIC

AIC ≈ CV

BIC ← wants "correct" model

theoretically NOT for mixtures

Hold for

- parametric $\mathcal{F} = \{f_\theta\}$ with θ a vector of parameters, e.g. (μ, σ^2), γ, a, b (for logistic)
- log-likelihood $l(\theta) = -\ln P(x^{1:n}|f_\theta)$
- $f^{ML} \in \mathcal{F}$ estimated by Maximum Likelihood
- (for BIC: $\frac{\partial^2 l}{\partial parameters}$ non-singular at f^{ML})

$f(x;\theta^{ML}) \in \mathcal{F}$ model family

Akaike's Information Criterion (AIC)

$$\text{score} \quad AIC(f^{ML}) = l(f^{ML}) - d, \tag{1}$$

log-likelihood

$\ell(\theta^{ML};\mathcal{D})$

where $d = \#parameters(f)$, and $n =$ the size of $\mathcal{D}$.

The Bayesian Information Criterion (BIC)

$$\text{score} \quad BIC(f^{ML}) = +\ell(f^{ML}) - \frac{d}{2}\ln n, \tag{2}$$

good to fit data better

penalize for more variance

with $d = \#parameters(f)$

✓ Estimation < discrete / continuous

✓ Model Selection

→ Stat. estimators as R.V

data

$$\{x^{1:n}\} = \mathcal{D} \xrightarrow{ML} \theta^{ML}$$

$$x \in S_x, \qquad \{x^{1:n}\} \in S_x^n, \qquad \theta^{ML} \in S_{\theta^{ML}}$$

$$P_x \xrightarrow[iid]{} P_x^n \xrightarrow[P_x]{n} P_{\theta^{ML}},\ E[\theta^{ML}],\ Var\,\theta^{ML}$$

assume known

Wanted: θ^{true} $\quad 0$

$\underbrace{\text{for } n \to \infty}_{\text{consistency}}$

- $E[\theta^{ML}] = \theta^{true}$ for finite n

Unbiasedness

$$E[\theta^{ML}] - \theta^{true} = \text{bias}$$

a special form of bias (refers to a parameter)

Normal (μ, σ^2)

- $\boxed{\mu^{ML} = \frac{1}{n}\sum_{i=1}^{n} x^i}$

$$\boxed{E[\mu^{ML}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x^i\right] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{E[x^i]}_{\mu} = \mu}$$

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

$\rightarrow$ unbiased

randomness in sample !!

- $$\boxed{Var\ \mu^{ML} = Var\left[\frac{1}{n}\sum x^i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{Var(x^i)}_{\sigma^2} = \frac{n\sigma^2}{n^2} = \frac{1}{n}\sigma^2}$$

if σ^2 smaller, n can be smaller, too

iid (independent!)

$Var \searrow 0$ for $n \rightarrow \infty$