STAT

5/16/23

Lecture 15

Bayesian ~~Est~~imation

- for multinomial

(for Normal)

Regression

Vote: Bayesian for $N(\mu, \sigma^2)$ YES ~~NO~~

Given $\mathcal{D} = \{x^{1:n}\} \subset S = \{1, \ldots m\}$

Model family $\{\underbrace{(\theta_{1:m})}_{\in \Delta_m}$ probability distrib. over $S\} = \mathcal{F}$

$\Delta_m = \{\theta_{1:m} \geq 0, \sum_{j=1}^{m} \theta_j = 1\}$ simplex

Prior $\rightarrow$ over $\mathcal{F}$

$S_\theta \equiv \Delta_m$ $\quad f_0$ = density on Δ_m

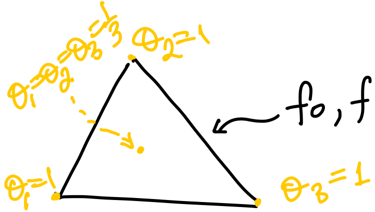
Want: Posterior

f = density on Δ_m

Bayesian estimation

- $S=\{1,2,\dots m\}$, Parameters $\theta_{1:m} \geq 0$

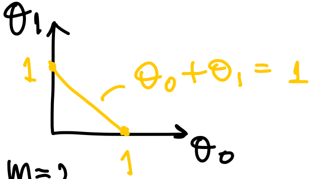
$$\underbrace{\theta_1+\theta_2+\dots+\theta_m=1}_{\text{prior } f_0(\theta_{1:m})}$$



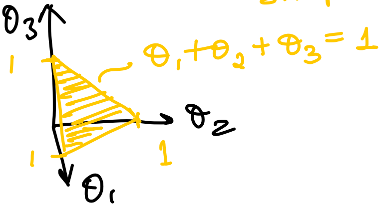
f_0, f over Δ_m

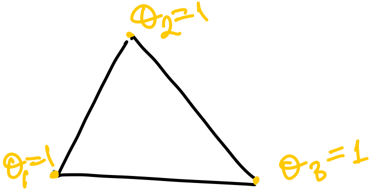
$m=3$ $\quad \theta_1+\theta_2+\theta_3=1$

Probability simplex Δ_m



$m=2$





$m = 3 \qquad \theta_1 + \theta_2 + \theta_3 = 1$

Likelihood $(\mathcal{D} \mid \theta_{1:m}) =$

$$= \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_m^{n_m} \leftarrow \text{counts}$$

$$n = n_1 + \cdots + n_m$$

$$\mathcal{P} = \{ \text{Diri}(\alpha_{1:m}), \quad \alpha_{1:m} > 0 \}$$

conjugate prior

$$\alpha = \alpha_1 + \cdots + \alpha_m$$

$$\text{Diri}(\theta_{1:m} \mid \underbrace{\alpha_{1:m}}_{\text{hyper params}}) = \frac{1}{Z} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_m^{\alpha_m - 1}$$

$\theta_{1:m} \in \Delta_m$

$$\boxed{Z = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_m)}{\Gamma(\alpha)}}$$

$$\boxed{\alpha = \alpha_1 + \cdots + \alpha_m}$$

Posterior

$$f(\theta_1,..\theta_m) \propto \underbrace{f_0(\theta_1..\theta_m)}_{\text{Dirichlet}(\alpha_1,..\alpha_m)} \; \underbrace{L(\mathcal{D}\,|\,\theta_1..\theta_m)}_{\theta_1^{n_1}..\theta_m^{n_m}}$$

CALCULUS

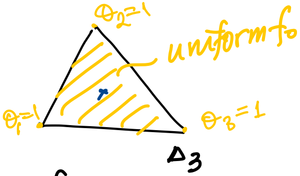
$$\propto \underbrace{\theta_1^{\alpha_1+n_1-1}\,\theta_2^{\alpha_2+n_2-1}\cdots\theta_m^{\alpha_m+n_m-1}}_{!!\ \text{Dirichlet}(\alpha_1+n_1,\ \ldots,\ \alpha_m+n_m)}$$

Ex

$m=3 \Rightarrow \alpha_{1,2,3}$ hyperparams

punkten

$\mathcal{D} = \{1,1,2,1,1\}$ iid $\Rightarrow$ $n=5$, $n_1=4$, $n_2=1$, $n_3=0$



Prior uniform = Dirichlet$(1,1,1) = f_0$

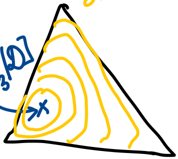
$\alpha_1'-1$ $\alpha_2'-1$ $\alpha_3'-1$

Posterior $f \propto 1 \cdot \underbrace{\theta_1^4\theta_2^1}_{\text{likelihood}} = \theta_1^4\,\theta_2^1\,\theta_3^0 = \text{Diri}(5,2,1)$

$\tilde{f}_0 = \text{Diri}(2,2,2)$

$\theta_2 = 1$

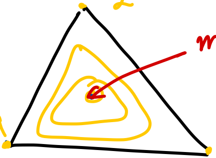
$E[\theta_{1:3}|D]$



$\theta_1 = 1$ $\theta_3 = 1$

$\theta_2 = 1$

$\max f_0 = \text{uniform}\{1,2,3\}$



$\theta_1 = 1$ $\theta_3 = 1$

$\text{Diri}(\underset{\downarrow 1}{5}, \underset{\downarrow 2}{2}, \underset{\downarrow 3}{1}) \Rightarrow \alpha' = 5+2+1 = 8$ fict. sample size

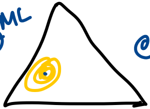
$$\underbrace{\tilde{\theta}_1 = \frac{5}{8},\ \tilde{\theta}_2 = \frac{2}{8},\ \tilde{\theta}_3 = \frac{1}{8}}_{\text{Expected } \theta_{1:3} \text{ under } f}$$

$$E[\theta_j] = \frac{\alpha_j}{\alpha}$$

Dirichlet

$n \nearrow \Rightarrow$

θ^{ML}



concentrates near θ^{ML}

Bayesian prediction

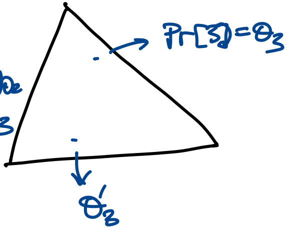
Ex: $S = \{1, 2, 3\} \Rightarrow X$

$f(\theta_1, \theta_2, \theta_3) = \text{Dirichlet}(5, 2, 1)$

$\Pr[X=3] = ?$

$\Pr[X=3] = E_f[\theta_3]$

$$= \int_{\Delta_m} \theta_3 \frac{1}{Z} \theta_1^{5-1} \theta_2^{4-1} \theta_3^{1-1} \, d\theta_1 d\theta_2 d\theta_3$$



$P[\text{event}] = p(\theta)$ depends on $(\theta_{1,2,3})$

sequence $(1, 3, 3)$

$x \le 2$

- - - -

$$= \int_{\Delta_m} p(\theta) \cdot \underbrace{f(\theta)}_{\text{Dirichlet}} \, d\theta$$

Bayesian Statistics in Real Life

- when not sure about model family
  - averaging more "wrong" models
  - more flexible modeling
- when not sure about θ in model family $\mathcal{F}$
  - not enough data – e.g. smoothing
  - noisy data ← reduce variance
  - extreme values
- reason about single case
- data in stages (on-line)

$f(\theta)$ prior $+ \mathcal{D}_1 \to f(\theta \mid \mathcal{D}_1) \to f(\theta \mid \mathcal{D}_1 \cup \mathcal{D}_2) \to \cdots$

prior $\Leftrightarrow$ $\mathcal{D}_2$ $\mathcal{D}_3$

(transfer; domain adaptation)

Lecture Notes I – Regression & Prediction

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

May, 2023

Prediction ←

Linear regression

Linear regression for non-linear f

Reading: Ch.

Prediction

inputs

outputs

- **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$
- **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- **Outputs** $y^{1:n}$

- **Goal** Learn/estimate $f(x)$ **predictor** for y

- By type of output
  - **Classification** if $y \in S_Y$ discrete
  - $y \in \{0, 1\}$ (or $\{\pm 1\}$) binary classification
  - $y \in \{1, 2, \ldots m\}$ multiclass classification

  - **Regression** if $y \in \mathbb{R}$ continuous

($x \in \mathbb{R}$ so far)

$x \in \mathbb{R}^d$ typical

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \begin{array}{l} \leftarrow \text{weight} \\ \leftarrow \text{height} \\ \leftarrow \text{age} \\ \leftarrow \text{allergies} \\ \ldots \end{array}$$

- inputs
- features
- attributes
- covariates
...

"d features"

Prediction

- **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$
- **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- **Outputs** $y^{1:n}$

- **Goal** Learn/estimate $f(x)$ **predictor** for y

- By type of output
  - **Classification** if $y \in S_Y$ discrete
  - $y \in \{0, 1\}$ (or $\{\pm 1\}$) binary classification
  - $y \in \{1, 2, \ldots m\}$ multiclass classification

  - **Regression** if $y \in \mathbb{R}$ continuous

- **Model family** $\mathcal{F} = \{f_\theta : \mathbb{R}^d \to S_Y, \theta \in \Theta\}$
  - $\mathcal{F} = \{$ linear functions $\}$ linear regression/classification
  - $\mathcal{F} = \{$ polynomes of degree 2, 3, ... $\}$ polynomial regression/classification
  - $\mathcal{F} = \{\frac{1}{1+e^{-\beta^T x + \beta_0}}\}$ logistic regression
  - neural network regression/classification
  - kernel regression/classification
  - $\mathcal{F} = \{$ monotonic functions $\}$ isotonic regression
  - support vector regression/classification
  - regression/classification *trees* (and random forests)