ORIE 4742 - Info Theory and Bayesian ML

Chapter 6: Intro to Bayesian Statistics

Februaru 20, 2020

Sid Banerjee, ORIE, Cornell

Bayesian basics

marginals and conditionals

let X and Y be discrete rvs taking values in $\mathbb{N}$. denote the joint pmf:

$$p_{XY}(x, y) = \mathbb{P}[X = x, Y = y]$$

marginalization: computing individual pmfs from joint pmfs as

$$p_X(x) = \sum_{y \in \mathbb{N}} p_{XY}(x, y) \qquad p_Y(y) = \sum_{x \in \mathbb{N}} p_{XY}(x, y)$$

conditioning: pmf of X given $Y = y$ (with $p_Y(y) > 0$) defined as:

$$\mathbb{P}[X = x | Y = y] \triangleq p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

more generally, can define $\mathbb{P}[X \in \mathcal{A} | Y \in \mathcal{B}]$ for sets $\mathcal{A}, \mathcal{B} \in \mathbb{N}$

see also this visual demonstration

the basic 'rules' of Bayesian inference

let X and Y be discrete rvs taking values in $\mathbb{N}$, with joint pmf $p(x, y)$

product rule

for $x, y \in \mathbb{N}$, we have: $p_{XY}(x, y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$

sum rule

for $x \in \mathbb{N}$, we have: $p_X(x) = \sum_{y \in \mathbb{N}} p_{X|Y}(x|y)p_Y(y)$

and most importantly!

Bayes rule

for any $x, y \in \mathbb{N}$, we have:

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{x \in \mathbb{N}} p_{Y|X}(y|x)p_X(x)}$$

see also this video for an intuitive take on Bayes rule

fundamental principle of Bayesian statistics

- assume the world arises via an underlying generative model $\mathcal{M}$
- use random variables to model all unknown parameters θ
- incorporate all that is known by conditioning on data D
- use Bayes rule to update prior beliefs into posterior beliefs

$$\underbrace{p(\theta|D,\mathcal{M})}_{\text{posterior}} \propto \underbrace{p(\theta|\mathcal{M})}_{\text{prior}}\underbrace{p(D|\theta,\mathcal{M})}_{\text{likelihood}}$$

- Physics - Newtonian dynamics, relativity
- <u>Note</u> - Bayesian ML DOES NOT believe the model parameters are random

pros and cons

in praise of Bayes

- conceptually simple and easy to interpret
- works well with small sample sizes and overparametrized models
- can handle all questions of interest: no need for different estimators, hypothesis testing, etc.

why isn't everybody Bayesian

- they need priors (subjectivity. . .)
- they may be more computationally expensive: computing normalization constant and expectations, and updating priors, may be difficult

basics of Bayesian inference

the likelihood principle

given model $\mathcal{M}$ with parameters Θ, and data D, we define:

- the prior $p(\Theta|\mathcal{M})$: what you believe before you see data
- the posterior $p(\Theta|D, \mathcal{M})$: what you believe after you see data
- the marginal likelihood or evidence $p(D|\mathcal{M})$: how probable is the data under our prior and model

  these three are probability distributions; the next is not

- the likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \overset{\Theta}{\theta})$: function of Θ summarizing data

the likelihood principle

given model $\mathcal{M}$, all evidence in data D relevant to parameters Θ is contained in the likelihood function $\mathcal{L}(\Theta)$

this is not without controversy; see Wikipedia article

REMEMBER THIS!!

given model $\mathcal{M}$ with parameters Θ, and data D, we define:

- the prior $p(\Theta|\mathcal{M})$: what you believe before you see data
- the posterior $p(\Theta|D, \mathcal{M})$: what you believe after you see data
- the marginal likelihood or evidence $p(D|\mathcal{M})$: how probable is the data under our prior and model
- the likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \Theta)$: function of Θ summarizing the data

the fundamental formula of Bayesian statistics

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$P(\theta|D) = \frac{P(D|\theta)\ P(\theta)}{P(D)}$$

also see: Sir David Spiegelhalter on Bayes vs. Fisher

Notes

- For discrete θ, $p(\theta|D)$, $p(\theta)$ are pmfs
  For continuous θ, $p(\theta|D) = f(\theta|D)$, $p(\theta) = f(\theta)$ (use pdfs)
- Similarly for discrete vs continuous data (for $p(D)$)
- Likelihood $p(D|\theta)$ is not a prob distn. It is a fn of θ that is parameterized by the data.
  - If D is continuous, use $f(D|\theta)$ - Note this is still some fn of θ.

example: the mystery Bernoulli rv

- data $D = \{X_1, X_2, \dots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

fix θ; what is $\mathbb{P}[\cancel{X_i}|\mathcal{M}]$ for any $i \in [n]$? $N_1 = \#$ of 1s , $N_0 = \#$ of 0s

D

$N_0 + N_1 = n$

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | \mathcal{M}, \theta] = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{N_1}(1-\theta)^{N_0}$$

$x_i \in \{0,1\}$

$\not\sim Bin(n,\theta)$

let $H = \#$ of '1's in $\{X_1, X_2, \dots, X_n\}$; what is $\mathbb{P}[H|\mathcal{M}, \theta]$?

$$\mathbb{P}[H = h | \mathcal{M}, \theta] = \binom{n}{h}\theta^h(1-\theta)^{n-h} \quad \sim Bin(n,\theta)$$

the Bernoulli likelihood function

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \theta)$: function of Θ summarizing the data

$$\boxed{\mathcal{L}(\theta) = \theta^{N_1} (1-\theta)^{N_0}}$$

Bernoulli: Likelihood Function

- Note - $\mathcal{L}(\theta)$ is NOT a distribution

(ie, $\int_0^1 \mathcal{L}(\theta) d\theta \neq 1$)

log-likelihood, sufficient statistics, MLE

- $\ell(\theta) = \log \mathcal{L}(\theta)$

$$\frac{d}{d\theta}\ell(\theta) = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}$$

$$\Rightarrow \theta^{MLE} = N_1 / N_1 + N_0$$

(For Bernoulli - $\ell(\theta) = \log\left(\theta^{N_1}(1-\theta)^{N_0}\right) = N_1 \log\theta + N_0 \log(1-\theta)$)

- (N_1, N_0) are sufficient statistics of D

(ie. $\mathcal{L}(\theta | D)$ = parametric fn of N_1 and N_0)

- MLE - $\underset{\theta\in[0,1]}{\arg\max}\ \mathcal{L}(\theta) = \underset{\theta\in[0,1]}{\arg\max}\ \ell(\theta) = \dfrac{N_1}{N_1+N_0} = \dfrac{N_1}{n}$

cromwell's rule

how should we choose the prior?

the zeroth rule of Bayesian statistics

never set $p(\theta|\mathcal{M}) = 0$ or $p(\theta|\mathcal{M}) = 1$ for any θ

- "I beseech you, in the bowels of Christ, think it possible that you may be mistaken." (Oliver Cromwell, 1650)
- Connected to philosophy of science (Falsifiability)

also see: Jacob Bronowski on Cromwell's Rule and the scientific method

from where do we get a prior?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

option 1: from the 'problem statement'

Mackay example 2.6

- eleven urns labeled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls
- urn u contains u red balls and $10-u$ blue balls
- select urn u uniformly at random and draw n balls with replacement, obtaining n_R red and $n-n_R$ blue balls

$$P(\theta) = \mathrm{Unif}\left\{\frac{0}{10}, \frac{1}{10}, \frac{2}{10}, \ldots, \frac{10}{10}\right\}$$

from where do we get a prior

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

option 2: the maximum entropy principle

choose $p(\theta|\mathcal{M})$ to be distribution with maximum entropy given $\mathcal{M}$

we know $\theta \in [0,1]$

- Maximum entropy prior on $[0,1] \equiv U[0,1]$

from where do we get the prior, take 2

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

option 3: easy updates via conjugate priors

- prior $p(\theta)$ is said to be conjugate to likelihood $p(D|\theta)$ if corresponding posterior $p(\theta|D)$ has same functional form as $p(\theta)$
- natural conjugate prior: $p(\theta)$ has same functional form as $p(D|\theta)$
- conjugate prior family: closed under Bayesian updating

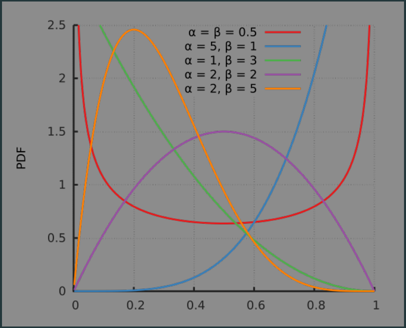
Note - The family of all distributions is trivially a conjugate prior ... we want more useful families

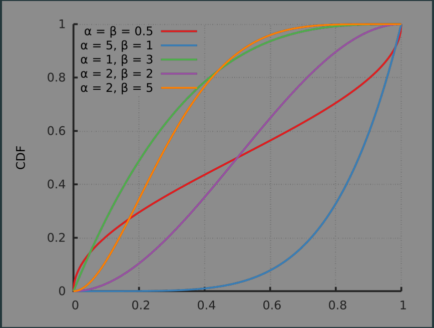
the Beta distribution

Beta distribution

- $x \in [0,1]$, parameters: $\Theta = (\alpha, \beta) \in \mathbb{R}^+$ ('# ones'+1,'# zeros'+1)
- pdf: $p(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$
- normalizing constant: $\frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ Gamma fn

Beta fn





Beta-Bernoulli prior and updates

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

Beta-Bernoulli model

- prior parameters: $\Theta_0 = (\alpha, \beta) \in \mathbb{R}^+$ (hyperparameters)
- Beta-Bernoulli prior: $Beta(\alpha, \beta) \sim p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- likelihood: $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$

  then via Bayesian update we get

- posterior:

$$p(\theta|D) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{N_1}(1-\theta)^{N_0} \sim Beta(\alpha+N_1, \beta+N_0)$$

the Beta distribution: getting familiar

$Beta(\alpha, \beta)$ distribution

$$p(x) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\frac{1}{B(\alpha,\beta)}} x^{\alpha-1}(1-x)^{\beta-1}$$

properties of $\Gamma(\alpha)$

$$\frac{1}{B(\alpha,\beta)} = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\,dy, \quad \boxed{\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)}$$

- If α is an integer. $\Gamma(\alpha) = (\alpha-1)!$

the Beta distribution: mean and mode

$Beta(\alpha, \beta)$ distribution

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- $$\mathbb{E}[X] = \int_0^1 x \, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cancel{\Gamma(\beta)}} \cdot \frac{\cancel{\Gamma(\beta)}\,\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}$$

Thus $\boxed{\text{mean of } Beta(\alpha,\beta) \text{ dist is } \frac{\alpha}{\alpha+\beta}}$

$$\text{mode} = \underset{\theta \in [0,1]}{\arg\max} \ \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

$$\frac{d}{dx}\left(x^{\alpha-1}(1-x)^{\beta-1}\right) = (\alpha-1)\, x^{\alpha-2}(1-x)^{\beta-1} - (\beta-1)\, x^{\alpha-1}(1-x)^{\beta-1} = 0$$

$$\Rightarrow \quad (\alpha-1)(1-x^*) = (\beta-1)\, x^*$$

$$\Rightarrow \quad x^* = \frac{\alpha-1}{\alpha+\beta-2} \quad (\text{for } \alpha>1,\ \alpha+\beta>2)$$

Thus $\boxed{\text{mode of Beta}(\alpha,\beta) \text{ dist is } \frac{\alpha-1}{\alpha+\beta-2}}$

Beta-Bernoulli model: what should we report?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

- Correct Answer - You should report Model, Prior, Posterior

- Decision theoretic answer - Ask for a loss fn, report θ which minimizes loss

decision theory

- Choose 'actions' to minimize a loss function (stats / ML)
  maximize a utility function (economics)

- Eg. Let θ be sample from posterior. Output $\hat{\theta}$ to minimize

1) $L(\theta,\hat{\theta}) = \mathbb{1}\{\theta \neq \hat{\theta}\}$ (L_0 loss) – $\hat{\theta}_{L_0}$ = mode of posterior distn
2) $L(\theta,\hat{\theta}) = |\theta - \hat{\theta}|$ (L_1 loss) – $\hat{\theta}_{L_1}$ = median of posterior distn
3) $L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$ (L_2 loss) – $\hat{\theta}_{L_2}$ = mean of posterior distn

In general, return $\arg\min_{\hat{\theta}} \mathbb{E}_{\theta \sim \text{posterior}} [\underbrace{L(\theta,\hat{\theta})}_{\text{loss fn}}]$

Beta-Bernoulli model: posterior mean

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior mean: $\mathbb{E}[\theta|\alpha, \beta, N_0, N_1] = \mathbb{E}\left[Beta(\alpha+N_1, \beta+N_0)\right]$

$$= \frac{\alpha+N_1}{\alpha+\beta+N_1+N_0} = \frac{\alpha+N_1}{m+n}$$

$$= \frac{\alpha}{m}\cdot\frac{m}{m+n} + \frac{N_1}{n}\cdot\frac{n}{m+n}$$

$$= \underbrace{w\cdot\frac{\alpha}{m}}_{\text{regularization}} + \underbrace{(1-w)}\cdot\underbrace{\frac{N_1}{n}}_{\text{'shrinkage' of MLE}}$$

Define $m = \alpha+\beta$

$n = N_1 + N_0$

$m \equiv$ 'number of prior samples'

$\frac{\alpha}{m} \equiv$ prior mean

$\frac{N_1}{n} \equiv$ data mean (also, MLE)

$w = \frac{m}{m+n} \equiv$ 'strength of prior' relative to data

Beta-Bernoulli model: posterior mode (MAP estimation)

maximum a posteriori

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_2)$

posterior mode: $\max_{\theta \in [0,1]} p(\theta|\alpha, \beta, N_0, N_1) = \dfrac{\alpha + N_1 - 1}{\alpha + \beta + N_1 + N_2 - 2}$

- If $\alpha = \beta = 1$ (ie, uniform prior)

then $\Theta_{MAP} = \dfrac{N_1}{N_1 + N_2} = \Theta_{MLE}$

In general, if prior is uniform, then $\Theta_{MLE} = \Theta_{MAP}$

Beta-Bernoulli model: posterior prediction (marginalization)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_2)$

posterior prediction: $\mathbb{P}[X = 1|D] = \int_0^1 p(\theta) \cdot \theta \cdot d\theta$

$$= \mathbb{E}[\theta] = \frac{\alpha + N_1}{\alpha + \beta + N_1 + N_2}$$

If $\alpha = \beta = 1$, $\boxed{\mathbb{P}[X = 1 | D] = \frac{N_1 + 1}{N_1 + N_2 + 2}}$ Laplace Estimator

(or 'add-one' smoothing)

the black swan

- If we observe $N_0 = n$, then what is $\mathbb{P}[X_{n+1}=1]$?
  - MLE $\equiv$ $\mathbb{P}_{MLE}[X_{n+1}=1] = 0$, $\mathbb{P}_{MLE}[X_{n+1}=0] = 1$
  - Laplace (ie, Bayesian update with Beta(1,1) prior)

$$\mathbb{P}_{Lap}[X_{n+1}=1] = \frac{1}{n+2}, \quad \mathbb{P}_{Lap}[X_{n+1}=0] = \frac{n+1}{n+2}$$

more '0's that we see, less unlikely the arrival of a '1'

however, not <u>impossible</u>! remember Cromwell's law