till now → from today!

ORIE 4742 - Info Theory and Bayesian ML

Chapter 6: Intro to Bayesian Statistics

March 15, 2021

Sid Banerjee, ORIE, Cornell

Main Idea - Assume θ is drawn from a distribution

World → data (depends on θ) → 'learn' θ

unknown parameters θ

'model' for what is unknown

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:

joint

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

marginal

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 $h(x,y) = h(x) + h(y|x)$, $H(X,Y) = H(Y) + H(X|Y)$

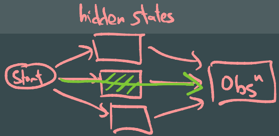
for $x, y \in \mathbb{N}$, we have: $p_{XY}(x,y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$

sum rule $H(X) = H(Y) + \sum_y p(y) H(X|Y=y)$

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

probabilistic

- 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

$$p(\theta|D,\mathcal{M}) \propto p(\theta|\mathcal{M})p(D|\theta,\mathcal{M})$$

Note : Bayesian ML DOES NOT believe that the Θ are random

- This is only for 'convenience'

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 $\boxed{\text{priors}}$ (subjectivity. . .) (but all methods are subjective...)
- they may be more computationally expensive: computing normalization constant and expectations, and updating priors, may be difficult

- Eg - MCMC (however - anytime you use a Bayesian ML method, you get much more info)

basics of Bayesian inference

(often hidden)

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

both are distribn over Θ

- the marginal likelihood or evidence $p(D|\mathcal{M})$: how probable is the data under our prior and model

not distrn

~~these three are probability distributions; the next is not~~

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

the likelihood principle (main axiomatic basis for Bayesian ML)

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

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

$$p(\theta|D,M) = \frac{p(D|\theta,M)\, p(\theta|M)}{p(D|M)}$$

also see: Sir David Spiegelhalter on Bayes vs. Fisher

Most often (>90%)

$$P(\theta|D,M) \propto p(D|\theta,M)\, p(\theta|M)$$

Notes

- Likelihood, evidence are not distributions ($\mathcal{L}(\theta)$ is just a fn of θ which summarizes the data)

$p(\theta)$, $p(\theta|D)$ are distributions over Θ

- $\mathcal{L}(\theta)$ is different for discrete vs continuous parameters Θ
  - If θ discrete, $\mathcal{L}(\theta|D) = p(\theta|D)$ (pmf)
    If θ contin, $\mathcal{L}(\theta|D) = f(\theta|D)$ (pdf)
- The evidence is different for discrete vs continuous D
  - If θ discrete, evidence $= p(D|M)$
    θ continuous, evidence $= f(D|M)$

example: the mystery Bernoulli rv

- 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

ie, $\mathbb{P}[X=1] = \theta$

fix θ; what is $\underbrace{\mathbb{P}[D|\mathcal{M}, \theta]}_{\lambda}$ for any $i \in [n]$?

Let $N_1 = \#$ of 1s in D
$N_0 = \#$ of 0s in D
$N_1 + N_0 = n$

$$\mathcal{L}(\theta) = P(D|M,\theta) = \mathbb{P}[X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n | M, \theta] = \theta^{N_1}(1-\theta)^{N_0}$$

$$\left(= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)} \right)$$

let N_1 = # of '1's in $\{X_1, X_2, \ldots, X_n\}$; what is $\mathbb{P}[H|\mathcal{M}, \theta]$?

$$\mathbb{P}[N_1 = k | \theta, M] = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

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

$$\boxed{\mathcal{L}(\theta) = \theta^{N_1}(1-\theta)^{N_0}} \qquad \theta \in [0,1]$$

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

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

log-likelihood, sufficient statistics, MLE

- $\ell(\theta) = \log \mathcal{L}(\theta) = N_1 \log \theta + N_0 \log(1-\theta)$

  for Bernoulli

  - (N_1, N_0) are sufficient statistics

    i.e., Given data D, $\mathcal{L}(\theta)$ completely determined by

    $N_1(D)$, $N_0(D)$

- MLE - max likelihood estimator

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

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

- Oliver Cromwell - 'I beseech you, < supplication to higher authority >, think it possible that you might be mistaken'

- Connected to falsifiability

also see:

– Jacob Bronowski on Cromwell's Rule and the scientific method

– Richard Feynman on the scientific method (at Cornell!)

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

$\theta = \frac{i}{10}$ with prob $\frac{1}{11}$ for each $i \in \{0, 1, \ldots, 10\}$

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 $\theta \in [0,1]$

Eg - If we know $\theta \in [0,1]$, then one

choice of prior $\equiv$ MaxEnt$([0,1])$

$= \text{Unif}([0,1])$

Eg - If $\theta \in \mathbb{N}_+$, $E[\theta] = \mu$ (≥ 1) $\Rightarrow$ $\text{Geom}(1/\mu)$

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 - One obvious family ≡ set of all distributions
(not useful ...)

Want - 'smallest' family which is closed under Bayesian updates

the Beta distribution

Beta distribution

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