ORIE 4742 - Info Theory and Bayesian ML

Bayesian ML: Revision of Basics

April 7, 2020

Sid Banerjee, ORIE, Cornell

Bayesian basics

unknown ← represented as random variable

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

$\Theta \to D \to \Theta$

prior posterior

- 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 (Bayes rule)

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\underbrace{\text{evidence}}_{\text{normalization}}} \propto \text{likelihood} \times \text{prior}$$

Bayesian statistics: three 'laws'

likelihood principle

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

- Eg - 2 expts (want to learn a $\underbrace{Ber(p_i)}_{\mathcal{M}}$ dist)

Binomial $Bin(9, p_i)$ [Expt 1 - Generate 9 iid samples $X_1, X_2, \ldots, X_9$, observe 1 success

Geometric $Geom(p_i)$ [Expt 2 - Wait till 1st success, need 9 trials

Both have the same likelihood

$\Rightarrow$ Both have same posterior (for any prior on p_i)

Bayesian statistics: three 'laws'

likelihood principle

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

Cromwell's rule

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

How to choose prior

Bayesian statistics: three 'laws'

likelihood principle

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

Cromwell's rule

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

choosing priors

- 'principled' choice: maximum entropy, 'objective' priors (Jeffreys prior)
- 'computational' choice: conjugate priors
  - prior $p(\theta)$ is 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 $p(D|\theta)$

fns of Θ

marginal likelihood (model evidence)

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

marginal likelihood

$$p(D|\mathcal{M}) = \frac{p(\theta)p(D|\theta)}{p(\theta|D)} = \frac{\text{prior} \times \text{likelihood}}{\text{posterior}}$$

- $p(D|M) \equiv$ probability of seeing D under model M (under the prior)

summarizing the posterior

model $\mathcal{M}+$ prior $p(\Theta)+$ data $D \Rightarrow$ posterior $p(\Theta|D)$

summarizing $p(\Theta|D)$

- posterior mean $\widehat{\theta}_{mean} = \mathbb{E}[\Theta|D]$
- posterior mode (or MAP estimate) $\widehat{\theta}_{MAP} = \arg\max_{\Theta} p(\Theta|D)$
- posterior median $\widehat{\theta}_{median} = \min\{\Theta : p(\Theta|D) \geq 0.5\}$
- Bayesian credible intervals: given $\delta > 0$, want (ℓ_Θ, u_Θ) s.t.

$$\mathbb{P}[\ell_\Theta \leq \Theta \leq u_\Theta | D] > 1 - \delta$$

- Ideal - Report posterior
- Marginalization, ie., Sample from posterior

decision theory

given posterior $p(\Theta|D)$ and loss function $L(\Theta, a)$

decision theoretic estimate $\Theta^\star$

choose 'action/estimate' $\Theta^\star$ to minimize expected loss under posterior

$$\hat{\theta}^* = \arg\min_a \mathbb{E}_{\Theta\sim p(\Theta|D)}\left[L(\Theta, a)\right]$$

example loss functions

- L_0 loss: $L(\Theta, a) = \mathbb{1}_{\{\Theta\neq a\}} \Rightarrow \Theta^\star = \widehat{\theta}_{mode}$
- L_1 loss: $L(\Theta, a) = |\Theta - a| \Rightarrow \Theta^\star = \widehat{\theta}_{median}$
- L_2 loss: $L(\Theta, a) = (\Theta - a)^2 \Rightarrow \Theta^\star = \widehat{\theta}_{mean}$

Important point - Bayesian update does not care about loss fn

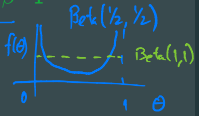
binary data and Beta-Bernoulli prior

- 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: $p(D|\theta) = x^{N_1}(1-x)^{N_0}$
- posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$
- marginal likelihood: let $m = \alpha + \beta$

$$p(D) = \frac{\Gamma(m)}{\Gamma(n+m)} \frac{\Gamma(N_1+\alpha)}{\Gamma(\alpha)} \frac{\Gamma(N_0+\beta)}{\Gamma(\beta)}$$



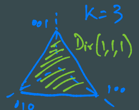
multiclass data and Dirichlet priors

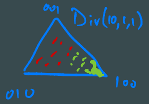
- for $i \in [K]$, data D contains N_i copies of type i
- model $\mathcal{M}$: X_i generated i.i.d. from $Multinomial(\theta_1, \theta_2, \ldots, \theta_K)$ distn

Dirichlet-Multinomial model

- prior parameters: $\Theta_0 = (\alpha_1, \alpha_2, \ldots, \alpha_K) \in \mathbb{R}_+^K$ (hyperparameters)
- Dirichlet prior: $Dir(\alpha_1, \alpha_2, \ldots, \alpha_K) \sim p(\theta) \propto \prod_{i=1}^K \theta_i^{\alpha_i - 1}$
- likelihood: $p(D|\theta) = \prod_{i=1}^K \theta_i^{N_i}$
- posterior: $p(\theta|D) \sim Dir(\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$
- marginal likelihood: let $m = \sum_{i=1}^K \alpha_i$

$$p(D) = \frac{\Gamma(m)}{\Gamma(n+m)} \prod_{i=1}^K \frac{\Gamma(N_i + \alpha_i)}{\Gamma(\alpha_i)}$$





normal-normal model for unknown μ

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown μ, known $\tau = 1/\sigma^2$

Precision

normal-normal model

- likelihood: $p(D|\mu, \tau) \propto \exp\left(-\tau \sum_{i=1}^n (x_i - \mu)^2/2\right)$ $\left(\prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_i-\mu)^2}\right)$
- prior: $\mu \sim \mathcal{N}(m_\mu, 1/\tau_\mu)$ (hyperparameters - m_μ, τ_μ) $\tau_\mu \to \tau_\mu + \frac{n}{2}\tau$
- posterior: let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $m_D = \frac{n\tau \cdot \bar{x} + \tau_\mu \cdot m_\mu}{n\tau + \tau_\mu}$ and $\tau_D = n\tau + \tau_\mu$ $\tau_\mu + n\tau$

$\frac{\tau_\mu}{\tau_\mu + n\tau} \cdot m_\mu + \frac{n\tau}{\tau_\mu + n\tau} \bar{x}$

$$p(\mu|D) \sim \mathcal{N}\left(m_D, 1/\tau_D\right)$$

$n\tau$

m_μ $\bar{x}$

- posterior predictive distribution:

$$p(x|D) \sim \mathcal{N}\left(m_D, 1/\tau + 1/\tau_D\right)$$

normal-gamma model for unknown τ

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, with unknown $\tau = 1/\sigma^2$, known μ

normal-gamma model

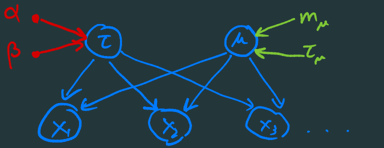
- likelihood: $p(D|\mu, \tau) \propto \left[\exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2 / 2\right)\right] \cdot \tau^{n/2}$
- prior for τ: $\tau \sim gamma(\alpha, \beta)$ hyper parameters: (α, β)
- posterior: let $\alpha_D = \alpha + \frac{n}{2}$ and $\beta_D = \beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$

$$p(\tau|D) \sim gamma\left(\alpha_D, \beta_D\right)$$

- posterior predictive distribution:

$$p(x|D) \sim \text{student-t}$$

- μ and τ unknown



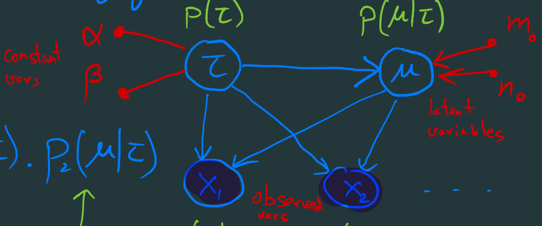
'Bayesian network' of model $\mathcal{M}$

- <u>Idea 1</u> - Prior $P(\mu, \tau) = P_1(\mu) \cdot P_2(\tau)$

<u>Problem</u> - Conditioned on X_1, μ and τ are <u>not</u> independent

$\Rightarrow$ prior is not a conjugate prior

- <u>Idea 2</u>



Prior - $p(\mu, \tau) = p_1(\tau) \cdot p_2(\mu|\tau)$

Gamma(α, β) $\qquad$ $\mathcal{N}(m_0, n_0\tau)$ $\quad$ $P(x_1|\mu, \tau) = P(x_2|\mu|\tau) = \ldots$

normal-(normal-gamma) model for unknown (μ, τ)

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, 1/\tau)$, unknown $\tau = 1/\sigma^2$, unknown μ

normal-(normal-gamma) model

likelihood Prior on μ Prior on τ

- likelihood: $D|\mu, \tau \sim \mathcal{N}(\mu, 1/\tau)$
- prior for (μ, τ): $\tau \sim gamma(\alpha, \beta)$ and $\mu|\tau \sim \mathcal{N}(m_0, 1/n_0\tau)$
- posterior: let $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $m_D = \frac{n\tau \cdot \overline{x} + n_0\tau \cdot m_\mu}{n\tau + n_0\tau}$ and $\tau_D = n\tau + n_0\tau$

$$p(\mu|\tau, D) \sim \mathcal{N}(m_D, \tau_D)$$

also let $\alpha_D = \alpha + \frac{n}{2}$, $\beta_D = \beta + \frac{1}{2}\sum_{i=1}^n (x_i - \overline{x})^2 + \frac{nn_0}{2(n+n_0)}(\overline{x} - m_0)^2$

$$p(\tau|D) \sim gamma(\alpha_D, \beta_D)$$

- posterior predictive distribution: $p(x|D) \sim$ student-t

naive Bayes classifier

- Collection of 'data sets' with 'class labels' (c, D)
- class label $c \in \{1, 2, \ldots, c\}$ (Eg- {spam, regular, imp})
- $D \sim P(\cdot \mid \theta_c)$
- Eg- $D \equiv$ bag of words (dictionary $\{1, \ldots, d\}$)
  - DNA, dict = {A, T, C, G}, dict = {triplets of bases}
  - document, dict = {words in language}
- Assumption $D \sim \mathrm{Dir}(\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cd})$

naive Bayes classifier

- $(\alpha_1, \alpha_2, \dots, \alpha_c)$
- $\forall c \in 1, \dots, C$ $(\alpha_{c1}, \alpha_{c2}, \dots, \alpha_{cd})$

$\xrightarrow[= (1, D_1)]{(c_1, D_1)}$

- $(\alpha_1 + 1, \alpha_2, \dots, \alpha_c)$
- $(\alpha_{11} + N_1, \dots, \alpha_{1d} + N_d)$

rest is same

$\xrightarrow{(c_2, D_2)}$ $\xrightarrow{(c_3, D_3)}$

- (α_s, α_r)

$(\alpha_{s1}, \alpha_{s2}, \dots, \alpha_{sd})$ $(\alpha_{r1}, \alpha_{r2}, \dots, \alpha_{rd})$

- Inference –

$P(D \in \text{spam}) = P(s)\, P(D \mid s)$

$P(D \in \text{reg}) = P(r)\, P(D \mid r)$

pick larger of the two

plan for remaining semester

- Bayesian networks (Graphical models)
  - directed Bayes nets
  - Markov random fields
  - Causal inference
- Approximate Bayesian update - use simulation to get approx posterior
  - Markov Chain Monte Carlo
- Gaussian processes - regression
- Mixture models - clustering, EM algorithm
- Sequential Decision theory - Bandits, MDP