Bayesian methods

- prob. theory
- data
- statistics

Bayesian approach

- model selection

[frequentist] & Bayesian

- bag of tools
- subjective Bayesian

- use probability everywhere when there is uncertainty

- focus on $p(\theta | \text{data})$ or $p(\text{data} | \theta) p(\theta)$
  - posterior
  - likelihood
  - prior

Caricature:

Bayesian is "optimist": she thinks you can get good models

$\Rightarrow$ obtain a method by doing probabilistic inference in model

Frequentist is "pessimist": use analysis tools

Example: biased coin

- Bayesian model: $X_i \in \mathbb{S}_{0,1}$
- Hypothesis: $X_i \sim \text{Bernoulli} (\theta)$
- $p(x_i | \theta) = \theta^{x_i} (1-\theta)^{1-x_i}$
- $\theta \sim \text{Uniform} [0, 1]$
- $x_i \sim \text{Uniform} [0, 1]$
posterior: \[ p(\theta|x,n) \propto \left( \prod_{i=1}^{n} p(x_i|\theta) \right) p(\theta) \]

\[ = \Theta \left( \frac{\sum x_i}{n} \right)^{n-\frac{2}{\alpha}} \left( 1 - \frac{\sum x_i}{n} \right)^{n-\frac{2}{\beta}} I_{(\alpha,\beta)}(\theta) \]

\[ \sim Beta(\theta|\alpha,\beta) \text{ where } \alpha = n_2 + 1 \]
\[ \beta = n - n_2 + 1 \]

Note: \( p(\epsilon) = Beta(\epsilon|\alpha_0,\beta_0) \)

Here \( p(\epsilon|dab) = Beta(\epsilon|\alpha_2 + \alpha_0, n_2 + \beta_0) \)

Here Beta(\epsilon|\alpha_0,\beta_0) is a "conjugate prior" to the Bernoulli likelihood model.

More generally, consider family \( F \) of dist. \( F = \frac{1}{2} p(\theta|x,\alpha) \) : \( \alpha \in \mathcal{R}^+ \)

Say that \( F \) is a "conjugate family" to observation model \( p(x|\epsilon) \)

If posterior \( p(\epsilon|x,\alpha) \in F \)

i.e. \( \exists \alpha^* \text{ s.t. } p(\epsilon|x,\alpha) = p(\epsilon|\alpha^*) \)

\[ \int_{\mathcal{R}^+} \frac{p(x|\epsilon)p(\epsilon|\alpha)}{\int_{\mathcal{R}^+} p(x|\epsilon)p(\epsilon|\alpha^*) d\epsilon} d\epsilon \]
Teaching Page 3

- Denote if use conjugate priors in a DGM,

then Gibbs sampling can be easy

- Bayesian in action?

question: What is the prior that next flip = 1?

\[ \text{Frequentist: } \Theta_{\text{ML}} = \frac{n_1}{n} \]

\[ \text{Bayesian: integrate out uncertainty} \]

\[ \text{using code, under} \]

\[ p(x_{n+1} \mid x_{1:n}) = \int \theta p(x_{n+1} \mid \theta) p(\theta \mid x_{1:n}) \, d\theta \]

\[ \text{predictive dist} \]

\[ p(x_{n+1} = 1 \mid x_{1:n}) = \int \theta p(\theta \mid x_{1:n}) \, d\theta \rightarrow \text{posterior mean?} \]

\[ \text{ESS | data} \frac{\alpha}{\alpha + \beta} = \frac{n_1 + 1}{n + 2} \quad (\text{with } \alpha_0 = \beta_0 = 1 \text{ i.e. uniform prior}) \]

\[ \Theta_{\text{posterior mean}} = \left( \frac{\theta_0}{2} \right) + \left( \frac{1}{2 n} \right) \]

\[ \Theta_{\text{ML}} \frac{n_1}{n} \Theta_{\text{prior}} 1 - \frac{n_1}{n} \]

\[ n \rightarrow \infty \]

\[ \frac{n_1}{n} \rightarrow 1 \quad \text{i.e. } \Theta_{\text{posterior}} \rightarrow \Theta_{\text{ML}} = \text{true } \theta \text{ mean} \]

\[ n \rightarrow 0 \text{ as } n \rightarrow \infty \]
Variance: \[ \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = \frac{1}{n} \left( 1 - \frac{\alpha}{n} \right) \sim O \left( \frac{1}{n} \right) \]
\[ = \beta \left( 1 - \beta \mu \right) \sim O \left( \frac{1}{n} \right) \]

"Posterior "contracts" around \( \mu \text{mean} \) \( \overset{\text{as}}{\rightarrow} \) \( \beta \mu \sim O \left( \frac{1}{n} \right) \)

"Bernstein von-Mises Thm"

\[ \Rightarrow \text{"Bayesian CLT"} : \text{basically say that if prior put non-zero mass on true parameter } \Theta \xrightarrow{\text{as } n \to \infty} (i.e. } x_i \sim p(x|\Theta)) \]

then posterior concentrates around \( \Theta \) as a Gaussian asymptotically

Recall from last lecture: multivariate model

\( X \sim \text{Mult}(\Theta) \) where \( \Theta \in \Delta_k \)

\[ \hat{\Theta}_{MLE} = \left( \frac{n \hat{\theta}_i}{n} \right)_{i=1}^k \]

Putting Dirichlet prior over \( \Theta \) \( \Theta \sim \text{Dir}(\alpha) \)

we get Dirichlet posterior \( \Theta | \text{data } \sim \text{Dir}(\alpha + \text{data}_i) \)

thus Dirichlet is conjugate prior to Multinoulli likelihood

\[ \text{Model selection} \]

say we want to choose between 2 DGMs

\[ M_1 \]

\[ M_2 \]
(note here that "M_1 < M_2")

as a frequentist:

\[
\hat{\Theta}_{M_2} = \text{argmax}_{\Theta} \log p(\text{data} | \Theta, \Sigma_1, \Sigma_2, \text{model} = M_2)
\]

\[
\hat{\Theta}_{M_1} = \text{argmax}_{\Theta} \log p(\text{data} | \Theta, \Sigma_1, \Sigma_2, \text{model} = M_1)
\]

how to choose between models?

can't just compare \( \log p(\text{data} | \hat{\Theta}_{M_2}, M=M_2) \) vs. \( \log p(\text{data} | \hat{\Theta}_{M_1}, M=M_2) \) because LHS < RHS since \( M_1 < M_2 \)

(i.e. you would always choose "bigger model")

\[\rightarrow \text{as frequentist, use cross-validation i.e. } \log p(\text{test data} | \hat{\Theta}_{M_2}(\text{train}) , M=M_2)\]

Bayesian alternative:

true Bayesian \( \rightarrow \) sum over models (integrate out uncertainty)

introduce prior over models \( p(M) \)

\[
P(x_{\text{new}} | D) = \sum_{M} p(x_{\text{new}} | D, M) p(M|D)
\]

\[
= \sum_{M} \left( \sum_{\Theta \in \Theta_{M}} p(x_{\text{new}} | \Theta, M) p(\Theta | D, M) p(M|D) d\Theta \right) p(M|D) p(M) p(0|M)
\]

Teaching Page 5
\[ p(\text{new data}) = \sum_M p(M|D) \left[ \int_{\Theta|D,M} p(\text{new} | \Theta, M) p(\Theta|D,M) d\Theta \right] \]

\[ p_{\text{model averaging}} \]

\[ p(\text{new} | M, D) \]

\[ \text{in model selection, forced to pick one model} \]

\[ \implies \text{pick model that maximizes } p(M|\text{data}) \propto p(\text{data}|M) p(M) \]

\[ \sum_{\Theta|D,M} p(\text{data}|\Theta, M) p(\Theta|D,M) d\Theta \]

\[ p(\text{data}|M) = \text{"marginal likelihood"} \]

\[ \frac{p(M=1|D)}{p(M=2|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_2)p(M_2)} \]

Favoring factor

"Uniform prior over models"; then we can pick among K models \( M_1, \ldots, M_k \)
key maximizing \( p(\text{data}|M=M_i) \)

"empirical Bayes"

"type II ML"

when \( H \) of models is "small", then this approach is fine (i.e. won't overfit)

Zoubin's cartoon: suppose \( M_1 \leq M_2 \leq M_3 \)

\[
p(D|M) = \sum_{M_i} p(D|M_i) p(M_i|\theta) p(\theta)
\]

\( p(D|M) \) is normalized over \( D \)

\( p(D|\theta, M_i) \)

Type II ML can still overfit when have many models

say e.g. \( p(D|M) = \sum_{\substack{M_2 M_1 \ldots \\theta \in \theta \ast \\alpha \ast \omega}} \)

\( S(D, M) = \frac{1}{\omega} \sum_{\theta \in \theta} \)

\( \int_{\omega} \)

how to compute marginal likelihood:

use approximations — variational inference sampling