Bayesian methods

From: theory $p(x|a)$

model

data

$\Theta$

statistics

"subjective Bayesian" -> use probability everywhere, there is uncertainty

"frequentist" = bag of tools $\Theta$

\[ \begin{align*}
\text{ML} & \quad \text{regularized ML} \\
\text{max. entropy} & \quad \text{moment matching}
\end{align*} \]

ERM

focus on $p(\Theta|\text{data})$ or $p(\text{data}|\Theta) p(\Theta)$

posterior likelihood prior

cartoon: Bayesian is "optimist" - they think you can get "good" models

$\Rightarrow$ obtain a method by doing proba. inference in model

frequentist is "pessimist" $\Rightarrow$ use analysis tools

Example: biased coin
Example: biased coin:

Bayesian model \( x_i \in \{0, 1\} \)

Hyperparameter for the prior

\( \theta \sim \text{Uniform}[0, 1] = \text{Beta}(1, 1) \)

\[ x_i | \theta \sim \text{Bernoulli}(\theta) \]

\[ p(x_i | \theta) = \theta^{x_i} (1 - \theta)^{1 - x_i} \]

Posterior:

\[ p(\theta | x_1:n) \propto \left( \prod_{i=1}^{n} p(x_i | \theta) \right) p(\theta) \]

\[ = \frac{\prod_{i=1}^{n} x_i \theta^{x_i} (1 - \theta)^{1 - x_i}}{n!} \text{Beta}(\alpha_0 + \sum_{i=1}^{n} x_i, n_0 + n - \sum_{i=1}^{n} x_i) \]

Note: if \( p(\theta) = \text{Beta}(\theta | \alpha_0, \beta_0) \)

\[ \Rightarrow p(\theta | \text{data}) = \text{Beta}(\theta | n_0 + \sum_{i=1}^{n} x_i, n - n_0 + \beta_0) \]

\( \Rightarrow \) "conjugate prior" to the Bernoulli likelihood model

More generally:

Consider a family \( F \) of dist. \( F = \{ p(\theta | \alpha) : \alpha \in \mathcal{A} \} \)

Say that \( F \) is a "conjugate family" to observation model \( p(x | \theta) \)
If posterior \( p(\theta | x, \alpha) \in F \) for any \( x \sim x_{1:n} \)

i.e. \( \exists \alpha \) s.t. \( p(\theta | x, \alpha) = p(\theta | \alpha) \)

Side note: if use conjugate priors in a DCM then Gibbs sampling can be easy

[\text{e.g. this is case in LDA topic model}]

∗ Bayesian in action:

question: what is the prob that next-flip = 1?  

\[
E[\theta | \text{data}] = \frac{\alpha}{\alpha + \beta} = \frac{n_{1} + 1}{n_{1} + n_{1} + n_{1} + 1} = \frac{n_{1} + 1}{n + 2}  
\]

[with \( \alpha_{0} = \beta_{0} = 1 \) and uniform prior]
\[ F(\theta|\text{data}) = \frac{\alpha}{\alpha + \beta} = \frac{n+1}{n+1+n-\alpha+1} = \frac{n+1}{n+2} \quad \text{[with } \alpha_0 = \beta_0 = 1 \text{]} \]

\[ \widehat{\text{Posterior mean}} = \frac{n_1}{n} \left[ \frac{\alpha}{n + 2} \right] + 1 \left[ \frac{\beta}{n + 2} \right] \]

\[ \widehat{\text{EML}} \quad \text{[prior]} \quad \widehat{\text{Prior mean}} \]

\[ P_n \to 1 \quad \therefore \quad \widehat{\text{Posterior}} \xrightarrow{n \to \infty} \widehat{\text{EML}} = \text{"true } \theta \text{"} = 0^* \]

Variance of a Beta:
\[ \frac{\alpha \beta}{(\alpha + \beta)^2(n + \alpha + \beta + 1)} = \text{EML} (1 - \text{EML}) O(n) \quad \xrightarrow{n \to \infty} 0 \]

Posterior "contracts" around \( \widehat{\text{Posterior mean}} \xrightarrow{n \to \infty} \text{EML} = 0^* \)

"Bernstein von-Mises thm."

"Bayesian CLT"; basically says that if prior put non-zero mass on true parameter \( \theta^* \) [i.e. \( x \sim \mathcal{N}(2, \theta^*) \)]

then posterior concentrates around \( \theta^* \) as a Gaussian asymptotically

recall from hwk 1: multivariate model
Pitfall: dist. is conjugate to multivariate model

Model selection:

say we want to choose between 2 DOM

\[ p(x_2 \mid x_1, \theta) \]

\[ p(x_3 \mid x_1, x_2, \bar{\theta}) \]

(note here that "M_1 \subseteq M_2")

as a frequentist:

\[ \hat{\Theta}_{M_1}^{ML} = \arg \max_{\Theta_1, \Theta_2} \log p(\text{data} \mid \Theta_0, \Theta_1, \Theta_2, \text{"model=M_1"}) \]

\[ \hat{\Theta}_{M_2}^{ML} = \arg \max_{\Theta_0, \Theta_1, \Theta_2} \log p(\text{data} \mid \Theta_0, \Theta_1, \Theta_2, \text{"model=M_2"}) \]

how to choose between models?

can't compare \( \log p(\text{data} \mid \hat{\Theta}_{M_1}, M=M_1) \) vs. \( \log p(\text{data} \mid \hat{\Theta}_{M_2}, M=M_2) \)

because LHS \( \leq \) RHS since \( M_1 \subseteq M_2 \)
Bayesian alternatives:

True Bayesian → Sum over models (integrate out uncertainty)

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

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

\[
= \sum_M \left( \sum_{\theta \in \Theta(M)} p(x_{\text{new}} | \theta, M) p(\theta | D, M) \right) \pi(M | D)
\]

\[
= \sum_M \pi(M | D) \left( \sum_{\theta \in \Theta(M)} p(x_{\text{new}} | \theta, M) p(\theta | D, M) \right)
\]

\[\text{[note: } p(x_{\text{new}} | \theta, M, D) = p(x_{\text{new}} | \theta, M) \text{]}\]

Closing model-averaging

Standard predictive dist. for one model

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

\[\star \text{ in model selection, forced to pick one model}\]
\[
p(M \mid \text{data}) \propto p(\text{data} \mid M) p(M)
\]

to compare two models, look at
\[
\frac{p(M = M_1 \mid \text{data})}{p(M = M_2 \mid \text{data})} = \frac{p(D \mid M_1) p(M_1)}{p(D \mid M_2) p(M_2)}
\]

Bayes factor

"uniform prior over models"; then we can pick among \(k\) models \(M_1, \ldots, M_k\) by maximizing \(p(\text{data} \mid M = M_i)\)

"empirical Bayes"

"type II ML"

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

Zoubin's cartoon: suppose \(M_1 \subseteq M_3 \subseteq M_3\)

\[
p(D \mid M) \text{ is normalized over } D
\]

\(\text{vs.} \)
Bayesian information criterion

\[ \text{BIC is a (rough) approximation of } \log p(\text{data}|M) \approx \log p(\text{data}|\hat{M}, M) \]  
-
\[ \frac{\text{dim}(\Theta_M)}{2} \]

- complexity penalty

\[ \text{use Laplace approximation} \]

\[ p(D|M) = \left\{ \int_{\Theta} \prod_{i=1}^{n} p(x_i|\theta, M) p(\theta|M) \right\} dt \]

\[ \exp \left( -n h(\theta) \right) \]

where
\[ h(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\hat{M}) + \log p(\hat{M}) \]

\[ \text{do Taylor expansion of this around } \hat{M} \]

\[ p(D|\hat{M}, M) \]  
- [can overfit]

Type II ML can still overfit when there are many models:

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

say e.g.

\[ p(D|M) = \sum_{M_1, M_2, M_3, \ldots} \]

how to compute marginal likelihood:

- use approximations
  - variational inference
  - sampling

Bayesian information criterion

\[ \text{BIC is a (rough) approximation of } \log p(\text{data}|M) \approx \log p(\text{data}|\hat{M}, M) \]  
-
\[ \frac{\text{dim}(\Theta_M)}{2} \]

- complexity penalty

\[ \text{use Laplace approximation} \]

\[ p(D|M) = \left\{ \int_{\Theta} \prod_{i=1}^{n} p(x_i|\theta, M) p(\theta|M) \right\} dt \]

\[ \exp \left( -n h(\theta) \right) \]

where
\[ h(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\hat{M}) + \log p(\hat{M}) \]

\[ \text{do Taylor expansion of this around } \hat{M} \]
Do Taylor expansion of this around \( \hat{\eta}_n \)

2 approximations: keep only terms which grow with \( n \) replace \( \hat{\eta}_n \) by \( \hat{\xi}_n \) get BIC

BIC is "consistent"

Gaussian networks:

\[ X \sim N(\mu, \Sigma) \quad \mu \in \mathbb{R}^p \quad \Sigma \in \mathbb{R}^{p \times p} \quad \Sigma > 0 \]

\[
p(x; \mu, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)
\]

\[
(x-\mu)^T \Sigma^{-1} (x-\mu) = x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x + \mu^T \Sigma^{-1} \mu
\]

\[
\text{tr}(x^T \Sigma^{-1} x) = \text{tr}(\Sigma^{-1} x x^T)
\]

\[
= \Sigma^{-1} \langle x x^T \rangle
\]

Canonical parametrization

\[
\Lambda \triangleq \Sigma^{-1}
\]

Linear form on matrices

\[
-2 \langle \Sigma^{-1} \mu, x \rangle
\]

\[
\sim 11 - \delta_{0}
\]

IFT6269 Page 9
\[ \Lambda = \Sigma^{-1} \]

precision matrix

sufficient statistics \[ T(x) = \begin{pmatrix} x \\ \frac{1}{2} xx^T \end{pmatrix} \]

canonical parameter \[ \hat{\eta}(x) = \begin{pmatrix} \eta \\ \Lambda \end{pmatrix} = \begin{pmatrix} \Sigma^{-1} \\ \Sigma^{-1} \end{pmatrix} \]

\[ p(x; \eta, \Lambda) = \exp(\eta^T x + \frac{1}{2} \eta \Lambda^{-1} \eta) \times \exp\left(\frac{1}{2} \log \det \Lambda + \frac{1}{2} \log |N|\right) \]

\[ \mathcal{D} = \{ (\eta, \Lambda) : \eta \in \mathbb{R}^p, \Lambda > 0, \quad \Lambda \in \mathbb{R}^{p \times p}, \quad \Lambda = \Lambda^T \} \]

useful exercise:\n\[ \nabla_\eta A(\eta, \Lambda) = E[x] = \mu = \Lambda^{-1} \eta \]
\[ \nabla_\Lambda A(\eta, \Lambda) = E[-xx^T] \]

UGM viewpoint:\n\[ p(x; \eta, \Lambda) = \exp\left( -\frac{1}{2} \sum_{ij} \Lambda_{ij} x_i x_j + \frac{1}{2} \sum_i \eta_i x_i - A(\eta, \Lambda) \right) \]
\[ p \in \mathcal{G}(J, b) \text{ where } E \nsubseteq \{ \forall i,j \text{ st. } d_{ij} \neq 0 \} \]

Zeros in precision matrix \( \Rightarrow \) conditional independence properties.

"Gaussian network"