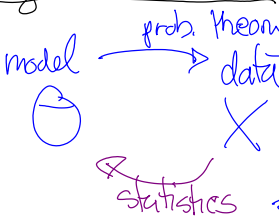
Lecture 23 - scribbles

Tuesday, November 28, 2017
14:34

today: • Bayesian approach
• model selection

Bayesian methods



$p(x|\theta)$

"frequentist" : bag of tools $\hat{\theta}$ { ML, regularized ML, max entropy, moment matching, ERM

"subjective Bayesian" → use probability everywhere there is uncertainty

→ focus on $p(\theta|\text{data}) \propto p(\text{data}|\theta)\, p(\theta)$

$p(\theta|\text{data})$: posterior; $p(\text{data}|\theta)$: likelihood; $p(\theta)$: prior

caricature: Bayesian is "optimist": she thinks you can get good models
⇒ obtain a method by doing probabilistic inference in model

frequentist is "pessimist" → use analysis tools

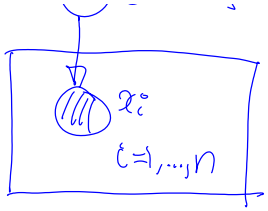
Example: biased coin

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

hyperparameter

$\theta \sim \text{Unif}[0,1]$ $\quad$ $x_i|\theta \sim \text{Bernoulli}(\theta)$ $\quad$ $p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$

$\theta \sim \text{Unif}[0,1]$



$\to$ Beta(1,1)

posterior: $p(\theta \mid x_{1:n}) \propto \left(\prod_{i=1}^{n} p(x_i \mid \theta)\right) p(\theta)$

$= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i} \mathbb{1}_{[0,1]}(\theta)$

$\propto$ Beta$(\theta \mid \alpha, \beta)$ where $\alpha = n_1 + 1$, $\beta = n - n_1 + 1$

note: if $p(\theta) = \text{Beta}(\theta \mid \alpha_0, \beta_0)$

then $p(\theta \mid \text{data}) = \text{Beta}(\theta \mid n_1 + \alpha_0, n - n_1 + \beta_0)$

here Beta$(\theta \mid \alpha_0, \beta_0)$ is a "conjugate prior" to the Bernoulli likelihood model

more generally, consider family F of dist.: $F = \{ p(\theta \mid \alpha) : \alpha \in A \}$

say that F is a "conjugate family" to observation model $p(x \mid \theta)$

if posterior $p(\theta \mid x, \alpha) \in F$

i.e. $\exists\, \alpha'$ s.t. $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$

$$\| \quad \frac{p(x \mid \theta)\, p(\theta \mid \alpha)}{\int_{\theta'} p(x \mid \theta')\, p(\theta' \mid \alpha)\, d\theta'}$$

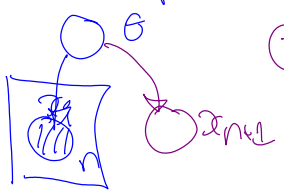
Sidenote: if use conjugate priors in a DGM,
then Gibbs sampling can be easy

---

* Bayesian in action:

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

frequentist: $\hat{\theta}_{ML} = \frac{n_1}{n}$



Bayesian: integrate out uncertainty

$$p(x_{n+1} \mid x_{1:n}) = \int_\theta p(x_{n+1} \mid \theta)\, p(\theta \mid x_{1:n})\, d\theta$$

using cond. indep.

"predictive dist"

posterior

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

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

$$\hat{\theta}_{\text{posterior mean}} = \underbrace{\frac{n_1}{n}}_{\hat{\theta}_{ML}} \underbrace{\left[\frac{n}{n+2}\right]}_{p_n} + \underbrace{\frac{1}{2}}_{\substack{\theta_{\text{prior}} \\ \text{mean}}} \underbrace{\left[\frac{2}{n+2}\right]}_{1-p_n}$$

$p_n \xrightarrow{n\to\infty} 1$

i.e. $\hat{\theta}_{\substack{\text{posterior} \\ \text{mean}}} \xrightarrow{n\to\infty} \hat{\theta}_{ML}$ = "true θ"

$\rightarrow 0$ as $n \to \infty$

Variance of a Beta : $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \left(\frac{n_1}{n}\right)\left(1-\frac{n_1}{n}\right)O\left(\frac{1}{n}\right)$

$= \hat{\theta}_{ML}(1-\hat{\theta}_{ML})O\left(\frac{1}{n}\right)$

posterior "contracts" around $\hat{\theta}_{\text{post. mean}} \overset{n\to\infty}{\rightsquigarrow} \hat{\theta}_{ML} \approx \theta^*$

"Bernstein von-Mises thm"

→ "Bayesian CLT" : basically says that if prior put non-zero mass on true parameter θ^* (→ i.e. $x_i \sim p(x|\theta^*)$) then posterior concentrates around θ^* as a Gaussian asymptotically

recall from hwk 1 : multinoulli model

$X|\theta \sim \text{Mult}(\theta)$ where $\theta \in \Delta_K$

$\hat{\theta}_{ML} = \left(\frac{n_\ell}{n}\right)_{\ell=1}^K$

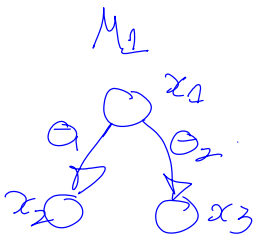
putting Dirichlet prior over θ $\theta \sim \text{Dir}(\alpha)$

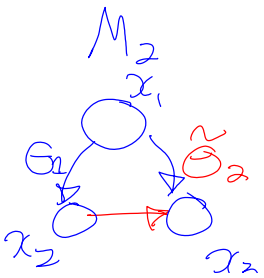
we get Dirichlet posterior $\theta|\text{data} \sim \text{Dir}((\alpha_\ell + n_\ell)_{\ell=1}^K)$

thus Dirichlet is conjugate prior to Multinoulli likelihood

Model selection:

say we want to choose between 2 DGM





(note here that "$M_1 \subseteq M_2$")

as a frequentist: $\hat{\theta}^{ML}_{M_1} = \underset{\theta_1,\theta_2}{\operatorname{argmax}} \log p(\text{data} \mid \theta_1, \theta_2, \text{model}=M_2)$ (cavalier notation)

$\hat{\theta}^{ML}_{M_2} = \underset{\theta_1,\theta_2}{\operatorname{argmax}} \log p(\text{data} \mid \theta_1, \theta_2, \text{model}=M_2)$ (different space)

how to choose between models?

can't just 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 $M_1 \subseteq M_2$

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

→ as frequentist, use cross-validation i.e. $\log p(\text{test data} \mid \hat{\theta}^{ML}_{M_2}(\text{train data}), M=M_1)$

Bayesian alternative:

true Bayesian → sum over models (integrate out uncertainty)

introduce prior over models $p(M)$

$$p(x_{new} \mid D) = \sum_M p(x_{new} \mid D, M)\, p(M \mid D)$$

$$= \sum_M \int_{\theta \in \Theta_M} p(x_{new} \mid \theta, M)\, p(\theta \mid D, M)\, p(M \mid D)\, d\theta$$

$p(\theta \mid D, M)\, p(M \mid D)$

$p(M \mid D) \propto p(M)\, p(D \mid M)$

$p(\theta \mid D, M)\, p(M \mid D)$

posterior on θ given data D and model M — marginal model probs.

$$p(x_{new} \mid data) = \sum_M p(M \mid D) \left[\int_{\theta \in \Theta_M} p(x_{new} \mid \theta, M)\, p(\theta \mid D, M)\, d\theta \right]$$

doing model averaging

Standard Bayesian predictive for one model

$p(x_{new} \mid M, D)$

(*) in model selection, forced to pick one model

$\Rightarrow$ pick model that maximizes $p(M \mid data) \propto p(data \mid M)\, p(M)$

$$\int_{\theta \in \Theta_M} p(data \mid \theta, M)\, p(\theta \mid M)\, d\theta$$

$p(data \mid M)$ = "marginal likelihood"

to compare two models, look at

$$\frac{p(M = M_1 \mid D)}{p(M = M_2 \mid D)} = \frac{p(D \mid M_1)}{p(D \mid M_2)} \frac{p(M_1)}{p(M_2)}$$

Bayes factor

prior ratio

"uniform prior over models"; then we can pick among K models $M_1, \ldots, M_K$

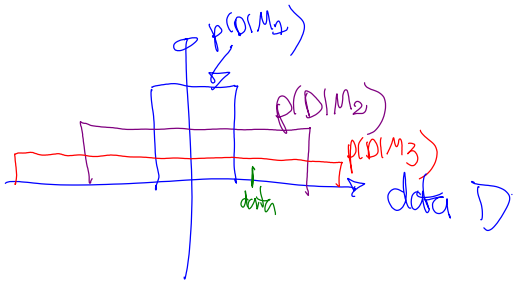
key maximizing $p(\text{data} \mid M = M_i)$

"empirical Bayes"

"type II ML"

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

Zoubin's cartoon: suppose $M_1 \subseteq M_2 \subseteq M_3$ $\quad p(\text{data} \mid \theta)$



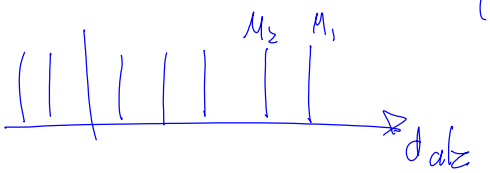
$p(D \mid M)$ is normalized over D

vs

$p(D \mid \hat{\theta}_{ML}(D), M)$

type II ML can still overfit when have many models say e.g. $p(D \mid M) = \delta(D, M)$

$\delta(D,M) = \begin{cases} 1 & \text{if } D = M \\ 0 & \text{o.w.} \end{cases}$



how to compute marginal likelihood:

use approximations ← variational inference
sampling