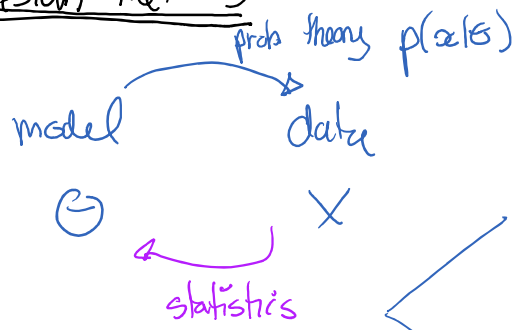
Lecture 23 - scribbles

Tuesday, November 27, 2018 14:39

today : • Bayesian approach
• model selection

Bayesian methods



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

"subjective Bayesian" → use probability everywhere there is uncertainty

↝ focus on $p(\theta|data)$ or $p(data|\theta)\,p(\theta)$

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

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

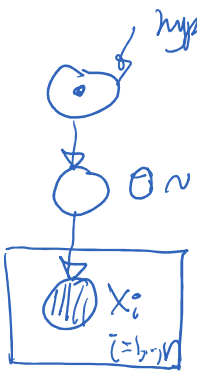
⇒ obtain a method by doing prob. inference in model

frequentist is "pessimist" → use analysis tools

Example : biased coin :

Example : biased coin :

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



$X_i \mid \theta \sim \text{Bernoulli}(\theta)$ $\quad p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$

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

$= \theta^{\underbrace{\sum_{i=1}^{n} x_i}_{n_1}} (1-\theta)^{n - \sum_{i=1}^{n} x_i} \mathbb{1}_{[0,1]}(\theta)$

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

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

↳ "conjugate prior" to the Bernoulli likelihood model

more generally;

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

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

if posterior $p(\theta|x,\alpha) \in F$ for any $x \sim X|\theta$

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

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

[e.g. → this is case in LDA topic model]

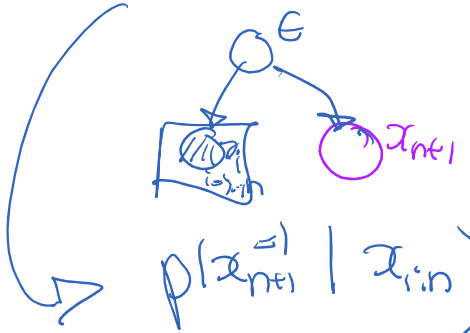
---

* Bayesian in action:

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

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

Bayesian: integrate out uncertainty



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

using cond. indep

"predictive dist."

$p(x_{n+1}|\theta)$ → θ

$p(\theta|x_{1:n})$ → posterior

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

$$\mathbb{E}[\theta|\text{data}] = \frac{\alpha}{\alpha+\beta} = \frac{n_1+1}{n_1+1+n-n_1+1} = \frac{n_1+1}{n+2}$$

[with $\alpha_0=\beta_0=1$ for uniform prior]

$$\mathbb{E}[\theta \mid \text{data}] = \frac{\alpha}{\alpha+\beta} = \frac{n_1+1}{n_1+1+n-n_1+1} = \frac{n_1+1}{n+2} \quad [\text{with } \alpha_0=\beta_0=1 \text{ for uniform prior}]$$

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

$p_n \xrightarrow{n\to\infty} 1$ i.e. $\hat{\theta}_{\text{posterior mean}} \xrightarrow{n\to\infty} \hat{\theta}_{ML} = \text{"free } \theta\text{"} = \theta^*$

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

$= \hat{\theta}_{ML}(1-\hat{\theta}_{ML})O\left(\frac{1}{n}\right) \xrightarrow{n\to\infty} 0$

posterior "contracts" around $\hat{\theta}_{\text{post mean}} \xrightarrow{n\to\infty} \hat{\theta}_{ML} = \theta^*$

"Bernstein von-Mises thm."

$\rightsquigarrow$ "Bayesian CLT": basically says that if prior put non-zero mass on true parameter θ^* [i.e. $x_i \sim p(x|\theta^*)$]

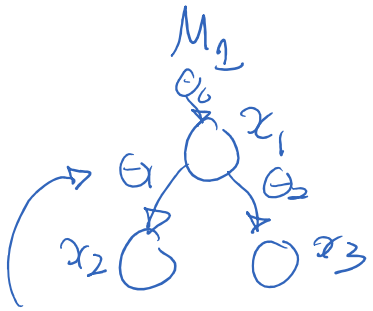
then posterior concentrates around θ^* as a Gaussian asymptotically

recall from book 1: multinomial model

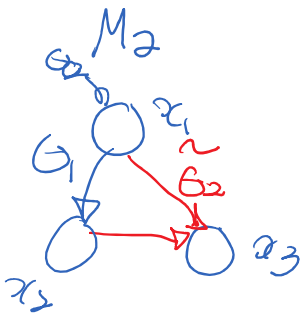
Dirichlet dist. is conjugate to multinoulli model

Model selection:

say we want to choose between 2 DGM



$p(x_2 | x_1, \theta_1)$



$p(x_3 | x_1, x_2, \tilde{\theta}_2)$

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

as a frequentist: $\hat{\theta}^{ML}_{M_1} = \underset{\theta_0, \theta_1, \theta_2}{\text{argmax}} \log p(\text{data} | \theta_0, \theta_1, \theta_2, \text{"model}=M_1\text{"})$

(caviar notation)

$\hat{\theta}^{ML}_{M_2} = \underset{\theta_0, \theta_1, \tilde{\theta}_2}{\text{argmax}}$ ... $| \theta_0, \theta_1, \hat{\theta}_2, \text{"model}=M_2\text{"})$

$\tilde{\theta}_2$: different space

how to choose between models?

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

because LHS $\leq$ RHS since $M_1 \subseteq M_2$

(ie you would always choose "bigger model")

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

Bayesian alternatives

free Bayesian → sum over models (integrate out uncertainty)

introduce prior over models $p(M)$

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

(D: data)

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

$p(M \mid D)$: marginal model prob.

$p(\theta \mid D, M)$: posterior on θ given data D & model M

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

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

$p(M \mid D)$: doing model averaging

$\int_{\theta \in \Theta_M} p(x_{new} \mid \theta, M)\, p(\theta \mid D, M)\, d\theta$: standard Bayesian predictive dist. for one model

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

15h35

⊛ in model selection, forced to pick one model

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

$p(data|M) = \int p(data|\theta, M)\, p(\theta|M)\, d\theta$ = "marginal likelihood" (likelihood)

to compare two models, look at

$$\frac{p(M=M_1|D)}{p(M=M_2|D)} = \underbrace{\frac{p(D|M_1)}{p(D|M_2)}}_{\text{Bayes factor}} \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior ratio}}$$

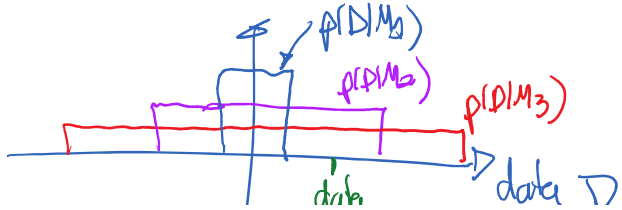
"uniform prior over models"; then we can pick among K models $M_1, \ldots, M_K$ by maximizing $p(data|M=M_i)$

"empirical Bayes"

"type II ML"

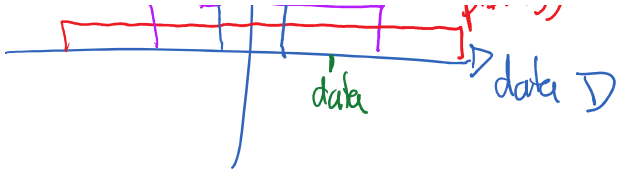
when # of models is "small" then this approach is fine (i.e. won't overfit)

Zoubin's cartoon: suppose $M_1 \subseteq M_2 \subseteq M_3$



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

vs.



vs.

$p(D \mid \hat{\theta}_{ML}(D), M)$ [can overfit]

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

how to compute marginal likelihood:

use approximations ← variational inference
sampling

Bayesian Information Criterion

BIC is a (rough) approximation of $\log p(data|M) \approx \log p(data|\hat{\theta}_{ML}, M) - \frac{d \log n}{2}$

$d = \dim(\Theta_M)$

$\frac{d \log n}{2}$ ← complexity penalty

use Laplace approximation

$$p(D|M) = \int \left[\prod_{i=1}^{n} p(x_i|\theta, M)\, p(\theta|M) \right] d\theta$$

$$= \int \exp(-n\, h(\theta))$$

where $h(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta, M) + \frac{\log p(\theta|M)}{n}$

→ do Taylor expansion of this around $\hat{\theta}_{MAP}$

↳ do Taylor expansion of this around $\hat{\theta}_{map}$

2 approximations: — keep only terms which grows with n
— replace $\hat{\theta}_{map}$ by $\hat{\theta}_{ML}$

} get BIC

BIC is "consistent"

Gaussian networks:

$X \sim N(\mu, \Sigma)$ $\quad \mu \in \mathbb{R}^p$ $\quad \Sigma \in \mathbb{R}^{p \times p}$ $\quad \Sigma \succ 0$

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

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = \underbrace{x^T \Sigma^{-1} x} - 2\mu^T \Sigma^{-1} \underbrace{x} + \mu^T \Sigma^{-1} \mu$$

$$\operatorname{tr}(x^T \Sigma^{-1} x) = \operatorname{tr}(\Sigma^{-1} x x^T)$$

$$= \langle \Sigma^{-1}, x x^T \rangle$$

linear form on matrices

canonical parameter

$\Lambda \triangleq \Sigma^{-1}$

$$-2 \langle \underbrace{\Sigma^{-1}\mu}_{\triangleq \eta}, x \rangle$$

$\rightsquigarrow \mu = \Sigma \eta$

parameter

$\Lambda \triangleq \Sigma^{-1}$

precision matrix

$-\eta$

$\Rightarrow \mu = \Sigma\eta$

$= \Lambda^{-1}\eta$

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

canonical parameter $\tilde{\eta}(\phi) = \begin{pmatrix} \eta \\ \Lambda \end{pmatrix} = \begin{pmatrix} \Sigma^{-1}\mu \\ \Sigma^{-1} \end{pmatrix}$

$$p(x;\eta,\Lambda) = \exp\left(\eta^T x + \langle \Lambda, -\tfrac{1}{2}xx^T \rangle - \underbrace{\left[\tfrac{1}{2}\eta^T\Lambda^{-1}\eta + \tfrac{N}{2}\log 2\pi - \tfrac{1}{2}\log|\Lambda|\right]}_{A(\eta,\Lambda)}\right)$$

valid for $(\eta,\Lambda)\in\Omega$

$$\Omega = \{(\eta,\Lambda) : \eta\in\mathbb{R}^p,\ \Lambda\succ 0,\ \Lambda\in\mathbb{R}^{p\times p},\ \Lambda=\Lambda^T\}$$

useful exercise:

$$\nabla_\eta A(\eta,\Lambda) = \mathbb{E}[x] = \mu = \Lambda^{-1}\eta$$

$$\nabla_\Lambda A(\eta,\Lambda) = \mathbb{E}[-\tfrac{xx^T}{2}]$$

UGM viewpoint:

$$p(x;\eta,\Lambda) = \exp\left(-\frac{1}{2}\sum_{ij}\Lambda_{ij}x_i x_j + \sum_j \eta_i x_i - A(\eta,\Lambda)\right)$$

$p \in \mathcal{J}(G)$ where $E \triangleq \{\{i,j\} \text{ s.t. } \Delta_{ij} \neq 0\}$

zeros in precision matrix $\Rightarrow$ cond. indep properties (*)

"Gaussian network"