Lecture 22 - scribbles

Friday, November 18, 2016
13:23

today: - ML in graph model
- Bayesian methods & model selection

Estimation in GM

DGM: parametric family $P_\Theta = \{ p_\theta(x) = \prod_i p(x_i | x_{\pi_i}, \theta_i) \}$ (independent parameterizations)

DGM $G=(V,E)$ $\quad \theta = (\theta_1, \dots, \theta_M) \in \Theta = \Theta_1 \times \dots \times \Theta_M$, $|V| = p$

i.e. no tying of parameters

$\Rightarrow$ MLE decouples in p independent problems

$\{x^{(i)}\}_{i=1}^n \quad p(\text{data} | \theta) = \prod_{i=1}^n p(x^{(i)} | \theta) = \prod_{i=1}^n \prod_{j=1}^p p(x_j^{(i)} | x_{\pi_j}^{(i)}, \theta_j)$

$\log[\quad] = \sum_{j=1}^p \underbrace{\left(\sum_{i=1}^n \log p(x_j^{(i)} | x_{\pi_j}^{(i)}, \theta_j) \right)}_{f_j(\theta_j)}$

example: for discrete R.V. $\Rightarrow \hat{\theta}_j^{(ML)}$ = proportion of observations of j $\frac{\#(x_j = k, x_{\pi_j} = \text{something})}{\#(x_{\pi_j} = \text{something})}$

⊛ if latent variable (unobserved)

$\Rightarrow$ use EM

UGM: example for exponential family:

$$p(x|\eta) = \exp\left(\sum_c \eta_c^T T_c(x_c) - A(\eta) \right)$$

$$\exp(\eta_c^T T_c(x_c)) = \psi_c(x_c)$$

gradient ascent on log-likelihood

$$\frac{1}{n} \sum_{i=1}^n \log p(x^{(i)} | \eta) = \sum_c \eta_c^T \left(\frac{1}{n} \sum_{i=1}^n T_c(x_c^{(i)}) \right) - A(\eta)$$

$\hat{\mu}_c$ (empirical moment)

$$\nabla_{\eta_c}(\quad) = \hat{\mu}_c - \underbrace{\mathbb{E}_{p(x|\eta)}[T_c(x_c)]}$$

$\rightarrow M_c(n)$

to compute this, need inference

Ising model

e.g. $T_{ij}(x_i, x_j)$ $\in \{0,1\}$ $= x_i x_j$

$M_{ij} = p(x_i = 1, x_j = 1 \mid n)$

perhaps use approximate inference — variational, sampling

Bayesian methods:

"frequentist": bag of tools

$\hat{\theta}$ { ML; regularized ML; max entropy; moment matching; empirical risk minimization

"subjective Bayesian" → use probability everywhere there is uncertainty

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

posterior; likelihood; prior

caricature: Bayesian is "optimist". Think 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

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

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

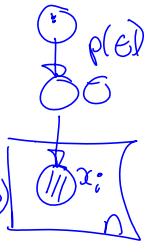
$\theta \sim \text{Unif}[0,1]$

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

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

n_1

$p(\theta) \cong \text{Unif}([0,1])$

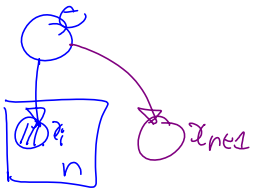


$\{ \text{Beta}(\alpha, \beta)$ where $\alpha = n_1 + 1$, $\beta = n - n_1 + 1$

questions: what is prob. of next flip = 1?

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

Bayesian, integrates out uncertainty



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

predictive distribution; posterior

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

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

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

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

variance of a Beta: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \left(\frac{n_1}{n}\right)\left(1-\frac{n_1}{n}\right)O\left(\frac{1}{n}\right) = \hat{\theta}_{ML}(1-\hat{\theta}_{ML}) \xrightarrow[n\to\infty]{} 0$

posterior "contracts" around $\hat{\theta}_{PM} \xrightarrow{n\to\infty} \hat{\theta}_{ML}$

$\hat{\theta}_{ML} \xrightarrow{a.s.} \theta^*$ (true parameter)

"Bernstein von-Mises thm" → "Bayesian CLT"

basically says that if prior put non-zero mass around the true model θ^* then posterior concentrates around θ^* as a Gaussian asymptotically

recall from hwk 1: multinomial model

$X \mid \theta \sim \text{Mult}(\theta, 1)$ where $\theta \in \Delta_K$

$$\hat{\theta}_{ML} = \frac{n_\ell}{n}$$

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

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

→ Dirichlet family is a "conjugate prior" for Multinomial likelihood

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

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

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

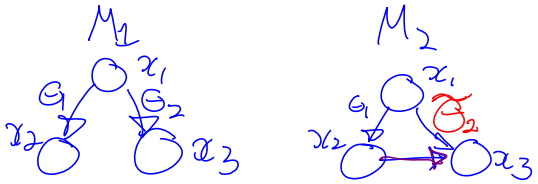
i.e. $\exists \alpha'$ s.t. $p(\theta|x,\alpha) = p(\theta|\alpha')$

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

sidenote: if use conjugate family pairs in a DGM; then Gibbs sampling can be easy

Model selection:

say want to choose between 2 DGM



(note here $M_1 \subseteq M_2$)

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

$\hat{\theta}_{M_2} = \underset{\theta_1,\tilde{\theta}_2}{\operatorname{argmax}} \log p(\text{data}|\theta_1,\tilde{\theta}_2,\text{model}=M_2)$

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

if compare $\log p(\text{data}|\hat{\theta}_{M_1}, M=M_1)$ vs. $\log p(\text{data}|\hat{\theta}_{M_2}, M=M_2)$?

since here M_1 "$\subseteq$" M_2 $\Rightarrow$ LHS $\leq$ RHS

i.e. cannot use likelihood directly to choose model

$\rightarrow$ instead, use cross-validation (i.e. $\log p(\text{test data}|\hat{\theta}^{ML}_{M_2}(\text{train data}), M=M_2)$...

Bayesian alternative:

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

introduce prior over models $p(M)$

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

data

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

posterior on parameter given data and model

$$= \sum_M p(M \mid \text{data}) \left[\int_\theta p(z_{new} \mid \theta, M)\, p(\theta \mid \text{data}, M)\, d\theta \right]$$

model averaging

standard Bayesian predictive dist. for one model

⊗ is model selection

forced to pick one model ; pick model which maximize $p(M \mid \text{data}) \propto p(\text{data} \mid M) p(M)$

"marginal likelihood"

$$p(\text{data} \mid M) = \int_\theta p(\text{data} \mid \theta, M)\, p(\theta \mid M)\, d\theta$$

to compare two models, look at

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

prior ratio

Bayes factor factor

choosing between k models $M_1, \ldots, M_k$

pick M_i which maximizes $p(\text{data} \mid M = M_i)$

"empirical Bayes"

"type II ML"

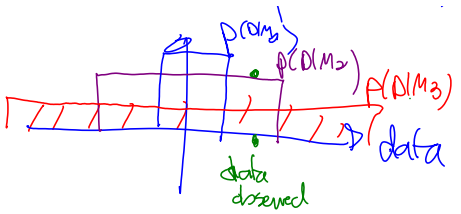
is this a good thing?

it is better than ML because $p(\text{data} \mid M = M_i)$ is normalized over the possible datasets

[vs. $p(\text{data} \mid \hat{\theta}_{ML}(\text{data}), M = M_i)$ which could be high on all datasets]

caveat : suppose $M_1 \subseteq M_2 \subseteq M_3$ (increasing complexity)

$p(D \mid M = M_i)$ = uniform on some support



depending on M_i

"type II ML" is less prone to overfitting than standard ML

(pick intermediate M_2)

② type II ML can still overfit if have many models eg. say $p(D|M) = \delta(D,M)$

how to compute Marginal likelihood?

usually need approximations ← variational inf.
sampling

Bayesian information criterion

is a (rough) approximation of $\log p(\text{data}|M)$

$$\approx \log p(\text{data}|\theta_{ML},M) - \frac{d}{2}\log(n)$$

($d = \dim(\theta_M)$)

complexity penalty

use Laplace approximation

$$p(D|M) = \int_\theta \underbrace{\prod_{i=1}^n p(x_i|\theta,M)\, p(\theta|M)}_{\exp(-n h(\theta))}\, d\theta$$

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 θ_{MAP}

2 approximations ← keep only term which grows with n
replace θ_{MAP} by θ_{ML}
} get BIC

BIC is "consistent"