Lecture 22 - Bayesian approach; model selection

Thursday, November 28, 2024 2:35 PM

today: • finish variational inference
• Bayesian
• model selection & causality

mean field continuation

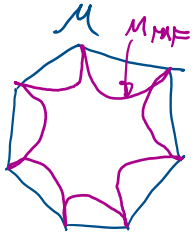
$\min_{q \in Q_{MF}} KL(q \| p)$

$\hookrightarrow \{ q : q(x) = \prod_i q_i(x_i) \}$

• $KL(\cdot \| p)$ is convex fct. of q

but Q_{MF} is non-convex constraint set

e.g. Ising model $\mu_{ij} = \mu_i \cdot \mu_j$ non-convex constraint (on moment parameterization)

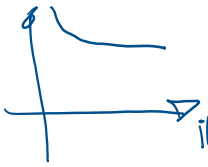


[see lecture 22 in 2017 for "marginal polytope perspective"]

lecture 22 Fall 2017 link

but can monitor progress

$KL(q^{(t)} \| p) + cst.$



pros & cons of variational methods vs sampling

⊕ optimization based ⇒ often faster to run & easier to debug

⊖ noisy ⇒ harder to debug
mixing problem for chains

⊖ biased estimate

$\mathbb{E}_{q(z)}[f(z)] \neq \mathbb{E}_p[f(z)]$

⊕ unbiased estimate

$\mathbb{E}\left[\mathbb{E}_{q(z)}[f(z)]\right] = \mathbb{E}_p[f(z)]$

with respect to random samples

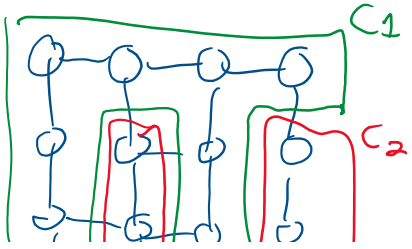
to make sure chain has mixed

$\frac{1}{N} \sum_{i=1}^{N} f(z^{(i)})$ $t = t_0$ in a chain

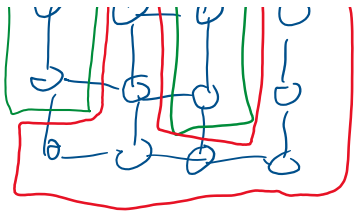
structured mean field

idea $q(z) = \prod_{j=1}^{K} q_j(z_{C_j})$ where $C_1, \dots, C_K$ is a partition of V

and q_j's are tractable distribution (for example tree UGM for q_j)

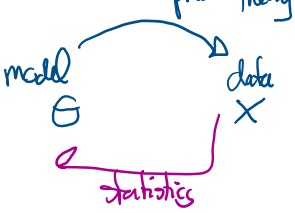


$q = q_1 \cdot q_2$



$q = q_1 \cdot q_2$

Bayesian methods

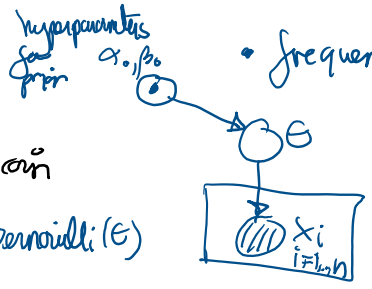


"frequentist": bag of tools $\hat{\theta}$
- MLE
- reg. MLE
- max. entropy
- moment matching
- ERM
- ...

"subjective Bayesian"
→ use prob. everywhere there is uncertainty

→ focus on $\underbrace{p(\theta|\text{data})}_{\text{posterior}} \propto \underbrace{p(\text{data}|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$

caricature:
- Bayesian is "optimist"
  they think you can get "good" models
  ⇒ obtain a method by doing inference in model
- frequentist is "pessimist" → use analysis tools



Example: biased coin

$X_i|\theta \sim \text{Bernoulli}(\theta)$

e.g. $\theta \sim \text{Unif}[0,1] = \text{Beta}(1,1)$

$p(\theta) = \text{Beta}(\theta|\alpha_0,\beta_0)$

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

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

$$= \theta^{\underbrace{\sum x_i}_{n_1}} (1-\theta)^{n-\sum x_i} \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \mathbb{1}_{[0,1]}(\theta)$$

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

↳ "conjugate prior" to the Bernoulli likelihood model

Conjugate priors:

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

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

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

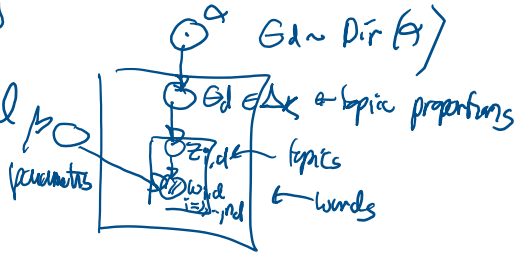
~~conjugate prior~~ ... prior $p(\theta)$

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

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

side note: if use conjugate priors in a DGM than Gibbs sampling might be easy

e.g. LDA topic model

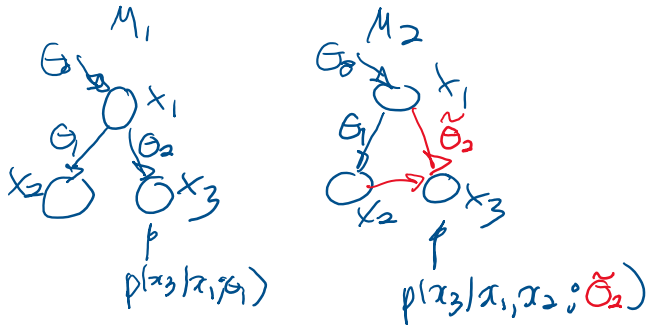


example hwk 1
Dirichlet prior is conjugate to multinomial likelihood model

15h34

Model selection

say we want to choose between 2 DGM



[note here that "$M_1 \subseteq M_2$"]

"cascaded notation"

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

$\hat{\theta}^{MLE}_{M_2} = \underset{\theta_0,\theta_1,\tilde{\theta}_2}{\text{argmax}} \log p(\text{data}|\theta_0,\theta_1,\tilde{\theta}_2, \text{"model} = M_2\text{"})$

$\tilde{\theta}_2$ ∈ different space than θ_2

how to choose between models?

can't compare $\log p(\text{data}|\hat{\theta}^{MLE}_{M_1}, M=M_1)$ vs $\log p(\text{data}|\hat{\theta}^{MLE}_{2}, M=M_2)$

because LHS ≤ RHS since $M_1 \subseteq M_2$
(i.e. you would always choose "bigger model")

→ as a frequentist, use cross-validation or validation set

i.e. $\log p(\text{test data}|\hat{\theta}^{MLE}_{M_i}(\text{train data}), M_i)$

Bayesian alternatives

True Bayesian ⇒ sum over models (integrate out uncertainty about M)

True Bayesian ⇒ sum over models (integrate out uncertainty about M)

introduce a prior over models $p(M)$

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

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

- $p(x_{new}|D,M)$: standard Bayesian predictive dist. for one model
- $p(\theta_M|D,M)$: posterior given data D & model M
- $p(M|D)$: posterior over models

(*) in model selection, forced to pick model

⇒ pick model that maximizes $p(M|data)$ or $p(data|M)\,p(M)$

$p(data|M)$ = "marginal likelihood"

$$\int_{\theta_M} p(data|\theta_M,M)\,p(\theta_M|M)\,d\theta_M$$

to compare two models, look at

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

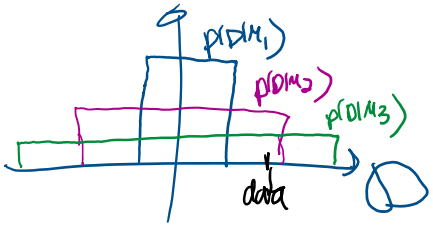
- $\frac{p(D|M_1)}{p(D|M_2)}$: Bayes factor
- $\frac{p(M_1)}{p(M_2)}$: prior ratio

"uniform prior over models" ⇒ then pick among K models $M_1,\dots,M_K$ by max $p(data|M=M_i)$

"empirical Bayes" "type II ML"

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

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



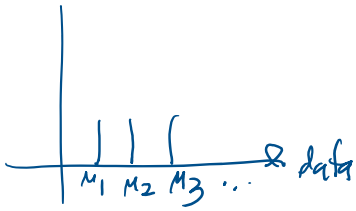
$p(D|M)$ is normalized over D

vs.

$p(D|\hat{\theta}_{MLE}(D),M)$ [can overfit badly]

but type II ML can still overfit if have too many models

say. e.g. $p(D|M) = \delta(D, M)$



how to compute marginal likelihood:

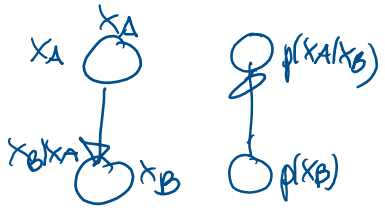
use approximations — variational inference
— sampling

simple approximation → Bayesian information criterion (BIC)

causality:

structural causal model: graph model + intervention model

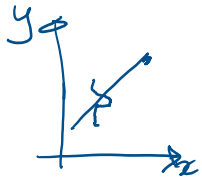
$$p(x) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}, \theta_i)$$



semantic of intervening on node J

$$p(x | \text{intervention on } J) = \left(\prod_{i \neq J} p(x_i | x_{\pi_i}, \theta) \right) p(x_J | \text{intervention})$$

identify causal direction — via parametric assumptions
— via interventions



see thoughts of Bernhard Schölkopf on causality:

https://arxiv.org/abs/1911.10500
(and references therein, e.g. his book:)

Elements of Causal Inference, 2017
By Jonas Peters, Dominik Janzing and Bernhard Schölkopf
https://mitpress.mit.edu/books/elements-causal-inference
(available for free online)