Lecture 23 - Bayesian approach; model selection

November 26, 2021 2:57 PM

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

mean field continuation

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

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

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

but Q_{MF} is a non-convex constraint set

Ising model

$\mu_{ij} = \mu_i \cdot \mu_j$

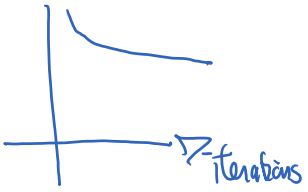
non-convex constraint

[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

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

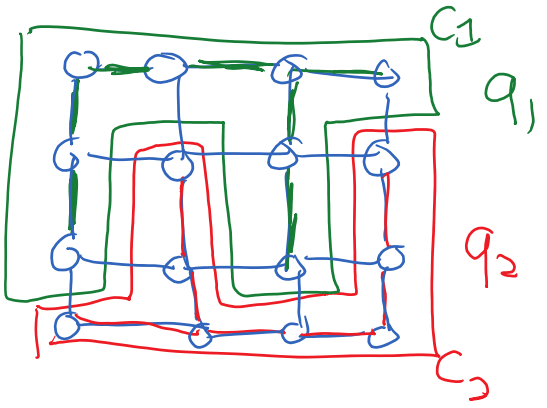
↳ with respect to random sample

structured mean field:

lecture 22 Fall 2017 link

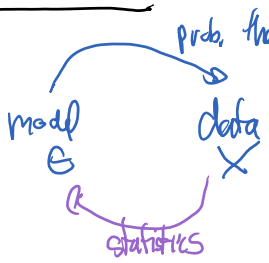
idea $q(z) = \prod_{j=1}^{k} q_j(z_{C_j})$ where $C_1, \dots, C_k$ is a partition of V

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



$q = q_1 \cdot q_2$

Bayesian methods:



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

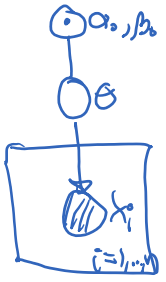
"subjective Bayesian"
→ use prob. everywhere there is uncertainty

→ 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.

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

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

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

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

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

$\sum x_i \equiv n_1$

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

"conjugate prior" to the Bernoulli likelihood model

Conjugate priors

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

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

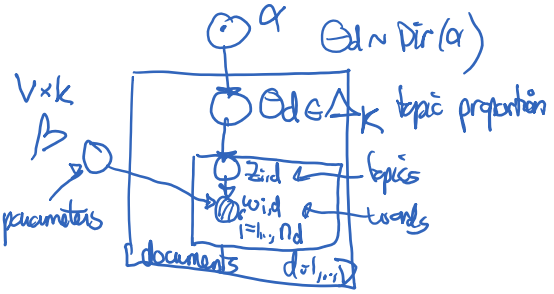
if posterior $p(\theta|x,\alpha) \in F$ for any $x \sim X|\theta$

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

Sidenote:

sidenote: if use conjugate prior in a DGM then Gibbs sampling can be easy [e.g. this is case in LDA topic model]

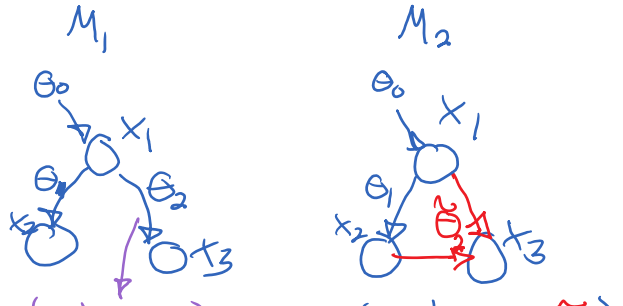
example hack 1: Dirichlet prior is conjugate for multinomial likelihood model



16h03

Model selection:

say we want to choose between 2 DGM



$p(x_3 | x_1; \theta_2)$ $\quad$ $p(x_3 | x_1, x_2; \tilde{\theta}_2)$

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

as a frequentist: ("canonical notation")

$$\hat{\theta}^{MLE}_{M_1} = \underset{\theta_0,\theta_1,\theta_2}{\arg\max} \log p(\text{data} \mid \theta_0, \theta_1, \theta_2, \text{"model} = M_1\text{"})$$

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

↳ different space than θ_2

how to choose between models?

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

because LHS ≤ RHS since $M_1 \subseteq M_2$

(ie. you would always choose "bigger model")

→ as frequentist, use cross-validation or validation set ie. $\log p(\text{test data} \mid \hat{\theta}^{MLE}_{M_2}(\text{train data}), M = M_2)$

Bayesian alternatives:

true Bayesian → sum over models (integrate out uncertainty about M)

introduce a prior over models $p(M)$

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

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

- $p(M|D)$: sum over posterior over models; doing model averaging
- $p(\theta|D,M)$: posterior on θ given data D, model M
- $p(x_{new}|\theta,M)$: standard Bayesian predictive dist. for one model

(*) in model selection forced to pick model

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

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

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

likelihood

to compare two models, look at

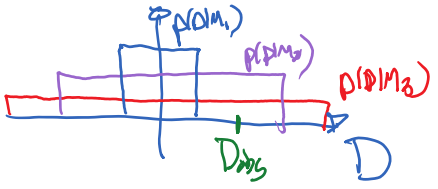
$$\frac{p(M=M_1|D)}{p(M=M_2|D)} = \underbrace{\frac{p(D|M_1)}{p(D|M_2)}}_{\text{Bayes factor}} \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior ratio}}$$

"Uniform prior over models" $\Rightarrow$ then pick among k models $M_1, \ldots, M_k$ by $p(data|M=M_i)$

"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:

how to compute marginal likelihood:

use approximations → variational inference
sampling

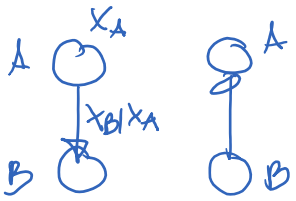
simple approximation → Bayesian information criterion

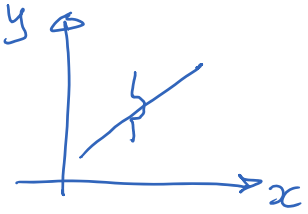
Causality:

structural causal model: graph model + intervention model

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

identify causal direction: via parametric assumptions
via interventions





semantic of intervening on node J

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

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)