Lecture 22 - scribbles

Friday, November 29, 2019 13:56

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

mean field continuation:

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

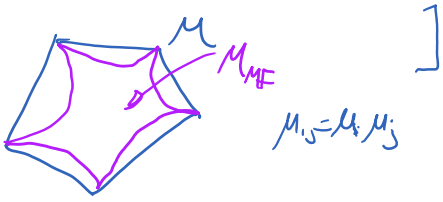
- $KL(\cdot||p)$ is a convex of q

but Q_{MF} is a non-convex constraint set $\rightarrow \mu_{ij} = \mu_i \cdot \mu_j$

$\Rightarrow$ can get stuck in local minima

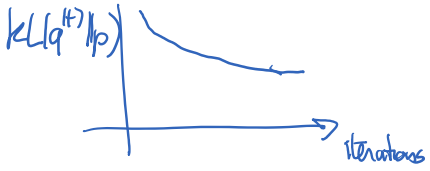
[see lecture 22 in 2017, for "marginal polytope"

lecture 22 Fall 2017 link



]

but can monitor progress by



pros & cons of variational methods vs. sampling

⊕ optimization based $\Rightarrow$ often faster to run & easier to debug

⊖ biased estimate

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

⊖ noisy $\Rightarrow$ hard to debug

mixing problem for chains

⊕ unbiased estimate

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

↳ with respect to random sample

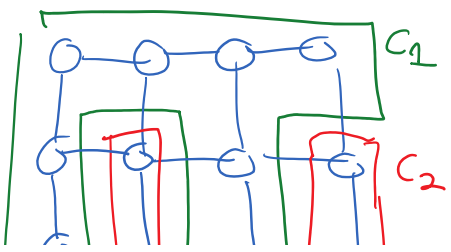
structured mean field: (see lect. 22 2017)

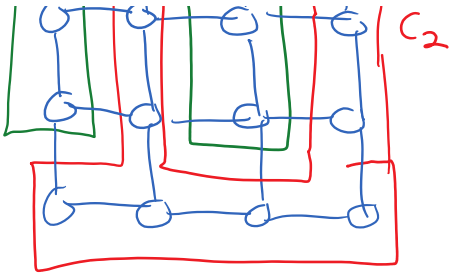
lecture 22 Fall 2017 link

idea $q(z) = \prod_{j=1}^{K} q_j(z_{C_j})$

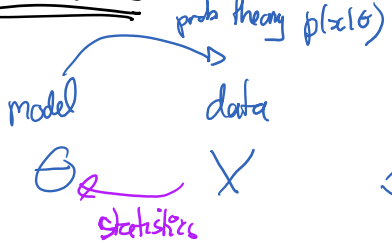
where $C_1, \ldots, C_K$ partition of V

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





Bayesian methods



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

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

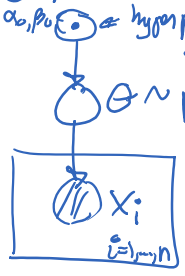
→ focus on $p(\theta|\text{data})$ or $p(\text{data}|\theta)\,p(\theta)$ (posterior; likelihood; prior)

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:

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



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

If prior $p(\theta) = \text{Beta}(\theta|\alpha_0,\beta_0)$ $= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \mathbb{1}_{[0,1]}(\theta)$, with $\sum_{i=1}^n x_i = n_1$

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

↳ "conjugate prior" to the Bernoulli likelihood model

more generally;

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

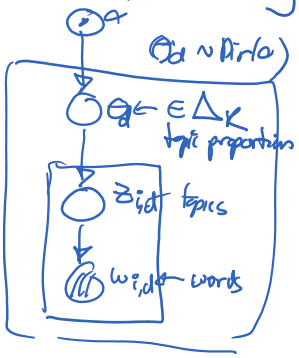
say that F is a "conjugate family" to observation model $p(x|\theta)$

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

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

Sidenote:

if use conjugate prior in a DGM then Gibbs sampling can be easy [eg → this is case in LDA topic model]

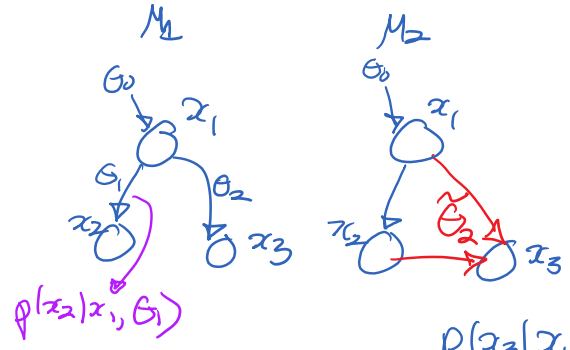


15h05

example from hwk 1
Dirichlet prior
is conjugate
for multinomial likelihood model

Model selection:

say we want to choose between 2 DGM



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

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

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

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

how to choose between models?

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

because LHS ≤ RHS since $M_1 \subseteq M_2$

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

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

Bayesian alternatives:

true Bayesian → sum over models (integrate out uncertainty)

introduce prior over model $p(M)$

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

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

marginal model prob.

posterior on θ given data D & model M

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

standard Bayesian predictive dist for one model

doing model averaging

⊛ in model selection, forced to pick one model

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

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

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

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

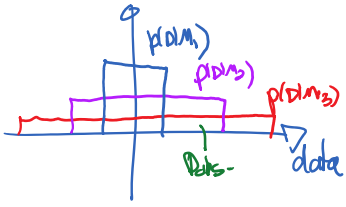
prior ratio

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

"empirical Bayes" "type II ML"

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

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



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

vs.

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

type II ML can still overfit when have many models

can e.g. $p(D \mid M) = \delta(D, M)$

many models

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

how to compute marginal likelihood:

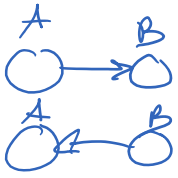
use approximations ← variational inference
sampling

simple approximation → Bayesian information criterion

Causality:

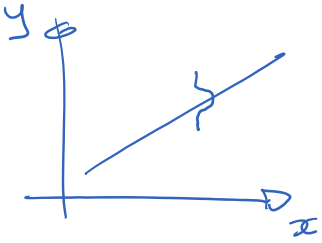
structural causal model: graph model + intervention model

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



---

or: learn graph through parametric assumption



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)