Lecture 24 - Bayesian approach; model selection

Friday, November 27, 2020 13:31

today:
- finish variational
- Bayesian
- model selection & causality

mean field continuation

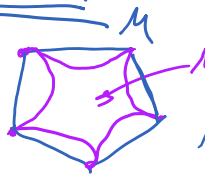
$\min_{q \in Q_{MF}} KL(q \| p)$ → $\{q : q(x) = \prod_i q_i(x_i)\}$

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

but Q_{MF} is a non-convex constraint set ⇒ can get stuck in local minima

Ising model → $\mu_{ij} = \mu_i \cdot \mu_j$? non-convex constraint

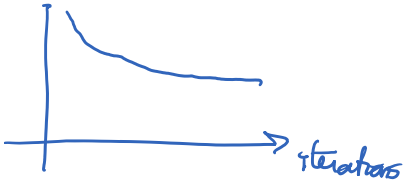
[see lecture 22 in 2017, for "marginal polytope"



lecture 22 Fall 2017 link

but can monitor progress

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



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

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

⊕ unbiased estimate

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

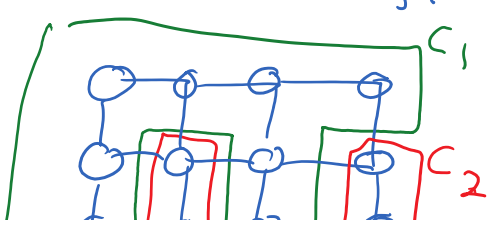
↳ with respect to random sample

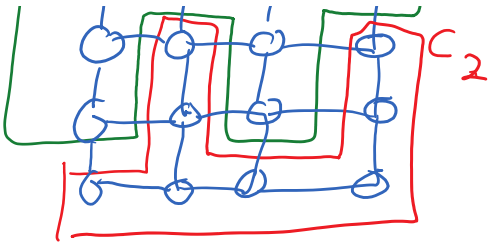
structured mean field :

lecture 22 Fall 2017 link

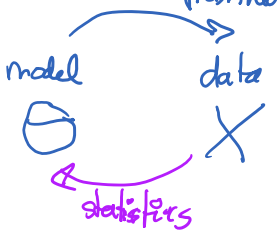
idea $q(z) = \prod_{j=1}^{K} q_j(z_{C_j})$ where $C_1, \dots, C_K$ is a partition of V

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





Bayesian methods:



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

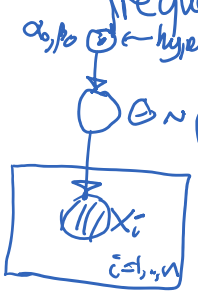
"subjective Bayesian"
→ use proba everywhere there is uncertainty

→ 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.

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

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

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

$$\text{posterior}: p(\theta|x_{1:n}) \propto \left(\prod_i p(x_i|\theta)\right) p(\theta)$$

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

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

↳ "conjugate prior" to the Bernoulli likelihood model

more generally,

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

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

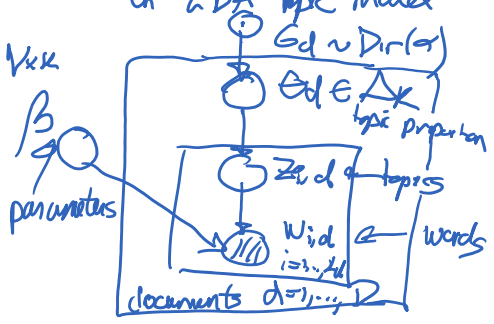
if posterior $p(\theta|x,\alpha) \in F$ for any $x \sim X|\theta$

ie. $\exists$ fn $\alpha'(x,\alpha)$ 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

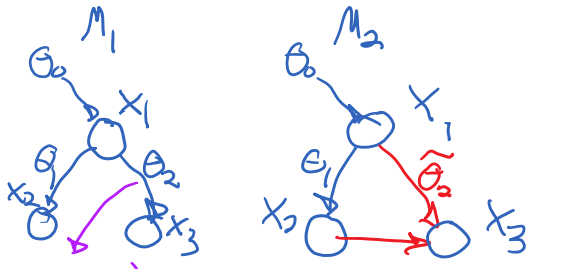


example hwk 1:
Dirichlet prior
is conjugate
for multinomial likelihood model

14h33

model selection:

say we want to choose between 2 DGM



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

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

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

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

how to choose between models?

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

because LHS ≤ RHS since $M_1 \subseteq M_2$

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

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

Bayesian alternatives :

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

introduce a prior over models $p(M)$

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

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

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

(*) in model selection, forced to pick model

$\Rightarrow$ pick model that maximizes $p(M \mid \text{data})$ or $p(\text{data} \mid M)\, p(M)$

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

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

likelihood

to compare two models, look at

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

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

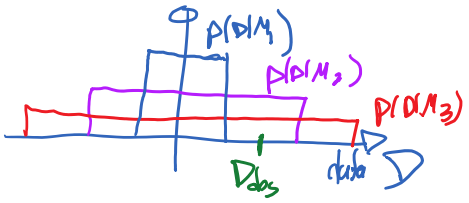
"uniform prior over models"; $\Rightarrow$ then pick among k models $M_1, \ldots, M_k$ by max. $p(\text{data} \mid M = M_i)$

"empirical Bayes" "type II ML"

when # of models is "small", then this approach is fine

i.e. won't overfit)

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

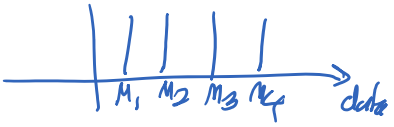


$p(D|M)$ is normalized over D

vs.

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

but $\mathbb{M}_{ML}$ can still overfit if have many models



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

how to compute marginal likelihood:

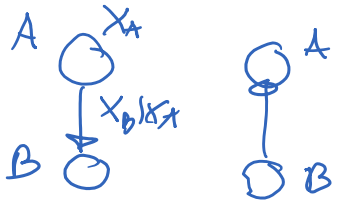
use approximations ← variational inference
sampling

simple approximation → Bayesian information criteria

Causality:

structural causal model: graph model + intervention model

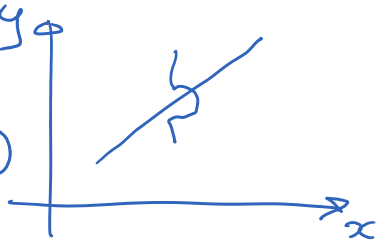
$$p(x) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}, \theta_i)$$



identify causal direction — via intervention
via parametric assumptions

Semantic of intervening on node J

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



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)