Cours 10 - scribbles

Wednesday, December 2, 2015
9:00

today: · Bayesian approach
· model selection

variational methods vs. sampling

recall for mean field

$$\min_{q \in Q_\pi} KL(q \,\|\, p) \quad \} \text{ non-convex}$$

p ← target dist.

Q_π ↳ fully factorized distributions i.e. $q(x) = \prod_i q_i(x_i)$

in case where $p(x)$ was from Ising model
→ can coordinate descent on q_i's

approximate marginal on node $q_i(x_i = 1) = \hat{\mu}_i$

say we have converged to a stationary pt. $\hat{\mu}_i^*$
usually $\hat{\mu}_i^* \neq p(x_i = 1)$

"biased" / "inexact"

in contrast: sampling methods are usually asymptotically unbiased

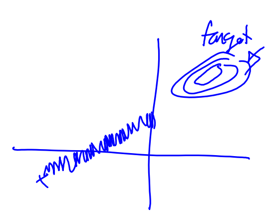
example: if use Gibbs sampling to get $\vec{x}^{(t)}$ (samples)

then $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} x_i^{(t)} = p(x_i = 1)$

from Ergodic thm

"asymptotically unbiased"

problem: "mixing time" ← how long it takes to forget initial conditions



sometimes can be very long "sticky chain"

⇒ slow convergence of Monte Carlo estimate

* in practice, you can not use the first few samples to reduce the bias of estimate

"burn in time"

summary:

variational approach:

- often faster than sampling, but it is inexact

e.g. QMR-DT network seconds for variational vs hours for sampling

- easier to debug the variational

Bayesian statistics

$p(x \mid \theta)$ — can observe (x), model (θ)

model Θ $\xrightarrow{\text{prob. theory}}$ data X

statistics } → "subjective Bayesian" use prob theory (Bayes rule)

→ "frequentist" [traditional statistics]

bag of tools:
- ML
- regularized ML
- max entropy
- moment matching
- etc ...

caricature: Bayesian is "optimist": think that can come up with good models, → obtain a method by pulling the Bayesian crank

frequentist is more pessimistic: → use analysis tools

Bayesian: $p(x \mid \theta)$ "likelihood model"

$p(\theta)$ ← "prior"

posterior: $\underbrace{p(\theta \mid x)}_{\text{posterior}} = \dfrac{p(x \mid \theta)\, p(\theta)}{p(x)}$ Bayes rule

$p(x)$] "marginal likelihood" → normalization

"Summarize belief of Bayesian"

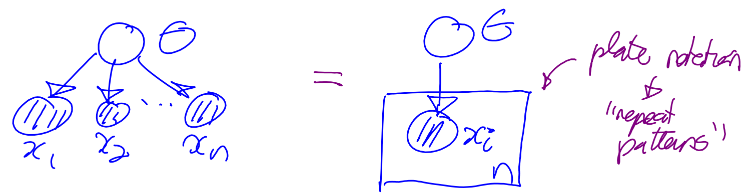
$\boxed{\text{example: biased coin}}$

Bayesian model $\quad x_i \in \{0,1\}$

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

$$\theta \sim \text{Unif}[0,1] \quad \text{(uniform prior)}$$

graph model:

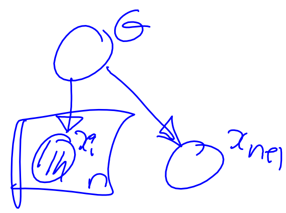


posterior:
$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)\, p(\theta)$$
$$= \left(\prod_{i=1}^{n} p(x_i \mid \theta)\right) p(\theta)$$
$$= \theta^{\underbrace{\sum_{i=1}^n x_i}_{\triangleq n_1}} (1-\theta)^{n - \sum_{i=1}^n x_i}\, \mathbb{1}_{[0,1]}(\theta)$$

Beta(α, β) distribution where $\alpha = n_1 + 1$
$\beta = n - n_1 + 1$

question: what is probability of next flip $= 1$?

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



Bayesian integrates out the uncertainty

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

predictive distribution

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

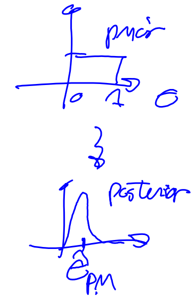
mean of Beta(a,β) is $\frac{a}{a+\beta} = \frac{n_1+1}{n+2}$ here

notice that for $n=0 \rightarrow$ get $\frac{1}{2}$

smoothed version of ML

$$\hat{\theta}_{\substack{\text{posterior}\\ \text{mean}}} = \frac{n_1}{n}\left[\frac{n}{n+2}\right] + \frac{1}{2}\left[\frac{2}{n+2}\right] = \frac{n_1+1}{n+2}$$

$$= \hat{\theta}_{ML}\, p_n + \hat{\theta}_{prior}\,(1-p_n)$$



convex combination of $\hat{\theta}_{ML}$ & $\hat{\theta}_{prior}$

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

variance of a beta: $\frac{a\beta}{(a+\beta)^2(a+\beta+1)} = \left(\frac{n_1}{n}\right)\left(1-\frac{n_1}{n}\right)\underbrace{O\left(\frac{1}{n}\right)}_{\xrightarrow{n\to\infty} 0}$

$\hat{\theta}\,(1-\hat{\theta})$

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

true parameter

$\hat{\theta}_{ML} \xrightarrow{a.s.} \theta^*$

"Bernstein von-Mises thm"

$\rightarrow$ "Bayesian central limit thm." which basically says that if prior puts non-zero mass around true model θ^*, then posterior concentrates around θ^* as a Gaussian asymptotically

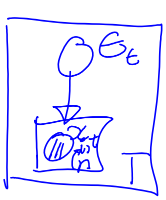
for a Gaussian mean & mode are the same

so can approximate $\mathbb{E}[\theta \mid \text{data}]$

with $\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid \text{data})$

[fine for large n]

<u>revisiting the example:</u>



J coins picked randomly each flipped n times...

as a frequentist, empirical distribution on $x_{1:n}$ will converge as $T \to \infty$

to $p(x_1, \dots, x_n) = \int_\Theta \left(\prod_{i=1}^n p(x_i|\theta) \right) p(\theta)\, d\theta$

distribution of coins in jar

(continuous) mixture distribution

here, $X_1, \dots, X_n$ are not independent

on the other hand, $p(x_1, \dots, x_n) = p(x_{\pi(1)}, \dots, x_{\pi(n)})$

for any permutation $\pi : 1{:}n \to 1{:}n$

$X_1, \dots, X_n$ are "exchangeable" weaker than independence

De Finetti's representation thm. :

$X_1, X_2, \dots$ is infinitely exchangeable

$\iff$ $\exists$ unique $p(\theta)$ on some space Θ

s.t. $p(x_1, \dots, x_n) = \int_\Theta \left(\prod_{i=1}^n p(x_i|\theta) \right) p(\theta)\, d\theta$

Multinomial model

e.g. modeling words...

$X|\theta \sim \mathrm{Mult}(\theta, 1)$ where $\theta \in \Delta_K$

i.e. $\sum_{\ell=1}^{K} \theta_\ell = 1$

$\theta_\ell \geq 0$

$\hat{\theta}_\ell^{ML} = \frac{n_\ell}{n}$ if $K > n$

then some $\hat{\theta}_\ell^{ML} = 0$ for some ℓ

$\Rightarrow$ overfitting

as a Bayesian, put prior on $\Delta_K = \Theta$

a convenient property of prior family is "conjugacy"

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

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

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

if posterior $\underset{\in F}{p(\theta|x,\alpha)} = \dfrac{p(x|\theta)\,p(\theta|\alpha)}{p(x|\alpha)}$

hyperparameter

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

for multinomial:

likelihood $\displaystyle p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$

$$= \prod_{\ell=1}^{k} \theta_\ell^{\sum_{i=1}^{n} x_{i,\ell}} \qquad x_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

if use prior $\displaystyle \propto \prod_{\ell=1}^{k} \theta_\ell^{\text{stuff}}$

Dirichlet distribution $\rightarrow$ dist. over Δ_K

$$\mathrm{Dir}(\theta|\alpha_1,\dots,\alpha_K) = \frac{1}{B(\vec{\alpha})} \prod_{\ell=1}^{k} \theta_\ell^{\alpha_\ell - 1}$$

convention

valid $\alpha_\ell > 0$

$$\mathbb{E}[\theta_\ell|\alpha_1,\dots,\alpha_K] = \frac{\alpha_\ell}{\sum_j \alpha_j}$$

$$\text{variance}(\theta_\ell) = O\left(\frac{1}{\sum_j \alpha_j}\right)$$

- $\alpha_\ell = 1 \rightarrow$ get uniform distribution
- $K=2 \rightarrow$ get Beta distribution
- $\alpha_\ell < 1$ $\cup$ shape distribution (no mode)
- $\alpha_\ell \geq 1 \ \forall \ell$ $\cap$ unimodal bump

* for multinomial model: if prior: $p(\theta|\alpha) = \mathrm{Dir}(\theta|\vec{\alpha})$

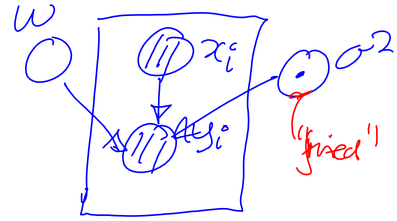
posterior $\displaystyle p(\theta|x_1,\dots,x_n) \propto \prod_{\ell=1}^{k} \theta_\ell^{n_\ell + \alpha_\ell - 1}$

$$= \mathrm{Dir}\left(\theta \,\middle|\, (n_\ell + \alpha_\ell)_{\ell=1}^{K}\right)$$

posterior mean: $\mathbb{E}[\theta_\ell|\text{data}] = \dfrac{n_\ell + \alpha_\ell}{\dots}$ "prior..

$n + \sum_j a_j$ "counts"

Bayesian linear regression



say $y = w^T x + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

observation model: $p(y \mid x, w) = N(y \mid w^T x, \sigma^2)$

prior: $p(w) = N(w \mid 0, \frac{I}{\lambda})$ ← λ precision

(conjugate)

posterior: $p(w \mid y_{1:n}, x_{1:n})$ is also Gaussian

with covariance $\hat{\Sigma}_n = \lambda I + \frac{X^T X}{\sigma^2}$ $\quad \underset{n}{\overset{d}{X}}$ design matrix $\quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

posterior mean $\hat{\mu}_n = \hat{\Sigma}_n^{-1} \left(\frac{X^T \vec{y}}{\sigma^2} \right)$

↳ same as in ridge regression with $\tilde{\lambda} = \sigma^2 \lambda$

as a Bayesian:

compute predictive dist. $p(y_{new} \mid x_{new}, x_{1:n}, y_{1:n})$

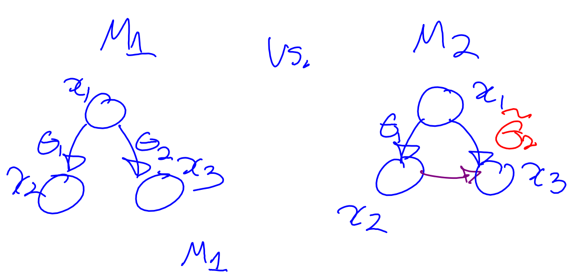
$$= \int_w \underbrace{p(y_{new} \mid x_{new}, w)}_{\text{Gaussian obs. model}} \underbrace{p(w \mid \text{data})}_{\text{Gaussian posterior}} \, dw$$

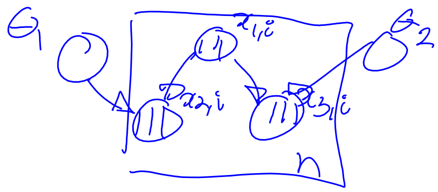
$$= N(y_{new} \mid \hat{\mu}_n^T x_{new}, \sigma^2_{predictive}(x_{new}))$$

$$\sigma^2_{predictive}(x_{new}) = \underset{\text{noise model}}{\sigma^2} + \underbrace{x_{new}^T \hat{\Sigma}_n x_{new}}_{\text{from posterior covariance}}$$

Model selection

say want to choose between



as a Bayesian 

vs. ...

as a frequentist,

$$\hat{\theta}^{ML}_{M_1} = \underset{\theta_1, \theta_2}{\operatorname{argmax}} \log p(\text{data} \mid \theta_1, \theta_2, M_1)$$

$$\hat{\theta}^{ML}_{M_2} = \underset{\theta_1, \underbrace{\tilde{\theta}_2}_{\text{different space}}}{\operatorname{argmax}} \log p(\text{data} \mid \theta_1, \tilde{\theta}_2, M_2)$$

compare $\log p(\text{data} \mid \hat{\theta}^{ML}_{M_1}, M_1)$ vs. $\log p(\text{data} \mid \hat{\theta}^{ML}_{M_2}, M_2)$?

here M_1 "$\subseteq$" $M_2 \Rightarrow$? $\leq$ ↗

here, likelihood is useless...

instead, use cross-validation

* here, Bayesian alternative:

true Bayesian, sum over models
(integrate out the uncertainty)

$$p(x_{new} \mid \underset{\text{data}}{D}) = \sum_M \int_\theta p(x_{new} \mid \theta, M) \underbrace{p(M, \theta \mid D)}_{p(\theta \mid D, M)\, p(M \mid \text{data})} d\theta$$

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

model averaging

standard predictive dist. for one model

* if force to pick model:

$$\text{pick model which maximize } p(M \mid \text{data}) \propto \underbrace{p(\text{data} \mid M)}_{\text{"marginal likelihood"}} p(M)$$

$$p(D \mid M) = \int_\theta p(D \mid \theta, M)\, p(\theta \mid M)\, d\theta$$

here, sloppy: mean $p(M = m_1 \mid D)$, $p(M = m_2 \mid D)$

$p(M=M_2|D)$

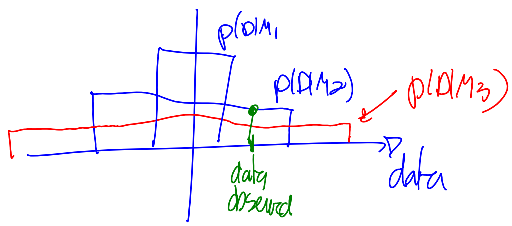
Bayes factor $\quad \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)} \frac{p(M_1)}{p(M_2)}$

pick $\quad \hat{M} = \arg\max_{M} p(D|M) \leftarrow$ "empirical Bayes"
"type 2 ML"

too many models $\rightarrow$ can still overfit

cartoon why marginal likelihood works vs. ML:

$M_1 \subseteq M_2 \subseteq M_3$



BIC criterion $\rightarrow$ approximation $p(D|M)$