Probability Density Estimation

Machine Learning - 10601

Geoff Gordon, Miroslav Dudík

(partly based on slides of Carlos Guestrin and Tom Mitchell)

http://www.cs.cmu.edu/~ggordon/10601/

September 14, 2009

Last time...

Inference in factor graphs/Bayes nets:

$P(M,Ra,O,W,Ru) \propto \phi_1(M)\ \phi_2(Ra)\ \phi_3(O)\ \phi_4(Ra,O,W)\ \phi_5(M,W,Ru)\ /Z$

$P(W \mid Ra=F, Ru=T) = ?$

(1) Incorporate evidence:

$P(M, Ra=F, O, W, Ru=T) \propto \dots$

(2) Eliminate nuisance nodes:

$P(W, Ra=F, Ru=T) \propto \sum_M \sum_O \dots = \phi'(w)$

(3) Normalize:

$P(W \mid Ra=F, Ru=T) = \phi'(w) / \sum_w \phi'(W=w)$

Last time…

Benefits of factored representations:

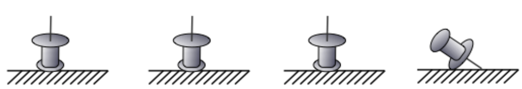
- efficient inference (sometimes)
- fewer parameters to estimate

approximate

LEARNING SIMPLER

less info needed

Last time: maximum likelihood



heads w/prob **θ**

N tosses

H heads, N−H tails

p(H | N, θ) = $\binom{N}{H}\theta^{H}(1-\theta)^{N-H}$ binomial dist.

$\max_{\theta} p(H|N,\theta)$

$\hat{\theta}_{ML} = \frac{H}{N}$

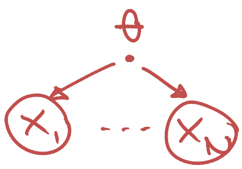
$\binom{N}{H} = \frac{N!}{H!(N-H)!}$

Are we learning?

Task: probability estimation (prediction)

Performance measure: likelihood

Experience: thumbtack tosses



$$x_n = \begin{cases} 1 & \text{if heads w/prob } \theta \\ 0 & \text{if tails w/prob } 1-\theta \end{cases}$$

$$p(x_1,\dots,x_N|\theta) = \prod_n p(x_n|\theta)$$

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

$$\max_\theta \frac{1}{N}\sum_n \log p(x_n|\theta)$$

AVG $\approx E_{\theta^*}[\log p(x|\theta)]$

$x \sim p(x|\theta^*)$

$x_n \sim p(x|\theta^*)$

true

Are we learning?

APPROX

TRUE PERF. MEASURE

θ fixed

$$\frac{1}{N}\sum_n \log p(x_n|\theta) \approx E_{\theta^*}[\log p(x|\theta)]$$

x random var

$$\arg\max_\theta E_{\theta^*}[\dots] = \theta^*$$

Maximum likelihood estimation

- **expected log likelihood**
  maximized by **true distribution**
- **average log likelihood of data**
  approximates **expected log likelihood**

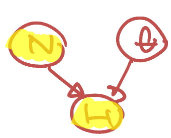
~~**GREAT!**~~ for small sample sizes
approximation poor

Bayesian approach

initial belief over values of **θ**

e.g. **θ**~ uniform over **[0,1]**

$p(\theta) = 1,$ $\quad \int_0^1 p(\theta)\, d\theta = 1$



$p(\theta \mid N, H) \propto p(\theta, N, H)$

$= \cancel{p(N)}\, \cancel{p(\theta)}\, p(H \mid N, \theta)$

as a fct of θ: const.; $p(\theta) = 1$; $p(H \mid N, \theta) = \cancel{\binom{N}{H}}\, \theta^H (1-\theta)^{N-H}$; $\binom{N}{H}$ const.

$\propto \theta^H (1-\theta)^{N-H}$

Bayesian approach

θ, N parameters

H data, $\mathcal{D}$

p(θ) prior

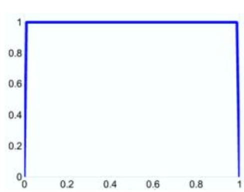
p($\mathcal{D}$ | θ, N) likelihood

p(θ |$\mathcal{D}$, N) posterior

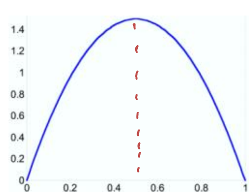
Priors and posteriors

posterior $\propto \theta^{H}(1-\theta)^{N-H}$

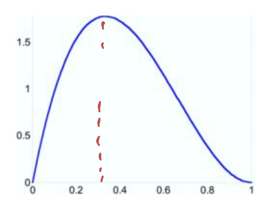
prior: p(θ)



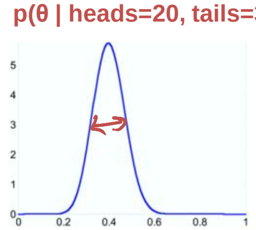
posterior:
p(θ | heads=1, tails=1)



posterior:
p(θ | heads=1, tails=2)



posterior:
p(θ | heads=20, tails=30)



MORE OBS

↳ less uncertainty
(usually)

Beta family

$\in [0,1]$

$\boxed{\text{Beta}(\theta \mid \underline{\alpha}, \underline{\beta}) \propto \underline{\theta^{\alpha-1}(1-\theta)^{\beta-1}}}$

- uniform: $p(\theta) = 1$, $\alpha = \beta = 1$
- after H heads, T tails: $p(\theta \mid H, T) \propto \underline{\theta^{H}(1-\theta)^{T}}$

$N-H$, $\alpha' = H+1$, $\beta' = T+1$

HYPERPARAMS

$p(D \mid \theta) = \boxed{\theta^{H}(1-\theta)^{T}}$

Say $p(\theta) = \text{Beta}(\theta \mid \alpha, \beta) \propto \boxed{\theta^{\alpha-1}(1-\theta)^{\beta-1}}$

- after H heads, T tails: $p(\theta \mid H, T) \propto p(\theta, \underbrace{H, T}_{D})$

$= p(\theta)\, p(D \mid \theta)$

$= \theta^{H+\alpha-1}(1-\theta)^{T+\beta-1}$

$= \text{Beta}(H+\alpha, T+\beta)$

Conjugacy

- if posterior has the same form as prior, we say: prior is conjugate relative to likelihood
- e.g.: **Binomial** and **Beta** are conjugate families

Conjugacy: simple Bayesian inference

Bayesian updating

p(θ) = Beta(θ | α, β) prior

observe **H** heads, **T** tails

p(θ | H, T) = Beta(θ | α+H, β+T) ~~old posterior~~ new prior

observe additional **H'** heads, **T'** tails

p(θ | H, T, H', T') = $\mathrm{Beta}(\theta \mid \underbrace{\alpha+H+H'}_{\alpha'}, \underbrace{\beta+T+T'}_{\beta'})$ new posterior

PRIOR = summary of past experience

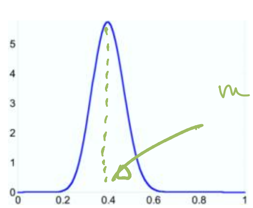
Predicting next outcome

observe **H** heads, **T** tails

next observation **X**

- max likelihood $\hat{\theta}_{ML} = \frac{H}{H+T}$ $\quad p(X=\text{heads} \mid \hat{\theta}_{ML}) = \hat{\theta}_{ML}$

- prior **p(θ) = Beta(θ | α, β)**
  posterior **p(θ | H, T) = Beta(θ | α+H, β+T)**

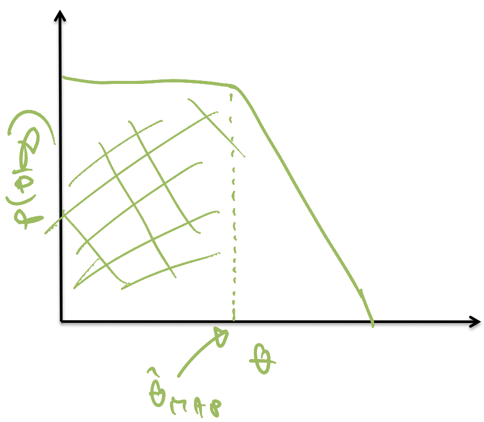


maximum a posteriori (MAP)

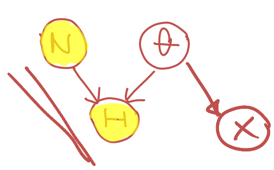
$\hat{\theta}_{MAP}$

if $p(\theta)$ uniform then $\hat{\theta}_{ML} = \hat{\theta}_{MAP}$

What if posterior looks like…



Bayesian prediction



p(θ) = Beta(θ | α, β)

$$x = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$

$$P(X \mid N, H)$$

$$\propto p(x, N, H)$$

$$= \int p(x, N, \theta, H)\, d\theta$$

$$= \int \underbrace{p(N,H)}_{\text{const.}}\; \underbrace{p(\theta \mid N, H)}_{\propto\, \theta^{H+\alpha-1}(1-\theta)^{T+\beta-1}}\; \underbrace{p(x \mid \cancel{N}, \theta, \cancel{H})}_{\propto\, \theta^{x}(1-\theta)^{1-x}}\, d\theta$$

$$\propto \int \theta^{H+\alpha+x-1} (1-\theta)^{T+\beta+(1-x)-1}\, d\theta$$

Bayesian prediction

LAPLACE SMOOTHING (for UNIFORM & BINOM)

$$p(x \mid N, H) \propto \int \underbrace{\theta^{\overbrace{H+\alpha+x}^{\alpha'}-1} (1-\theta)^{\overbrace{T+\beta+(1-x)}^{\beta'}-1}}_{\text{Beta dist}} \, d\theta$$

NORMALIZATION OF BETA

BUT NOT OF $p(x \mid N, H)$

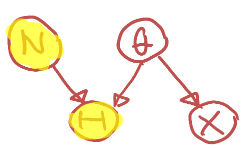
$$\int \theta^{\alpha'-1} (1-\theta)^{\beta'-1} \, d\theta = \frac{(\alpha'-1)!\,(\beta'-1)!}{(\alpha'+\beta'-1)!} \qquad (\alpha'-1)+(\beta'-1)+1$$

$$p(x \mid N, H) \propto \begin{cases} \dfrac{\overset{(H+1)H!}{(H+1)!}\, T!}{(H+T+2)!} & \text{for } x=1 \\[2ex] \dfrac{H!\,\overset{(T+1)T!}{(T+1)!}}{(H+T+2)!} & \text{for } x=0 \end{cases}$$

$$p(x \mid N, H) = \begin{cases} \dfrac{H+1}{H+T+2} & \text{for } x=1 \\[2ex] \dfrac{T+1}{H+T+2} & \text{for } x=0 \end{cases}$$

NORMALIZE →

Bayesian prediction vs MAP



MAP: $\max_{\theta} p(\theta \mid N, H)$

Bayesian prediction:

$$p(X \mid N, H) \propto \int \overbrace{p(X \mid \theta)}^{p(x \mid \theta, \cancel{N}, \cancel{H})} \; \overbrace{p(\theta \mid N, H)}^{\text{const. } p(N,H)} \, d\theta$$

$$= \propto p(x, N, H, \theta)$$

DIFFICULT TO CALCULATE: APPROXIMATE

What you should know

Maximum likelihood estimation (MLE)

- approximates maximization of **expected log likelihood**
- **expected log likelihood** maximized by **true distribution**
- approximation poor on small data sets

Bayesian posterior, MAP, Bayesian prediction

- **posterior** reflects uncertainty in the parameter
- **conjugacy**: posterior has the same form as the prior
  e.g., **Beta** and **Binomial**
- **prior** as a summary of **previous experience (observations)**
- **maximum a posteriori** can suffer similar problems as MLE — STAT ISSUES
- **Bayesian prediction** can be intractable — COMPUT. ISSUES