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:

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

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

(1) Incorporate evidence:

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

(2) Eliminate nuisance nodes:

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

(3) Normalize:

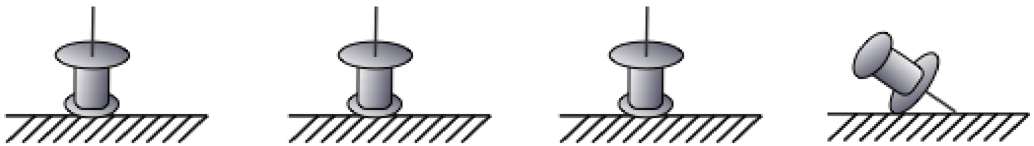
$P(W \mid Ra=F, Ru=T) =$

Last time…

Benefits of factored representations:

- efficient inference
- fewer parameters to estimate

Last time: maximum likelihood



heads w/prob **θ**

N tosses

H heads

p(H | N, θ) =

Are we learning?

Task:

Performance measure:

Experience:

Are we learning?

Maximum likelihood estimation

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

GREAT!

Bayesian approach

initial belief over values of **θ**

e.g. **θ** ~ uniform over **[0,1]**

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

Bayesian approach

θ, N parameters

H data, $\mathscr{D}$

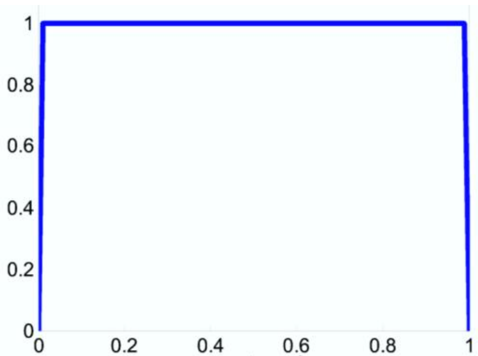
p(θ)

p($\mathscr{D}$ | θ, N)

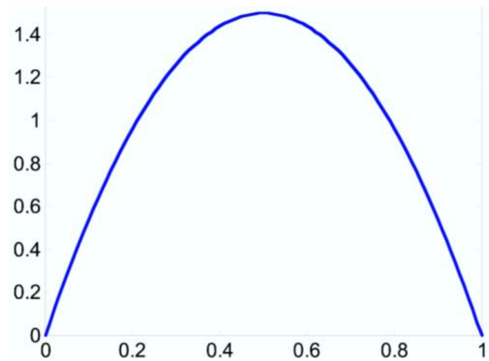
p(θ |$\mathscr{D}$, N)

Priors and posteriors

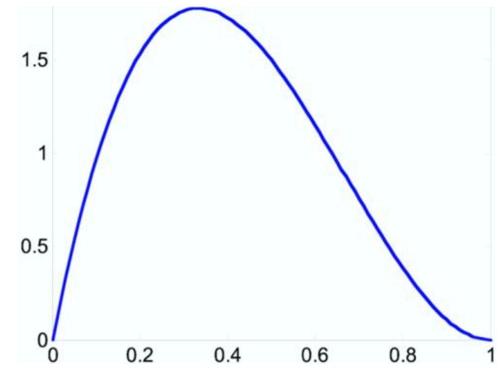
prior: p(θ)



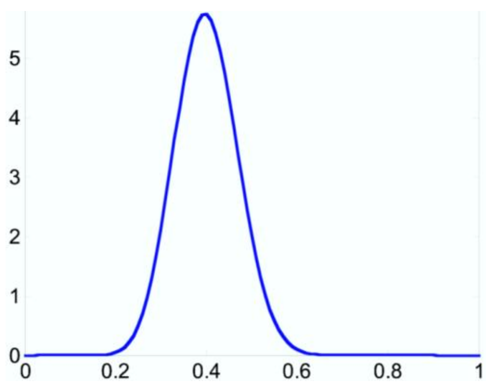
posterior:
p(θ | heads=1, tails=1)



posterior:
p(θ | heads=1, tails=2)



posterior:
p(θ | heads=20, tails=30)



Beta family

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

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

Say **$p(\theta) = \text{Beta}(\theta \mid \alpha, \beta)$**

- after **H** heads, **T** tails: **$p(\theta \mid H, T) =$**

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

$p(\theta) = \text{Beta}(\theta \mid \alpha, \beta)$

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

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

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

$p(\theta \mid H, T, H', T') =$

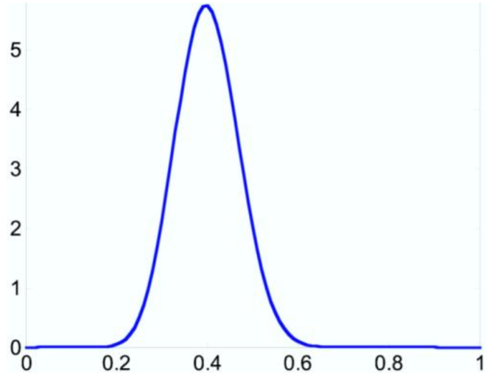
Predicting next outcome

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

next observation **X**

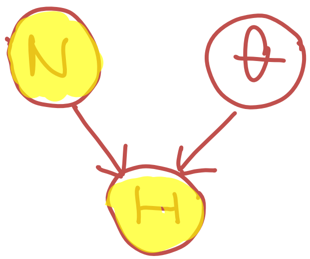
- max likelihood

- prior **$p(\theta) = \text{Beta}(\theta \mid \alpha, \beta)$**
  posterior **$p(\theta \mid H, T) = \text{Beta}(\theta \mid \alpha+H, \beta+T)$**



What if posterior looks like…

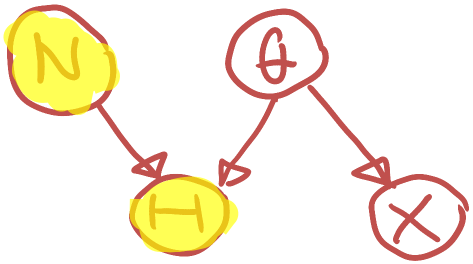
Bayesian prediction



p(θ) = Beta(θ | α, β)

Bayesian prediction

Bayesian prediction vs MAP



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

Bayesian prediction:

$$p(X \mid N, H) \propto \int p(X \mid \theta)\, p(\theta \mid N, H)\, d\theta$$

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
- **Bayesian prediction** can be intractable

Learning to classify text documents

- classify which emails are spam?
- classify which emails promise an attachment?
- classify which web pages are student home pages?

As a subroutine:

- for each category, learn the distribution over the documents belonging there

“Bag of words” approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Multinomial and Dirichlet

Multinomial:
$$p(N_1, N_2, \dots, N_K \mid \theta_1, \dots, \theta_K) = \frac{N!}{N_1! \cdots N_K!} \theta_1^{N_1} \cdots \theta_K^{N_K}$$

Dirichlet:
$$p(\theta_1, \dots, \theta_K \mid \alpha_1, \dots, \alpha_K) \propto$$

Subtle points:

- dictionary is potentially infinite
- need to estimate "missing mass"

Gaussian distribution