CS786
Lecture 18: July 2nd, 2012

Bayesian Parameter Estimation

[KF Chapter 17]

Bayesian Learning

- Idea: compute posterior over hypotheses given data
$$\Pr(H|d) \propto \Pr(H)\Pr(d|H)$$
- Recall candy example: $H \in \{h_1, \ldots, h_5\}$
  - h_1: 100% lime, h_2: 75% lime, …, h_5: 0% lime
  - Finitely many hypotheses
- What if we have infinitely many hypotheses?
  - E.g., continuum of parameters

Simplified Candy Example

- One-node Bayesian network:
- Parameter: $\theta = \Pr(F = cherry)$

- Continuum of hypotheses: $0 \leq \theta \leq 1$
- What is a good family of distributions over the [0,1] interval?

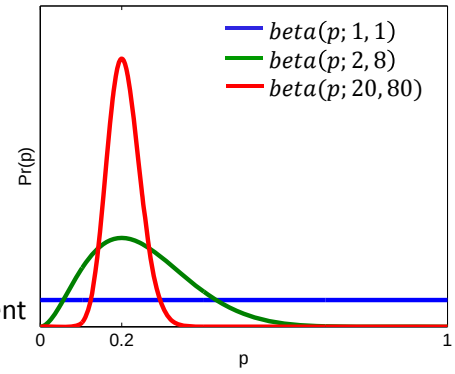
Beta Distribution

- Distribution over Bernoulli
- Definition: $beta(p; \alpha, \beta) = kp^{\alpha-1}(1-p)^{\beta-1}$

  where $k = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$
- Mean: $\frac{\alpha}{\alpha+\beta}$
- Precision: $\alpha + \beta$
- Interpretation:

  $\alpha - 1$: freq. of p-event

  $\beta - 1$: freq. of $(1-p)$-event



Conjugate Prior

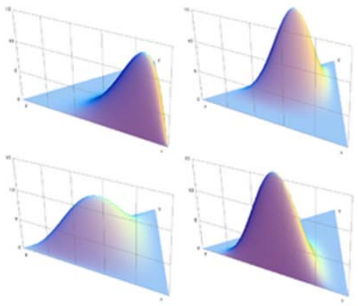
- Prior: Let $\Pr(\theta) = beta(\theta; \alpha, \beta)$
- Suppose that we eat c cherries and l limes
- Posterior: $\Pr(\theta|c, l) \propto \Pr(\theta)\, \theta^c (1-\theta)^l$
  $= k\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^c(1-\theta)^l$
  $= k\theta^{\alpha+c-1}(1-\theta)^{\beta+l-1}$
  $\propto beta(\theta; \alpha + c, \beta + l)$
- Definition: a family of distributions is said to be a **conjugate prior** when it is closed under Bayesian learning for a certain type of likelihood distribution.
  - E.g., the Beta distribution is a conjugate prior for the Bernoulli distribution

Dirichlet Distribution

- Generalization of the Beta to multivalued variables
- Let $\boldsymbol{\theta} =< \theta_1, \theta_2, \ldots, \theta_m >$ such that $\sum_{i=1}^m \theta_i = 1$
- **Dirichlet**: $Dir(\boldsymbol{\theta}; \boldsymbol{n}) = k \prod_{i=1}^m \theta_i^{n_i - 1}$

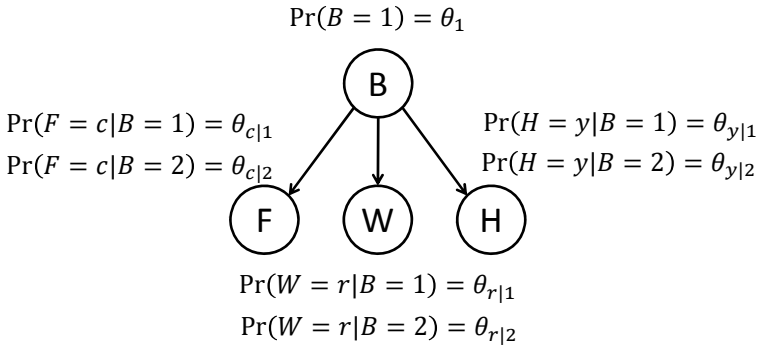
  Where $\boldsymbol{n} =< n_1, n_2, \ldots, n_m >$

  and $k = \frac{\Gamma(\sum_i n_i)}{\prod_i \Gamma(n_i)}$



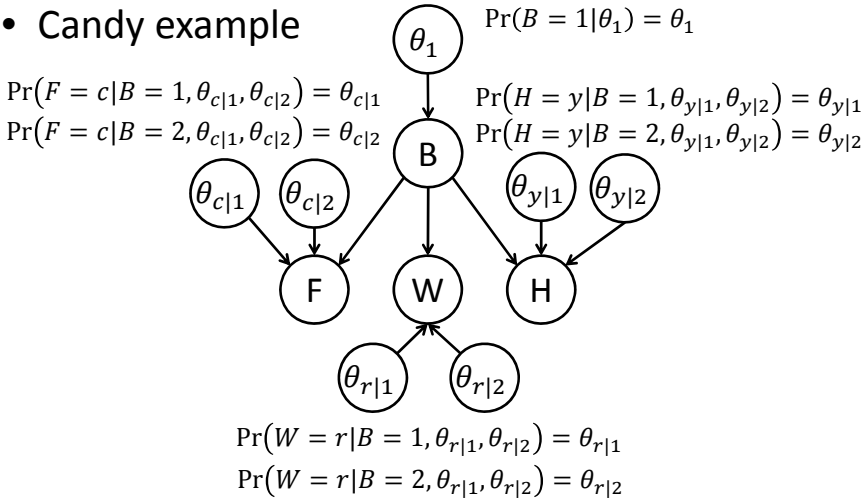
Bayesian Network

- Candy example



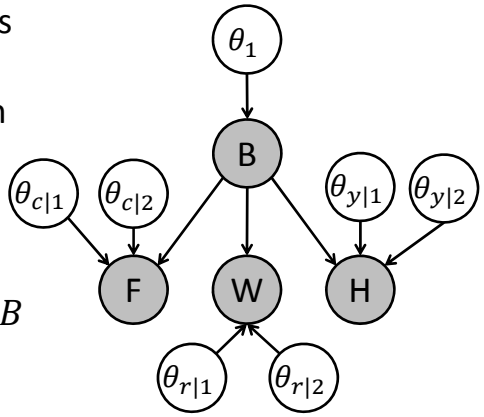
Augmented Bayesian Network

- Candy example



Supervised Learning

- D-separation allows us to estimate the parameters of each CPT in isolation
- E.g., $\theta_{c|1}$ and $\theta_{c|2}$ depends on F and B only



Supervised Learning

- Start with Dirichlet priors over all θ's.
  e.g., $\Pr(\theta_{c|1}) = beta(\theta_{c|1}; \alpha, \beta)$
- Compute posterior over each parameter
  e.g., $\Pr(\theta_{c|1}|data) \propto \Pr(\theta_{c|1}) \Pr(data|\theta_{c|1})$
  $= k\theta_{c|1}^{\alpha-1}(1-\theta_{c|1})^{\beta-1}\theta_{c|1}^{\#(c,1)}(1-\theta_{c|1})^{\#(l,1)}$
  $= k\theta_{c|1}^{\alpha+\#(c,1)-1}(1-\theta_{c|1})^{\beta+\#(l,1)-1}$
  $\propto beta(\theta_{c|1}; \alpha+\#(c,1), \beta+\#(l,1))$

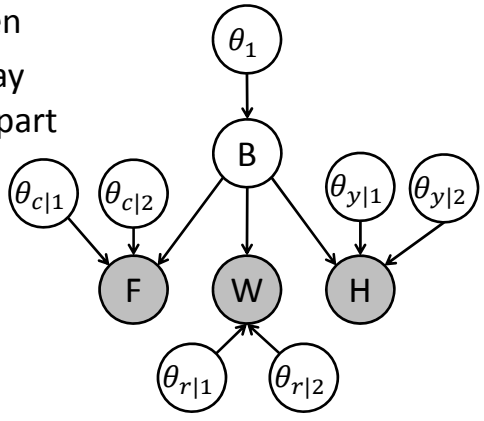
Prediction

- Query: $\Pr(F = c|B = 1)$?
- Average over hypotheses

$$\begin{aligned}
&\Pr(F = c|B = 1\,) \\
&= \int_{\theta_{c|1}} \Pr\big(F = c\big|B = 1, \theta_{c|1}\big)\Pr(\theta_{c|1}|data)\, d\theta_{c|1} \\
&= \int_{\theta_{c|1}} \theta_{c|1}\Pr(\theta_{c|1}|data)\, d\theta_{c|1} \\
&= E_{beta(\theta_{c|1};\, \alpha+\#(c,1),\beta+\#(l,1))}\big[\theta_{c|1}\big] \\
&= \frac{\alpha+\#(c,1)}{\alpha+\beta+\#(c,1)+\#(l,1)} \qquad \leftarrow \textbf{smoothed relative frequencies}
\end{aligned}$$

Unsupervised Learning

- Suppose B is hidden
- Each parameter may depend on a large part of the network
- E.g. $\theta_{c|1}$ depends on everything



Unsupervised Learning

- Posterior distribution of each parameter:
  - **Closed form:** mixture of Dirichlets
  - **Intractable:** # of mixture components grows exponentially with the amount of data.
- Solution: approximations
  - Gibbs sampling
  - Expectation propagation

Gibbs Sampling

- Sample each hidden variable in turn given values for all the other variables
- E.g. sample

$B \sim \Pr(B|F,W,H,\boldsymbol{\theta})$

(regular inference query)

$\theta_{c|1} \sim \Pr(\theta_{c|1}|F,W,H,\boldsymbol{\theta}_{\sim c|1})$

(sample from posterior Dirichlet)