Lecture 22: April 13, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Review

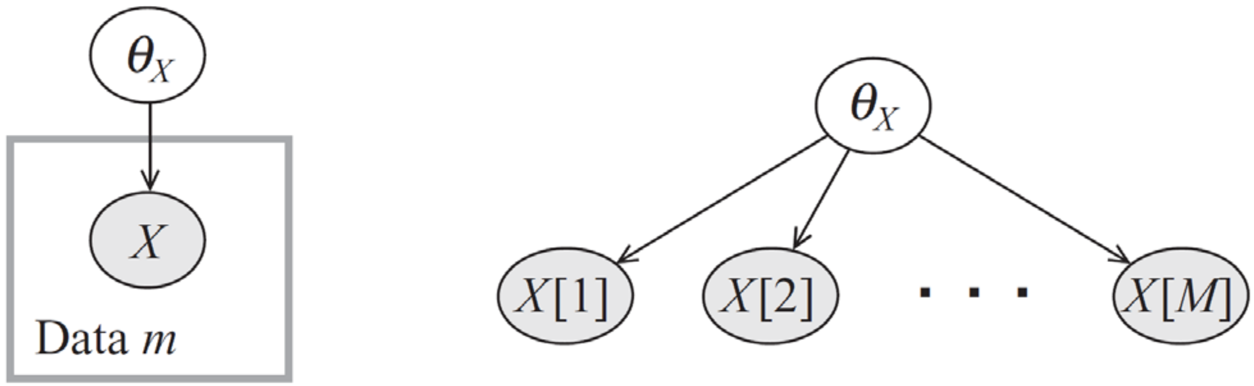
- HW #6B due today
- HW#7 to be assigned this week (last assignment)
- Previous Lecture
  - Issues in learning
  - Maximum Likelihood Estimate
- Today's objective
  - Bayesian parameter learning

Bayesian Parameter Estimation

- In MLE, we learn a value for each parameter θ. In Bayesian learning, we treat θ also as a random variable with some distribution. Our task is to learn this distribution and probabilities of variables as a function of θ.
- Consider example of coin flip
  - If "heads" 7 out of 10 trials, is P(H) = .7? Could a fair coin also not produce this outcome (with some probability)?
- Can make use of prior knowledge of parameters, such as whether the coin is fair (or not)
- Parameter distributions are modified based on observations; as number of observations increases, they will start to dominate the priors

Detour to Plate Models

- Figure (a) is “short-hand for Fig (b);
- (a) is plate model; (b) *ground* Bayesian Network



(a) (b)

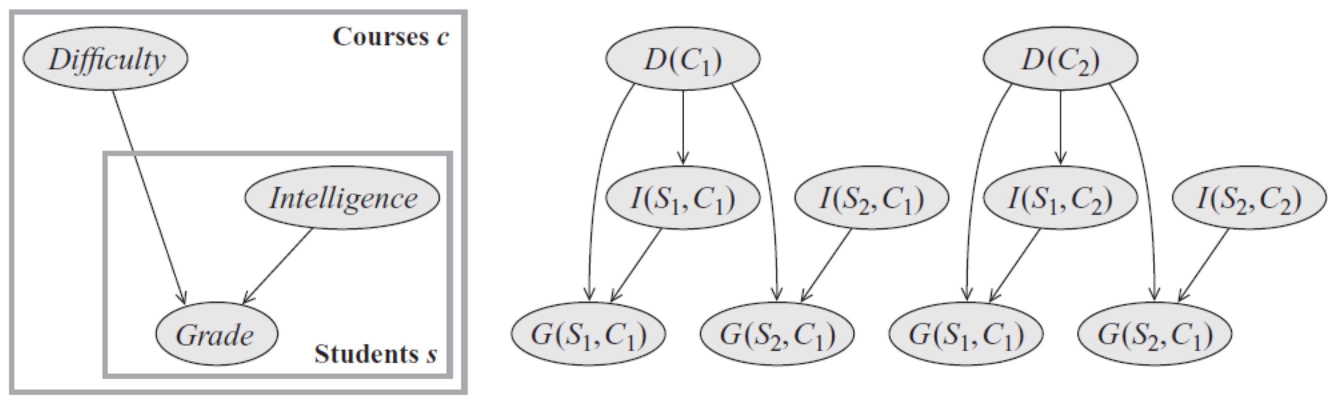
- Grey box shows a “plate”: implies that samples X[i] are iid, dependent only on θ_x.
- More complex plate models on next slide

Nested and Intersecting Plates

Single Plate

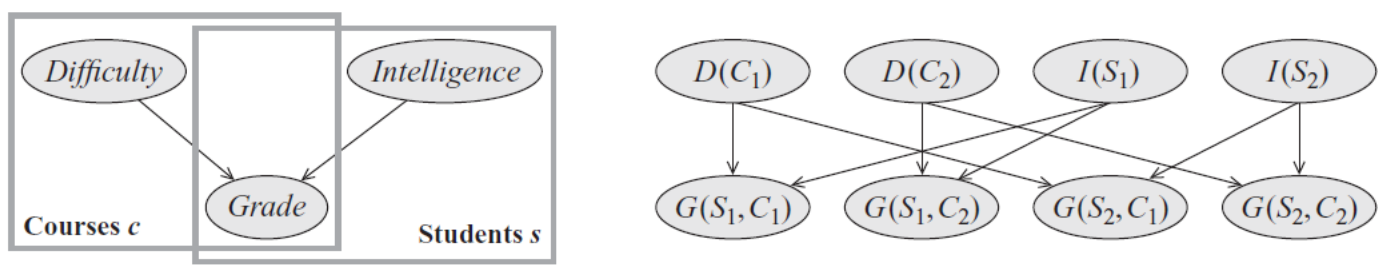
(a)

Nested Plates



(b)

Intersecting Plates

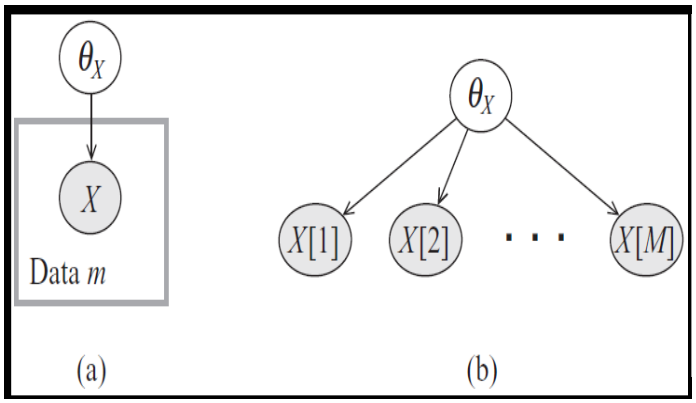


(c)

ground BNs

Joint Probabilistic Model

- $P(x[m]|\theta) = \theta$ if $x[m] = x^1$;
  $= 1- \theta$ if $x[m] = x^0$
  - Note the use of "|" above in place of ":" as θ is now also a random variable



$$
\begin{aligned}
P(x[1], \ldots, x[M], \theta) &= P(x[1], \ldots, x[M] \mid \theta) P(\theta) \\
&= P(\theta) \prod_{m=1}^{M} P(x[m] \mid \theta) \\
&= P(\theta) \theta^{M[1]} (1-\theta)^{M[0]},
\end{aligned}
$$

$$
P(\theta \mid x[1], \ldots, x[M]) = \frac{P(x[1], \ldots, x[M] \mid \theta) P(\theta)}{P(x[1], \ldots, x[M])}
$$

Note: update of θ , numerator is product of prior and likelihood; denominator is a normalizing constant

Prediction

$$
\begin{aligned}
& P(x[M+1] \mid x[1], \ldots, x[M]) = \\
&= \int P(x[M+1] \mid \theta, x[1], \ldots, x[M]) P(\theta \mid x[1], \ldots, x[M]) d\theta \\
&= \int P(x[M+1] \mid \theta) P(\theta \mid x[1], \ldots, x[M]) d\theta,
\end{aligned}
$$

- For the thumbtack example, assuming **uniform prior** over θ in [0,1]

$$
P(X[M+1] = x^1 \mid x[1], \ldots, x[M]) = \frac{1}{P(x[1], \ldots, x[M])} \int \theta \cdot \theta^{M[1]} (1-\theta)^{M[0]} d\theta.
$$

$$
= \frac{M[1]+1}{M[1]+M[0]+2}.
$$

- Similar to MLE but with a correction factor. As M becomes large, the two estimators become close to each other.

General Formulation

- $P(D,\boldsymbol{\theta}) = P(D\,|\boldsymbol{\theta})\,P(\boldsymbol{\theta})$
- $P(\boldsymbol{\theta}\,|\,D) = P(D\,|\boldsymbol{\theta})\,P(\boldsymbol{\theta})\,/\,P(D)$
- $P(D) = \int_{\Theta} P(D\,|\boldsymbol{\theta})\,P(\boldsymbol{\theta})\,d\boldsymbol{\theta}$
- P(D) is called marginal likelihood

Choice of Priors

- Use of priors in certain forms can make the calculation of posteriors more convenient.
  - If the two have the same form, they are called *conjugate priors*.
- For Bernuolli distribution (one binary variable), *beta* functions are conjugate priors
  - Beta functions have two parameters, by varying them we can get a uniform or highly skewed distributions
- For multinomial distributions, Dirichlet distributions are conjugate priors
- As these are commonly used priors, it is good to develop some familiarity with them.

Beta Distributions

- β distribution is characterized by two parameters, α_1 and α_0, both are positive real numbers. Distribution is given by:

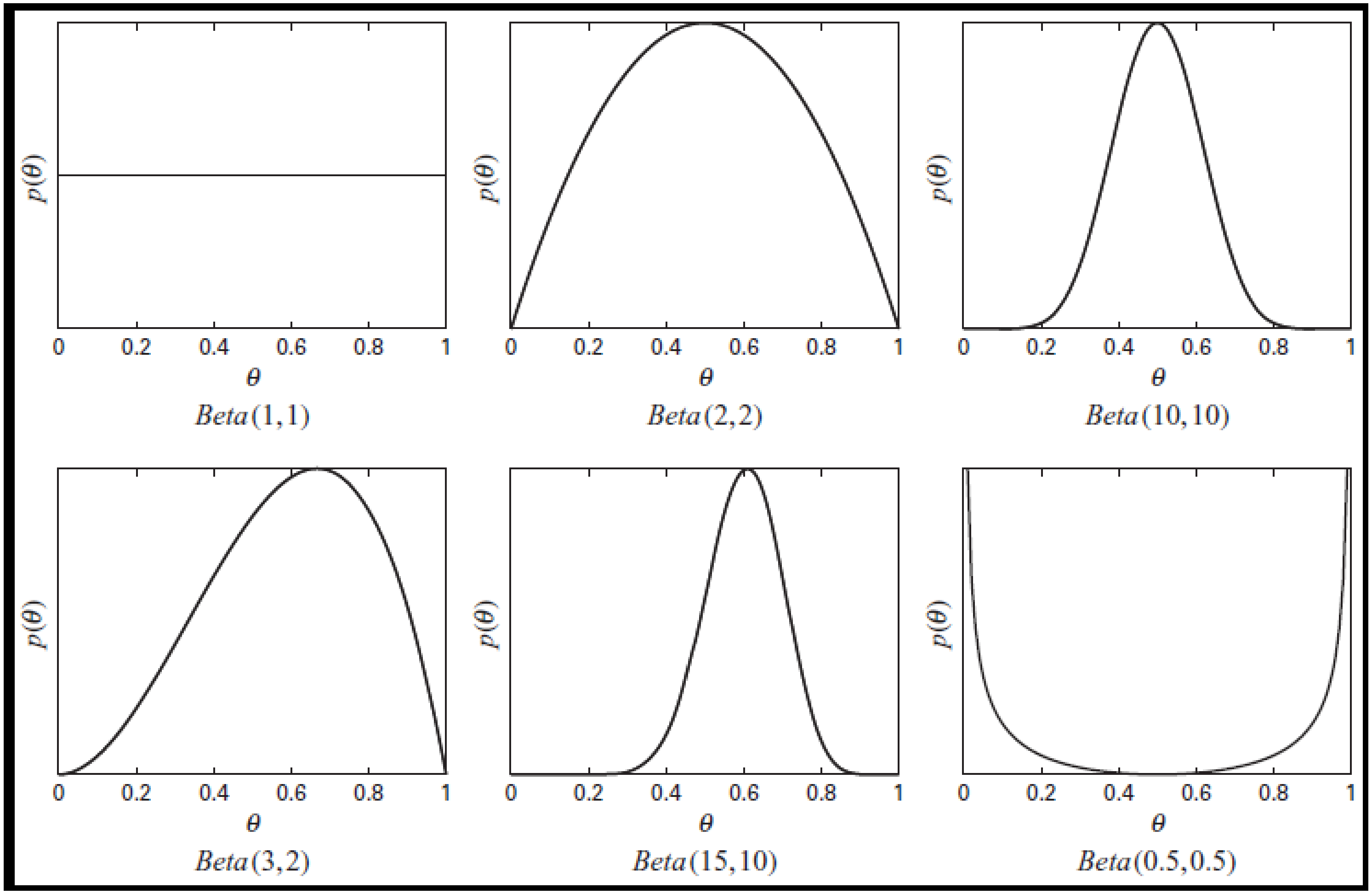
$$\theta \sim \text{Beta}(\alpha_1, \alpha_0) \textit{ if } p(\theta) = \gamma \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

- γ is a normalizing constant given by:

$$\gamma = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \qquad \Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \textit{ is the } \text{Gamma function.}$$

- $\Gamma(1) = 1$; $\Gamma(x+1) = x\,\Gamma(x)$; $\Gamma(n+1) = n!$ if n is integer
- α_1 and α_0, both can be viewed as *hyperparameters* (which control the distribution of parameter of interest θ)
- Examples of β distributions on next slide; note higher values of parameters give more peaked distributions

Beta Distribution Examples



Prediction Probabilities

- Distribution of one outcome:

$$\begin{aligned} P(X[1] = x^1) &= \int_0^1 P(X[1] = x^1 \mid \theta) \cdot P(\theta) d\theta \\ &= \int_0^1 \theta \cdot P(\theta) d\theta = \frac{\alpha_1}{\alpha_1 + \alpha_0}. \end{aligned}$$

- Updating θ based on observations:

$$\begin{aligned} P(\theta \mid x[1], \ldots, x[M]) &\propto P(x[1], \ldots, x[M] \mid \theta) P(\theta) \\ &\propto \theta^{M[1]}(1-\theta)^{M[0]} \cdot \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \\ &= \theta^{\alpha_1 + M[1] - 1}(1-\theta)^{\alpha_0 + M[0] - 1}, \end{aligned}$$

- Note: above is $\beta\,(\alpha_1+M[1], \alpha_0+M[0])$; conjugate function
- Predict probability of $(M+1)^{th}$ sample :

$$P(X[M+1] = x^1 \mid x[1], \ldots, x[M]) = \frac{\alpha_1 + M[1]}{\alpha + M} \qquad \alpha = \alpha_1 + \alpha_0$$

- Role of α_1 and α_0 can be viewed as adding some pseudo sample counts to the observed sample counts.
  - Their influence becomes smaller as number of samples grows
  - Allows for prior knowledge to be incorporated naturally.

Multinomial Distributions

- Parameter vector is $\theta = <\theta_1, \theta_2 \ldots \theta_k>$
- Hyperparameters defined by *Dirichlet distribution*

$$\boldsymbol{\theta} \sim Dirichlet(\alpha_1, \ldots, \alpha_K) \text{ if } P(\boldsymbol{\theta}) \propto \prod_k \theta_k^{\alpha_k - 1} \qquad \alpha \text{ to denote } \sum_j \alpha_j$$

- Note: beta distribution is a special case of Dirichlet.
- Working through the algebra, it can be shown that posterior is also Dirichlet.

$$P(\theta|D) = \mathit{Dirichlet}\,(\alpha_1+M[1],\ldots,\alpha_k+M[k])$$

- Prediction: $P(x[M+1] = x^k \mid \mathcal{D}) = \dfrac{M[k] + \alpha_k}{M + \alpha}$
- Role of α_k can again be viewed as adding some pseudo sample counts
  - Most effective when the number of samples is small

Bayesian Estimation of Bayesian Networks

- Consider table CPDs. Probability of each variable X_i is a function only of its parents, say U.
  - Thus distribution of samples is multinomial again; if priors are specified by Dirichlet, so are the posteriors.
  - Skipping mathematical derivations, we get:

$$P(X_i[M+1] = x_i \mid \boldsymbol{U}[M+1] = \boldsymbol{u}, \mathcal{D}) = \frac{\alpha_{x_i|\boldsymbol{u}} + M[x_i, \boldsymbol{u}]}{\sum_i \alpha_{x_i|\boldsymbol{u}} + M[x_i, \boldsymbol{u}]}$$

- How to choose the Dirichlet coefficients?
  - Should they be same for each CPD? Can human expert provide these?
- KF book claims that experience indicates that using priors gives much more “stable” results than MLE.

Next Topic Choices

- Learning parameters of a Markov Network
  - More complex than for BN as likelihood doesn't factor out
- Structure Learning
- Learning with incomplete data
  - We will do the last one first, to enable our last HW assignment

Partially Observed Data

- It is common that the sample data will not have values for all the variables in a network
  - In medical diagnosis, one does not perform all the tests
  - In speech recognition, we may not have annotations of the "phones"
  - *Hidden* nodes whose values are *never* observed
  - Data *Missing Completely At Random* (MCAR),
  - Whether data is missing or not in not dependent on the values of the variables in the data
  - *e.g.* coin in coin toss falls off of a table; patient records lost…
  - For MCAR, we can just ignore missing data for computing the probability distribution
    - We can also compute the probability of missing data incidents (how often coin falls off the table).

MNAR and MAR

- Missing Not At Random (MNAR)
  - Whether data is missing or not depends on the values of the unobserved variables: for example
    - Observer does not favor “heads” so does not report them
    - In a medical trial, patients not deriving benefit drop out
  - Difficult to detect MNAR or to train a model for it
- Missing At Random (MAR)
  - Less restrictive than MCAR but also not MNAR
  - Given the observed values, missing values do not depend on the unobserved values
    - e.g. physician does not order EKG for a patient who comes reporting a broken leg (missing EKG based on observation)
  - With MAR, we can still maximize likelihood of observable variables to learn model parameters
- We consider methods that apply when MAR holds

Likelihood Function with Missing Data

- Consider a network G with set of variables **X**
- At m[th] instance, let **O**[m] be the observed variables with values **o**[m]; **H**[m] be the set of missing or hidden variables
- $L(\theta : D) = P(D|\theta) = \sum_H P(D, H : \theta) = \prod_{m=1,M} P(o[m] \mid \theta)$
  $= \prod_{m=1,M} \sum_{h[m]} P(o[m],h[m] : \theta)$ for *iid* samples
- Consider a chain A → B → C; let B be "hidden"
  - Suppose we observe a^0 and c^0
  - $P(a^0,c^0) = \sum_B P(a^0,B,c^0)$
    $= P(a^0) \{P(b^0|a^0) P(c^0|b^0) + P(b^1|a^0) P(c^0|b^1)\}$
  - Probability of *k* such samples is $P(a^0,c^0)^k$
  - likelihood function over all samples is a product of such terms for various assignments of A and C
    - Each term in product, however, is some of other likelihood functions, so when we take a log, terms do not separate out

Next Class

- Read sections 19.1 and 19.2 of the KF book