Primer: Structured Probability Distributions

Note Title 4/2/2006

Fundamentals of Bayesian Inference.

Two variables. joint probability $P(a,b)$

conditional prob $p(a|b)$ & $p(b|a)$

marginal probs $p(a)$ & $p(b)$

$$P(a,b) = P(a|b)P(b) = P(b|a)P(a)$$

Implies Bayes Rule

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Replace a by d (data)

b by $h \in \mathcal{H}$ set of hypotheses.

Hypothesis Space.

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}$$

Compare Hypotheses.

E.G. Are data tosses from fair coin ($Pr(H) = 0.5$)

or a biased coin ($Pr(H) = 0.9$)

If HH....H - probably think biased.

HHTHTHTTH - probably think fair.

(Page 2) To address this problem formally
Let θ be the probability that the coin produces head.

Hypothesis

h_0 : $\theta = 0.5$ unbiased

h_1 = $\theta = 0.9$ biased.

Hypothesis Space $\mathcal{H} = \{h_0, h_1\}$

$$P(d|\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

N_H no. of heads in data d

N_T no. of tails " " "

Posterior odds of the hypotheses

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)\,P(h_1)}{P(d|h_0)\,P(h_0)}$$

Gives 357:1 in favor of h_0 from HH ... H

165:1 in favor of h_1 from HHTHTHTTHT..

Comparing Infinitely Many Hypotheses

Suppose $0 \le \theta \le 1$.

$$p(\theta|d) = \frac{P(d|\theta)\,p(\theta)}{P(d)}$$ ← prior

$$p(d) = \int_0^1 P(d|\theta)\,p(\theta)\,d\theta.$$

(3.) Example: If $p(\theta) = 1$.

$$p(\theta|d) = \frac{(N_H+N_T+1)!}{N_H! N_T!}\theta^{N_H}(1-\theta)^{N_T}$$

Beta (N_H+1, N_T+1)

If $p(\theta) = \text{Beta}(V_H+1, V_T+1)$ V_H & V_T positive integer.

$$p(\theta|d) = \frac{(N_H+N_T+V_H+V_T+1)!}{(N_H+V_H)!(N_T+V_T)!}\theta^{N_H+V_H}(1-\theta)^{N_T+V_T}$$

Beta (N_H+V_H+1, N_T+V_T+1).

Comparing Hypotheses of Different Complexity.
~Model Selection.

h_0: hypothesis $\theta = 0.5$

h_1: hypothesis θ drawn from uniform distribution on θ.

$$p(d|h_0) = (0.5)^{N_H+N_T}$$

$$p(d|h_1) = \int_0^1 p(d|\theta, h_1)\, p(\theta|h_1)\, d\theta$$

$p(\theta|h_1) = 1$.

$$= \frac{N_H! N_T!}{(N_H+N_T+1)!}$$

Can apply Bayes rule as before:

Important. Occam's razor

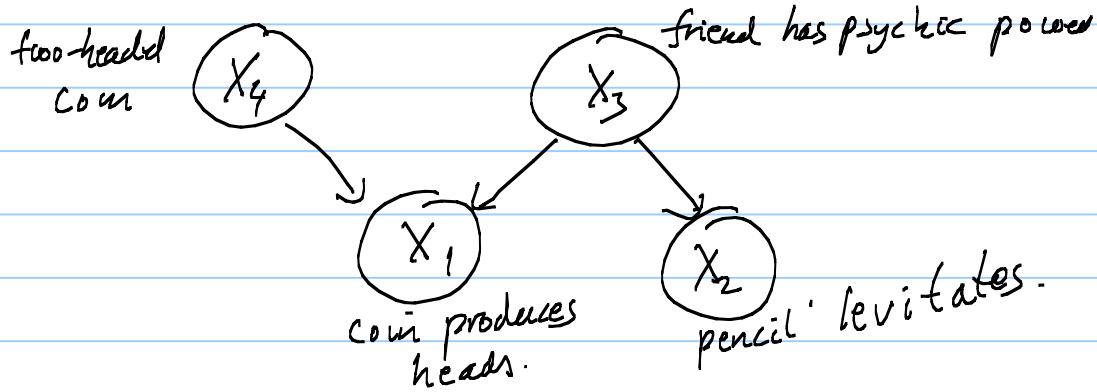
complex hypotheses have more degrees of freedom & can be adapted to data. But integrating over θ prevents this

(4) Representing Structured Probability Distributions

Probabilistic models can define the joint distribution for a set of random variables.

Friend claims psychic powers — test on coin tossing
— test on pencil levitating

X_1 represent truth of coin being flipped to give heads
X_2 " pencil levitating.
X_3 " friend having psychic powers
X_4 " use of a "two-headed coin.



Represent the joint distribution

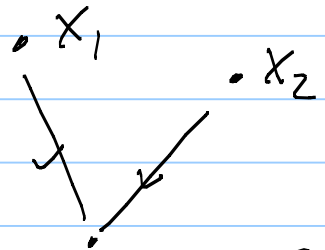
$$P(X_1, X_2, X_3, X_4) = P(X_1 | X_3, X_4) P(X_2 | X_3) p(X_3) p(X_4)$$

Markov Property: Graph represents the dependencies between variables. Direct & Indirect Relationships.
If you know X_3, the knowing X_4, won't give info on X_2
But, if you don't know X_3 - it will.

(Page 5)

Simple Example: Deterministic Special Case.

2 Gamma to 1 Beta



$P_\gamma(x_1)$
$P_\gamma(x_2)$

Gamma Distributions.

Delta Function.

$$P(z|x_1,x_2) = \delta\left(z - \frac{x_1}{x_1+x_2}\right)$$

$$P(x_1,x_2,z) = P(z|x_1,x_2)\,P_\gamma(x_1)\,P_\gamma(x_2).$$

Marginal

$$P(z) = \int dx_1 \int dx_2\, P(x_1,x_2,z)$$

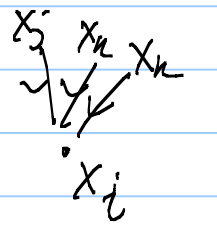
$\propto$ Beta Distribution

Relates to "special relationship" between Beta & Gamma for Sampling (see lecture 2)

Directed Bayes Network

$$P(X_1, X_2 \ldots, X_n) = \prod_i P(X_i \mid Pa(X_i))$$

where $Pa(X_i)$ are the parents of X_i



Markov Condition

Conditioned on parents, each variable is independent of all other variables except its descendents

Factorization → allows us to use fewer numbers than directly specifying the full distribution $p(X_1 \ldots X_n) = 2^n - 1$

E.g. for psychic problem.

Need to specify 8 numbers, not $2^4 - 1 = 15$

Computations are also simplified by exploiting the structure.

Eg. $$P(X_1=1) = \sum_{X_2}\sum_{X_3}\sum_{X_4} P(X_1=1, X_2, X_3, X_4)$$

$$= \sum_{X_2}\sum_{X_3}\sum_{X_4} P(X_1=1|X_3, X_4)\, P(X_2|X_3) P(X_3) P(X_4)$$

$$= \sum_{X_3}\sum_{X_4} P(X_1=1|X_3, X_4)\, P(X_3) P(X_4).$$

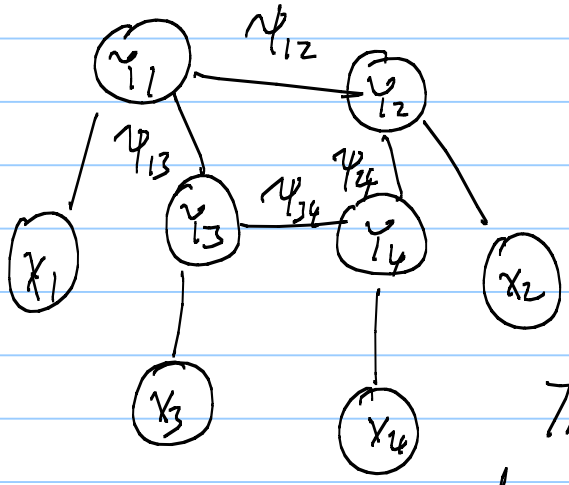
Sum over X_2 can be done automatically.

If the graph has no closed loops (tree) then dynamic programming (DP) can be used. (later in course).

Directed Graphs have many uses in Artificial Intelligence & Statistics Communities.

Undirected Graphical Models

Markov Random Fields (MRFs)



Undirected edges depict neighbourhood structures on the graph.

These indicate the probability dependencies Markov Condition

Each fully connected neighbours is associated with a potential function

The distribution is the product of the potential functions:

$$p(\underline{x}|\underline{y}) = \prod_i p(x_i|y_i)\, p(\underline{y})$$

$$\text{with} \quad p(\underline{y}) = \frac{1}{Z} \prod_{i,j \in \mathcal{N}} \psi_{ij}(y_i, y_j) \prod_i \psi_i(x_i)$$

Note: Potentials not probabilities (unlike Directed Graphs).

Example: Ising Spin Model.

$X_1 \; X_2 \; X_3 \; \cdots \; X_N$ $\quad X_i \in \{-1, 1\}$

$$E[\underline{X}] = -J \sum_{i=1}^{N} X_i X_{i+1}$$

If $X_i = X_{i+1}$ then $X_i X_{i+1} = 1$ $\quad \{(1)^2 = 1, \; (-1)^2 = 1$

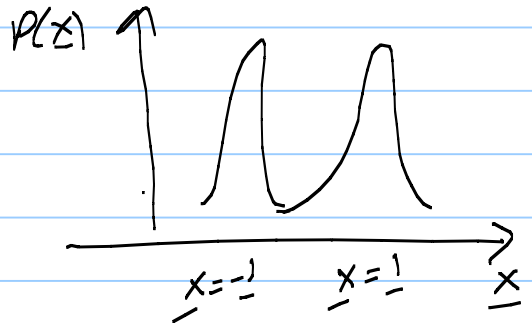
$$P[\underline{X}] = \frac{1}{Z[J]} e^{J \sum_{i=1}^{N} X_i X_{i+1}} = \frac{1}{Z} e^{-E[\underline{X}]}$$

Most probable states are lowest energy

$X_1 = X_2 = \ldots = X_N = 1 \qquad \underline{X} = \underline{1}$

or $X_1 = X_2 = \ldots = X_N = -1 \qquad \underline{X} = -\underline{1}$

The heights of the peaks depend on $J = 1/T$
Large J means sharp peaks
Small J means soft peaks.



Properties. $\mu = \frac{1}{N} \sum_{i=1}^{N} X_i$ (net magnetization)

Physicists like to study the average magnetization $\sum_{\underline{x}} \mu(\underline{x}) P[\underline{x}]$

It has a phase transition at critical J.
Behaviour is quite different below J_c and above J_c

Ising to Potts Model.

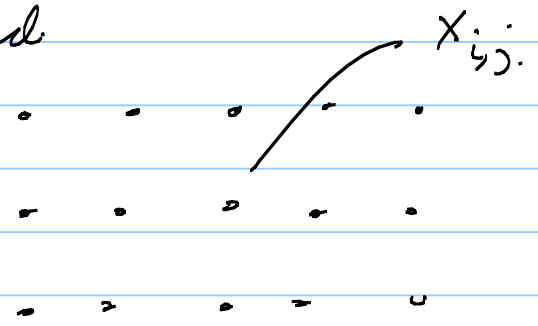
Extend to Potts Model. $x_i \in \{1, 2, \ldots M\}$

$$E[\underline{x}] = -J \sum_i \varphi(x_i, x_{i+1})$$

E.G. (Kronecker delta)

$$\varphi(x_i, x_{i+1}) = \delta_{x_i, x_{i+1}}$$

Can be two-dimensional



$$E[\underline{x}] = -J \sum_{ij} \left\{ \delta_{x_{ij}, x_{i,j+1}} + \delta_{x_{ij}, x_{i+1,j}} \right\}$$

Many applications of Potts models
- vision, encoding/decoding, etc.

Page 11

Hidden Markov Models (HMM)

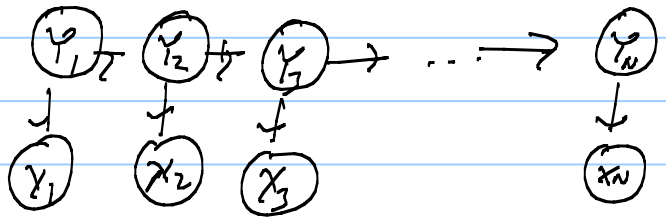
Used for speech and language processing

Sequence of T observations $\{x_t : t=1,\dots T\}$ generated by hidden states $\{y_t : t=1\dots T\}$

Joint Distribution:

$$P(\{y_t\},\{x_t\},w) = P(w)\,P(y_1|w)\,P(x_1|y_1,w) \prod_{t=2}^{T} P(y_t|y_{t-1},w)\,P(x_t|y_t,w).$$

word w.



Applying HMM's to recognize words requires

DP algorithms to: (i) learn $P(x_t|y_t,w)$ & $P(y_t|y_{t-1},w)$ for each w.

(ii) evaluate the probability

$$P(\{x_t\},w) = \sum_{\{y_t\}} P(\{y_t\},\{x_t\},w)$$ for each word

(iii) estimate $w^* = \arg\max_w \sum_{\{y_t\}} P(\{y_t\},w \mid \{x_t\})$

Probabilistic Context Free Grammars. (PCFG)

Define non-terminal nodes

S, NP, VP, AT, NNS, VBD, PP, IN, DT, NN

where S is a sentence.

VP is a verb phrase . . .

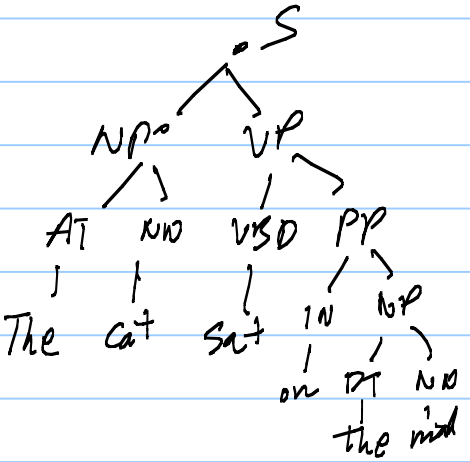
Terminal nodes are words from a dictionary

(eg. "the" "cat" "sat" "on" "the" "mat").

Define production rules which are applied to non-terminal nodes to generate child nodes

eg. S → NP, VP or NN → "cat"

Define probability distribution for the production rules.



Generate a sentence by starting with node S, and sampling the production rules.

Parse an input sentence by choosing the most probable parse tree.

Learn probabilities of rule

DP update.

Inference Algorithms

Want to infer values of latent/hidden variables conditioned on data $p(\underline{y}|\underline{x})$, calculate expectations, or sum out the hidden variables

$$p(\underline{x}|\theta) = \sum_{\underline{y}} p(\underline{x}, \underline{y}|\theta)$$

Expectation Maximization (EM) Algorithm
Markov Chain Monte Carlo (MCMC)
Dynamic Programming (trees only)

This will be covered later in the course.

Converting Undirected Graphs to Directed.

If the graph structure is a tree (no closed loops) then it is possible to convert an undirected graph to a directed graph.

If graph structure has closed loops, then conversion is possible by augmenting variables. But may not be worthwhile

Decision Theory & Control Theory.

Bayes Decision Theory introduces loss function $L(h, \alpha(d))$ for cost of making decision $\alpha(d)$ when input is d and true hypothesis is h.

Select decision rule $\alpha^*(\cdot)$ that minimizes risk, or expected loss,

$$R(\alpha) = \sum_{h,d} L(h, \alpha(d)) P(h,d)$$

Basis of rational decision making

Loss function often set $L(h, \alpha(d)) = 1$, if $\alpha(d) \neq h$.
$L(h, \alpha(d)) = 0$, if $\alpha(d) = h$

Then best decision rule is
maximum a posteriori (MAP)

$$\alpha^*(d) = \text{ARG MAX } P(h|d)$$

If $L(h, \alpha(d)) = (h - \alpha(d))^2$
then $\alpha^*(d) = \sum_h h P(h|d)$.

Can extend Bayes risk to dynamical systems whose decisions need to be made over time (ives. optimal control theory.