Primer: Structured Probability Distributions

Fundamentals of Bayesian Inference.

Two variables, joint probability
\[ P(a, b) \]

conditional prob
\[ P(a|b) \]
marginal probs
\[ P(a) \]

\[ P(a, b) = P(a|b)P(b) = P(b|a)P(a) \]

Implies Bayes Rule

\[ P(b|a) = \frac{P(a|b)P(b)}{P(a)} \]

Replace \( a \) by \( d \) (data)

\( b \) by \( h \) \in \text{set of hypotheses}

\[ P(h|d) = \frac{P(d|h)P(h)}{\sum P(d|h')P(h')} \]

Hypothesis Space.

Compare hypotheses.

E.g., are data tosses from fair coin (\( P(H) = 0.5 \)) or a biased coin (\( P(H) = 0.9 \))

If \( HH...H \) - probably think biased.

\( HHTHTHTH \) - probably think fair.
To address this problem formally, let \( \theta \) be the probability that the coin produces heads.

**Hypothesis**
- \( h_0 : \theta = 0.5 \) unbiased
- \( h_1 : \theta = 0.9 \) biased

**Hypothesis Space** \( \mathcal{H} = \{ h_0, h_1 \} \)

\[
P(d|h) = \Theta_{\mathcal{H}}(h | d)^N
\]

\( N_H \) no. of heads in data \( d \)
\( N_T \) no. of tails in \( d \)

**Posterior odds of the hypotheses**
\[
P(h_1 | d) = \frac{P(d | h_1) P(h_1)}{P(d | h_0) P(h_0)}
\]

Gives \( 357:1 \) in favor of \( h_0 \) from \( HHH \ldots \).

\( 165:1 \) in favor of \( h_1 \) from \( HTHTHTHTHTHTHTTT \).

Company indefinitely many hypotheses.

Suppose \( 0 \leq \theta \leq 1 \).

**Prior**
\[
P(\theta | d) = P(d | \theta) p(\theta)
\]

**Posterior**
\[
P(d) = \int \frac{1}{p(d)} P(d | \theta) p(\theta) d\theta
\]
\( \text{(3)} \)

Example: \( \forall \theta \in \mathcal{D} \), \( P(\theta) = 1 \).

\[
P(\theta | d) = \frac{\binom{N_{\theta} + N_T + 1}{N_{\theta}} \theta^{N_{\theta}} (1-\theta)^{N_T}}{N_{\theta}! N_T!}
\]

\( \text{Beta}(N_{\theta}+1, N_T+1) \)

\( \text{if} \ P(\theta) = \text{Beta}(V_{\theta}+1, V_T+1) \quad V_{\theta} \& V_T \)

\( \text{positive integer} \)

\[
P(\theta | d) = \frac{\binom{N_{\theta} + N_T + V_{\theta} + V_T + 1}{(N_{\theta} + V_{\theta})! (N_T + V_T)!} \theta^{N_{\theta} + V_{\theta}} (1-\theta)^{N_T + V_T}}{(N_{\theta} + V_{\theta})! (N_T + V_T)!}
\]

Company Hypotheses of Different Complexity.

Model Selection:

\( \text{h}_0 : \) hypothesis \( \theta = 0.5 \)

\( \text{h}_1 : \) hypothesis \( \theta \) drawn from uniform distribution

\[
P(d | h_0) = (0.5)^{N_{\theta} + N_T}
\]

\[
P(d | h_1) = \int_0^1 P(d | \theta, h_1) P(\theta | h_1) d\theta = N_{\theta}! N_T! \]

\[
\frac{(N_{\theta} + N_T + 1)!}{(N_{\theta} + V_{\theta})! (N_T + V_T)!}
\]

\[
P(\theta | h_1) = 1.
\]

Can apply Bayes rule as before:

Important: Occam's razor

\text{complex hypotheses have more degrees of freedom & can be adapted to data. But integrating over } \theta \text{ prevents this.}
(4) Representing Structural Probability Distribution

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

Friend claims psychic power — test on coin tossing
test on pencil levitation

\[ P(X_1, X_2, X_3, X_4) = P(X_1 | X_3, X_4) P(X_2 | X_3, p(X_3)p(X_4)) \]

Markov property: Graph represents the dependencies between variables. Direct & Indirect Relationship.

If you know \( X_3 \), the knowing \( X_4 \) can't give info on \( X_1 \).
But, if you don't know \( X_3 \), it will.
Simple Example: Deterministic Special Case.

2 Gamma to 1 Beta

\[ P(x_1) \]
\[ P(x_2) \]
\[ \delta \left( z - \frac{x_1}{x_1 + x_2} \right) \]

\[ P(x_1, x_2, z) = P(z | x_1, x_2) P(x_1) P(x_2) \]

Marginal

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

Beta Distribution

Relates to "special relationship" between Beta & Gamma for sampling (see lecture 2)
Directed Bayes Network.

\[ P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | \text{Pa}(x_i)) \]

Where \( \text{Pa}(x_i) \) are the parents of \( x_i \).

\[ \forall x_i \in \text{Pa}(x_i) \]

\[ x_i \text{ Markov Condition} \]

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

Factorization allows us to use fewer numbers than directly specifying the full distribution \( P(x_1, x_2, \ldots, x_n) \). For example, in a psychic problem:

\[ \text{Need to specify 8 numbers, not } 2^4 - 1 = 15 \]
Computations are also simplified by exploiting the structure.

Example: \( P(X_1 = 1) = \sum_{x_2} \sum_{x_3} \sum_{x_4} P(X_1 = 1, x_2, x_3, x_4) \)

\[ = \sum_{x_2} \sum_{x_3} \sum_{x_4} P(X_1 = 1 \mid x_2, x_3, x_4) P(x_2, x_3, x_4) \]

\[ = \sum_{x_2} \sum_{x_3} \sum_{x_4} P(X_1 = 1 \mid x_2, x_3, x_4) P(x_2) P(x_3) P(x_4). \]

Sum over \( x_2 \) can be done automatically.

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

(\text{later in course)\}.

\text{Directed Graphs have many uses in Artificial Intelligence & Statistics Community.}
Undirected Graphical Models

Markov Random Fields (MRFs)

Undirected edges define neighborhood structure on the graph.

These indicate the probability dependencies Markov Condition.

Each fully connected neighbors is associated with a potential function.

The distribution is the product of the potential functions:

\[
p(x|y) = \prod_i p(x_i|y_i) p(y_i)
\]

with

\[
p(y_i) = \frac{1}{Z} \prod_{j \in \text{neighbors}(i)} \psi_j(y_i, y_j)
\]

Note: Potentials not probabilities (unlike Directed Graphs).
Example: Ising Spin Model.

\[ E[X] = -J \sum_{i=1}^{N} x_i x_{i+1} \]

\[ P[X] = \frac{1}{Z} e^{-\frac{E[X]}{T}} = \frac{1}{Z} e^{-E[X]} \]

Most probable states are lowest energy

- \[ x_1 = x_2 = \ldots = x_N = 1 \]
- \[ x_1 = x_2 = \ldots = x_N = -1 \]

The heights of the peaks depend on \( J = K \)

Large J means sharp peaks

Small J means soft peaks.

Property: \[ \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \] (not magnetization).

Physicists like to study the average magnetization \[ \frac{1}{N} \sum_{i=1}^{N} P[X] \]

It has a phase transition at critical J. Behaviour is quite different below \( J_c \) and above \( J_c \)
Using to Potts Model.

Extend to Potts Model. \( x_i \in \{1, 2, \ldots, m \} \)

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

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

Can be two-dimensional

\[
\mathbb{E}[x] = -J \sum_{i,j} \left( \delta_{x_i, x_j} \times x_{i+1,j} 
+ \delta_{x_{i,j}, x_{i+1,j}} \right)
\]

Many applications of Potts models

- Vision, encoding/decoding, etc.
Hidden Markov Models (HMM) are used for speech and language processing. A sequence of \( T \) observations \( \langle x_t : t=1, \ldots, T \rangle \) are generated by hidden states \( \langle y_t : t=1, \ldots, T \rangle \).

The joint distribution is:

\[
P(\langle x_t, y_t \rangle : 1 \leq t \leq T, w) = P(w) P(y_1 \mid w) P(x_1 \mid y_1, w) \prod_{t=2}^{T} P(y_t \mid y_{t-1}, w) P(x_t \mid y_t, w).
\]

Applying HMM to recognize words requires an algorithm to:

1. Learn \( P(x_t \mid y_t, w) \) and \( P(y_t \mid y_{t-1}, w) \) for each \( w \).

2. Evaluate the probability:

\[
P_{\text{log}}(\langle x_t \rangle, w) = \sum_{\langle y_t \rangle} P(\langle y_t \rangle, \langle x_t \rangle, w)
\]

3. Estimate \( w^* = \arg \max_w \sum_{\langle y_t \rangle} P(\langle y_t \rangle \mid \langle x_t \rangle, w) \).
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. \( S \rightarrow NP, VP \) or \( NN \rightarrow "cat" \)).

Design probability distributions for the production rules.

\[ S \rightarrow NP, VP \]

\[ VP \rightarrow V \]

\[ V \rightarrow "sat" \]

\[ TP \rightarrow DT, NN, IN, PP \]

\[ PP \rightarrow "on", "at" \]

\[ DT \rightarrow "the", "a" \]

\[ NN \rightarrow "cat", "mat" \]

Generate a sentence by selecting with node \( S \), and sampling the production rules.

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

Learn probabilities of rules.
Inference Algorithms

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

\[
p(x|\theta) = \sum_y p(x,y|\theta)
\]

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

Select decision rule $d^*(x)$ that minimizes risk or expected loss,

$$R(x) = \sum_h L(h|x,d) \cdot P(h|d)$$

Basis of rational decision making.

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

Then best decision rule is maximum a posteriori (MAP)

$$d^*(x) = \arg \max P(h|x,d)$$

If $L(h|x,d) = (h - d(x))^2$ then

$$d^*(x) = \sum_h h \cdot P(h|x,d)$$

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