Lecture Slides for

INTRODUCTION TO

Machine Learning
2nd Edition

ETHEM ALPAYDIN
© The MIT Press, 2010

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e
Bayes Nets Quick Intro

- Topic of much current research
- Models dependence/independence in probability distributions
- Graph based - aka “graphical models” “belief nets”
- Two kinds - directed and undirected

Any joint distribution $P(X,Y,Z)$ can be factored:

$$P(X,Y,Z) = P(X) \ P(Y \mid X) \ P(Z \mid X,Y)$$

$$P(X=x, Y=y, Z=z) = P(X=x) \ P(Y=y \mid X=x) \ P(Z=z \mid X=x, Y=y)$$
Basic idea:

- Model a joint distribution as a DAG with attributes (features, RVs) at the nodes
- Sample distribution from sources
Graphical Models

- Aka Bayesian networks, probabilistic networks
- Concise way of represented complicated distributions with limited dependence
- *Nodes* are random variables
- *Arcs* are direct dependencies between random variables
- The *structure* is represented as a directed acyclic graph (DAG)
- The *parameters* are the conditional probabilities at the nodes (Pearl, 1988, 2000; Jensen, 1996; Lauritzen, 1996)
"Causes" and Bayes’ Rule

**Diagnostic inference:** Knowing that the grass is wet, what is the probability that rain is the “cause”?

\[
P(R \mid W) = \frac{P(W \mid R)P(R)}{P(W)} = \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \neg R)P(\neg R)}
\]

\[
= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\]
Conditional Independence

- $X$ and $Y$ are **independent** if
  \[ P(X,Y) = P(X)P(Y) \]

- $X$ and $Y$ are **conditionally independent given $Z$** if
  \[ P(X,Y|Z) = P(X|Z)P(Y|Z) \]
  or equivalently (why?)
  \[ P(X|Y,Z) = P(X|Z) \]

- Consider “stop at store”, “got milk”, “got candy”
Conditional Independence

- Conditional independence modelled by graph structure

- Three canonical cases: Head-to-tail, Tail-to-tail, head-to-head
Case 1: Head-to-Tail

\[ P(X,Y,Z) = P(X)P(Y|X)P(Z|Y) \]

- \[ P(\text{Cloudy}) = 0.4 \]
  - \[ P(R | \text{Cloudy}) = 0.8 \]
  - \[ P(R | \sim\text{Cloudy}) = 0.1 \]
  - \[ P(W | R) = 0.9 \]
  - \[ P(W | \sim R) = 0.2 \]

\[ P(W | C) = P(W | R)P(R | C) + P(W | \sim R)P(\sim R | C) \]
Case 2: Tail-to-Tail

- \( P(X,Y,Z) = P(X)P(Y|X)P(Z|X) \)
Case 3: Head-to-Head

- \[ P(X,Y,Z) = P(X)P(Y)P(Z|X,Y) \]

\[ P(W) = 0.52 \]
Causal vs Diagnostic Inference

Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

\[
P(W|S) = P(W|R,S) \cdot P(R|S) + P(W|\neg R,S) \cdot P(\neg R|S)
\]

\[
= P(W|R,S) \cdot P(R) + P(W|\neg R,S) \cdot P(\neg R)
\]

\[
= 0.95 \cdot 0.4 + 0.9 \cdot 0.6 = 0.92
\]

-- Sum over the unknowns
Causal vs Diagnostic Inference

Diagnostic inference: If the grass is wet, what is probability that the sprinkler is on?

\[ P(S \mid W) = \frac{P(W \mid S) \cdot P(S)}{P(W)} \]

\[ P(S \mid W) = \frac{0.184}{0.52} = 0.35 \]

and 0.35 is more than \( P(S) \)

Explaining away: Also knowing that it has rained decreases the probability that the sprinkler is on.

\[ P(S \mid R, W) = \frac{P(S \mid W, R)}{P(W \mid R)} \]

\[ = \frac{P(W \mid S, R) \cdot P(S \mid R) / P(W \mid R)}{0.95 \cdot 0.2 / 0.91} \approx 0.21 \]
Causes

Causal inference:

\[ P(W | C) = P(W | R, S) \cdot P(R, S | C) + P(W | \sim R, S) \cdot P(\sim R, S | C) + P(W | R, \sim S) \cdot P(R, \sim S | C) + P(W | \sim R, \sim S) \cdot P(\sim R, \sim S | C) \]

and use the fact that

\[ P(R, S | C) = P(R | C) \cdot P(S | C) \]

Diagnostic: \( P(C | W) = ? \)
Exploiting the Local Structure

\[ P(C) = 0.5 \]

\[ P(S \mid C) = 0.1 \]
\[ P(S \mid \neg C) = 0.5 \]

\[ P(R \mid C) = 0.8 \]
\[ P(R \mid \neg C) = 0.1 \]

\[ P(W \mid R, S) = 0.95 \]
\[ P(W \mid R, \neg S) = 0.90 \]
\[ P(W \mid \neg R, S) = 0.90 \]
\[ P(W \mid \neg R, \neg S) = 0.10 \]

\[ P(F \mid R) = 0.1 \]
\[ P(F \mid \neg R) = 0.7 \]

\[ P(C, S, R, W, F) = P(C)P(S \mid C)P(R \mid C)P(W \mid S, R)P(F \mid R) \]

\[ P(X_1, \ldots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents} \ (X_i)) \]
Bayes’ rule inverts the arc:

\[ P(C \mid x) = \frac{p(x \mid C)P(C)}{p(x)} \]
Given $C$, $x_j$ are independent:

$$p(x | C) = p(x_1 | C) \times p(x_2 | C) \times ... \times p(x_d | C)$$
Hidden Markov Model as a Graphical Model

\[ \pi = P(q^1) \]

\[ \begin{align*}
  q^1 & \xrightarrow{} & \pi \\
  O^1 & \xrightarrow{} &
\end{align*} \]

\[ A = P(q_t \mid q_{t-1}) \]

\[ \begin{align*}
  q^{t-1} & \xrightarrow{} & A \\
  O^{t-1} & \xrightarrow{} &
\end{align*} \]

\[ B = P(O_t \mid q_t) \]

\[ \begin{align*}
  q^t & \xrightarrow{} & B \\
  O^t & \xrightarrow{} &
\end{align*} \]
Linear Regression

\[
p(r'|x', r, X) = \int p(r'|x', w)p(w|X, r)dw
\]

\[
= \int p(r'|x', w)\frac{p(r|X, w)p(w)}{p(r)}dw
\]

\[
\propto \int p(r'|x', w)\prod_t p(r^t|x^t, w)p(w)dw
\]
d-Separation

- A path from node $A$ to node $B$ is blocked if
  a) The directions of edges on the path meet head-to-tail (case 1) or tail-to-tail (case 2) and the node is in $C$, or
  b) The directions of edges meet head-to-head (case 3) and neither that node nor any of its descendants is in $C$.

- If all paths are blocked, $A$ and $B$ are d-separated (conditionally independent) given $C$.

Given:
- $BCDF$ is blocked given $C$.
- $BEFG$ is blocked by $F$.
- $BEFD$ is blocked unless $F$ (or $G$) is given.
Belief Propagation (Pearl, 1988)

- Chain:

\[
P(X | E) = \frac{P(E \mid X)P(X)}{P(E)} = \frac{P(E^+, E^- \mid X)P(X)}{P(E)}
\]

\[
= \frac{P(E^+ \mid X)P(E^- \mid X)P(X)}{P(E)} = \alpha \pi(X) \lambda(X)
\]

\[
\pi(X) = \sum_U P(X \mid U) \pi(U)
\]

\[
\lambda(X) = \sum_Y P(Y \mid X) \lambda(Y)
\]
Trees

\[ \lambda(X) = P(E_X^- \mid X) = \lambda_Y(X) \lambda_Z(X) \]
\[ \lambda_X(U) = \sum_X \lambda(X) P(X \mid U) \]

\[ \pi(X) = P(X \mid E_X^+) = \sum_U P(X \mid U) \pi_X(U) \]
\[ \pi_y(X) = \alpha \lambda_Z(X) \pi(X) \]
Polyltrees

\[
\pi(X) = P(X \mid E_X^+) = \sum_{U_1} \sum_{U_2} \cdots \sum_{U_k} P(X \mid U_1, U_2, \ldots, U_k) \prod_{i=1}^{k} \pi_X(U_i)
\]

\[
\pi_{y_j}(X) = \alpha \prod_{s \neq j} \lambda_{y_s}(X) \pi(X)
\]

\[
\lambda_X(U_i) = \beta \sum_{X} \lambda(X) \sum_{U_{r \neq i}} P(X \mid U_1, U_2, \ldots, U_k) \prod_{r \neq i} \pi_X(U_r)
\]

\[
\lambda(X) = \prod_{j=1}^{m} \lambda_{y_j}(X)
\]

How can we model \(P(X \mid U_1, U_2, \ldots, U_k)\) cheaply?
Junction Trees

- If $X$ does not separate $E^+$ and $E^-$, we convert it into a junction tree and then apply the polytree algorithm.
Undirected Graphs: Markov Random Fields

- In a Markov random field, dependencies are symmetric, for example, pixels in an image.
- In an undirected graph, $A$ and $B$ are independent if removing $C$ makes them unconnected.
- **Potential function** $\psi_c(X_c)$ shows how favorable is the particular configuration $X$ over the clique $C$.
- The joint is defined in terms of the clique potentials

$$p(X) = \frac{1}{Z} \prod_c \psi_c(X_c) \text{ where normalizer } Z = \sum_X \prod_c \psi_c(X_c)$$
Factor Graphs

- Define new factor nodes and write the joint in terms of them

\[ p(X) = \frac{1}{Z} \prod_s f_s(X_s) \]
Learning a Graphical Model

• Learning the *conditional probabilities*, either as tables (for discrete case with small number of parents), or as parametric functions

• Learning the *structure* of the graph: Doing a state-space search over a *score function* that uses both goodness of fit to data and some measure of complexity
Influence Diagrams

- **Decision node**: choose class
- **Chance node**: $x$
- **Utility node**: $U$