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
Graphical Models

- Aka Bayesian networks, probabilistic networks
- **Nodes** are hypotheses (random vars) and the probabilities corresponds to our belief in the truth of the hypothesis
- **Arcs** are direct influences between hypotheses
- The **structure** is represented as a directed acyclic graph (DAG)
- The **parameters** are the conditional probabilities in the arcs (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?

\[
P(R | W) = \frac{P(W | R)P(R)}{P(W)}
\]

\[
= \frac{P(W | R)P(R)}{P(W | R)P(R) + P(W | \sim R)P(\sim R)}
\]

\[
= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\]
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
  \[ P(X|Y,Z) = P(X|Z) \]
- Three canonical cases: Head-to-tail, Tail-to-tail, head-to-head
Case 1: Head-to-Head

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

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

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

- $P(X, Y, Z) = P(X)P(Y)P(Z | X, Y)$
Causal vs Diagnostic Inference

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

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

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

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

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

\[
P(S|W) = 0.35 > 0.2 \cdot P(S)
\]

\[
P(S|R,W) = 0.21
\]

Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.
Causes

Causal inference:

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

and use the fact that

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

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

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

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

\[
P(C \mid x) = \frac{p(x \mid C)P(C)}{p(x)}
\]
Naive Bayes’ Classifier

Given $C, x_j$ are independent:

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

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

\[ q^1 \]

\[ O^1 \]

\[ A = P(q^t | q^{t-1}) \]

\[ q^{t-1} \]

\[ O^{t-1} \]

\[ B = P(O^t | q^t) \]

\[ q^t \]

\[ O^t \]
(a) Input-output HMM

(b) Factorial HMM

(c) Coupled HMM

(d) Switching HMM
Linear Regression

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

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

\[
\propto \int p(r' \mid x', w) \prod_t p(r^t \mid x^t, w)p(w)dw
\]
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 $C$.

$BCDF$ 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

\[ \lambda(X) = P(E_X^- | X) = \lambda_y(X) \lambda_z(X) \]

\[ \lambda_x(U) = \sum_X \lambda(X) P(X | U) \]

\[ \pi(X) = P(X | E^+_X) = \sum_U P(X | U) \pi_x(U) \]

\[ \pi_y(X) = \alpha \lambda_z(X) \pi(X) \]
Polytrees

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

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

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

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

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.

Tree of moralized, clique nodes
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

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