Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN
© The MIT Press, 2004

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml
CHAPTER 3: Bayesian Decision Theory
Probability and Inference

- Result of tossing a coin is ∈ {Heads,Tails}
- Random var $X ∈ \{1,0\}$
  Bernoulli: $P\{X=1\} = p_o^X(1 - p_o)^{(1 - X)}$
- Sample: $X = \{x_t\}_{t=1}^N$
  Estimation: $p_o = \# \{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t x_t / N$
- Prediction of next toss:
  Heads if $p_o > \frac{1}{2}$, Tails otherwise
Classification

- Credit scoring: Inputs are income and savings. Output is low-risk vs high-risk
- Input: $x = [x_1, x_2]^T$, Output: $C \in \{0,1\}$
- Prediction:

  choose \[
  \begin{cases} 
  C = 1 \text{ if } P(C = 1 \mid x_1, x_2) > 0.5 \\
  C = 0 \text{ otherwise}
  \end{cases}
  \]

  or equivalently

  choose \[
  \begin{cases} 
  C = 1 \text{ if } P(C = 1 \mid x_1, x_2) > P(C = 0 \mid x_1, x_2) \\
  C = 0 \text{ otherwise}
  \end{cases}
  \]
Bayes’ Rule

\[ P(C \mid x) = \frac{P(C)p(x \mid C)}{p(x)} \]

\[ P(C = 0) + P(C = 1) = 1 \]

\[ p(x) = p(x \mid C = 1)P(C = 1) + p(x \mid C = 0)P(C = 0) \]

\[ p(C = 0 \mid x) + P(C = 1 \mid x) = 1 \]
Bayes’ Rule: $K>2$ Classes

\[
P(C_i \mid x) = \frac{p(x \mid C_i)P(C_i)}{p(x)}
= \frac{p(x \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(x \mid C_k)P(C_k)}
\]

$P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

choose $C_i$ if $P(C_i \mid x) = \max_k P(C_k \mid x)$
Losses and Risks

- Actions: $\alpha_i$
- Loss of $\alpha_i$ when the state is $C_k: \lambda_{ik}$
- Expected risk (Duda and Hart, 1973)

$$R(\alpha_i \mid x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid x)$$

choose $\alpha_i$ if $R(\alpha_i \mid x) = \min_k R(\alpha_k \mid x)$


Losses and Risks: 0/1 Loss

\[ \lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases} \]

\[
R(\alpha_i \mid x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid x) \\
= \sum_{k \neq i} P(C_k \mid x) \\
= 1 - P(C_i \mid x)
\]

For minimum risk, choose the most probable class
Losses and Risks: Reject

\[ \lambda_{ik} = \begin{cases} 
0 & \text{if } i = k \\
\lambda & \text{if } i = K + 1, \quad 0 < \lambda < 1 \\
1 & \text{otherwise} 
\end{cases} \]

\[ R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda \]

\[ R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x}) \]

choose \( C_i \) if \( P(C_i \mid \mathbf{x}) > P(C_k \mid \mathbf{x}) \) \( \forall k \neq i \) and \( P(C_i \mid \mathbf{x}) > 1 - \lambda \)

reject otherwise
Discriminant Functions

choose $C_i$ if $g_i(x) = \max_k g_k(x)$

$$g_i(x) = \begin{cases} -R(\alpha_i \mid x) \\ P(C_i \mid x) \\ p(x \mid C_i)P(C_i) \end{cases}$$

$K$ decision regions $R_1, \ldots, R_K$

$$R_i = \{x \mid g_i(x) = \max_k g_k(x)\}$$

$g_i(x), i = 1, \ldots, K$
$K=2$ Classes

- Dichotomizer ($K=2$) vs Polychotomizer ($K>2$)
- $g(x) = g_1(x) - g_2(x)$
  
  \[
  \begin{cases}
  C_1 \text{ if } g(x) > 0 \\
  C_2 \text{ otherwise}
  \end{cases}
  \]

- Log odds:
  \[
  \log \frac{P(C_1 \mid x)}{P(C_2 \mid x)}
  \]
Utility Theory

- Prob of state $k$ given evidence $x$: $P(S_k|x)$
- Utility of $\alpha_i$ when state is $k$: $U_{ik}$
- Expected utility:

$$EU(\alpha_i \mid x) = \sum_k U_{ik} P(S_k \mid x)$$

Choose $\alpha_i$ if $EU(\alpha_i \mid x) = \max_j EU(\alpha_j \mid x)$
Value of Information

- Expected utility using $x$ only
  \[ EU(x) = \max_i \sum_k U_{ik} P(S_k | x) \]

- Expected utility using $x$ and new feature $z$
  \[ EU(x, z) = \max_i \sum_k U_{ik} P(S_k | x, z) \]

- $z$ is useful if $EU(x, z) > EU(x)$
Bayesian Networks

- Aka graphical models, probabilistic networks
- **Nodes** are hypotheses (random vars) and the prob 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 probs 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?

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

\[
= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\]
Causal vs Diagnostic Inference

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

\[
P(W|S) = P(W|R,S)P(R|S) + P(W|\sim R,S)P(\sim R|S) = 0.95 \cdot 0.4 + 0.9 \cdot 0.6 = 0.92
\]

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

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

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

Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.
Bayesian Networks: 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: \( P(C|W) = ? \)
**Bayesian Nets: Local structure**

\[
P(C) = 0.5
\]

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

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

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

\[
P(F \mid C) = ?
\]

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

\[
P(X_1, \ldots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents}(X_i))
\]
Bayesian Networks: Inference

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

\[ P(C,F) = \sum_S \sum_R \sum_W P(C,S,R,W,F) \]

\[ P(F|C) = \frac{P(C,F)}{P(C)} \quad \text{Not efficient!} \]

Belief propagation (Pearl, 1988)
Junction trees (Lauritzen and Spiegelhalter, 1988)
Bayesian Networks: Classification

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:

$$p(x|C) = p(x_1|C) \cdot p(x_2|C) \cdots p(x_d|C)$$
Influence Diagrams

- **Decision node**: choose class
- **Chance node**: x
- **Utility node**: U
Association Rules

- Association rule: $X \rightarrow Y$
- **Support** $(X \rightarrow Y)$:
  \[ P(X,Y) = \frac{\# \{ \text{customers who bought } X \text{ and } Y \} }{\# \{ \text{customers} \} } \]
- **Confidence** $(X \rightarrow Y)$:
  \[ P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{\# \{ \text{customers who bought } X \text{ and } Y \} }{\# \{ \text{customers who bought } X \} } \]

Apriori algorithm (Agrawal et al., 1996)