Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

Edited for CS536 Fall 05- Rutgers University
Ahmed Elgammal

CHAPTER 3:

Bayesian Decision Theory

Probability and Inference

- Result of tossing a coin is ∈ {Heads,Tails}
- Random var $X \in \{1,0\}$

  Bernoulli: $P\{X=1\} = p_o^X (1 - p_o)^{(1 - X)}$
- Sample: $\boldsymbol{X} = \{x^t\}_{t=1}^N$

  Estimation: $p_o = \#\{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t x^t / N$
- Prediction of next toss:

  Heads if $p_o > ½$, Tails otherwise

Classification

- Credit scoring: Inputs are income and savings.

  Output is low-risk vs high-risk
- Input: $\boldsymbol{x} = [x_1, x_2]^T$,Output: C ∈ {0,1}
- Prediction:

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

or equivalently

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

Bayes' Rule

$$\overset{\textit{posterior}}{P(\mathrm{C} \mid \boldsymbol{x})} = \frac{\overset{\textit{prior}}{P(\mathrm{C})}\,\overset{\textit{likelihood}}{p(\boldsymbol{x} \mid \mathrm{C})}}{\underset{\textit{evidence}}{p(\boldsymbol{x})}}$$

Prior: The knowledge we have as to the value of C before looking at the observation x

$$P(\mathrm{C}=0) + P(\mathrm{C}=1) = 1$$

Bayes' Rule

$$\overset{\textit{posterior}}{P(\mathrm{C} \mid \boldsymbol{x})} = \frac{\overset{\textit{prior}}{P(\mathrm{C})}\,\overset{\textit{likelihood}}{p(\boldsymbol{x} \mid \mathrm{C})}}{\underset{\textit{evidence}}{p(\boldsymbol{x})}}$$

- Likelihood: the conditional probability that an event belonging to C has associated observation value x
- Evidence: the marginal probability that an observation x is seen regardless of C

$$p(\mathbf{x}) = p(\mathbf{x} \mid \mathrm{C}=1)P(\mathrm{C}=1) + p(\mathbf{x} \mid \mathrm{C}=0)P(\mathrm{C}=0)$$

Bayes' Rule

prior *likelihood*

posterior

$$P(C \mid \boldsymbol{x}) = \frac{P(C)\,p(\boldsymbol{x} \mid C)}{p(\boldsymbol{x})}$$

evidence

- Combining the prior with what the data tells us, we can calculate the posterior probability P(C|x) after having seen the observation x
- The posterior sum up to 1

$$p(C=0 \mid \mathbf{x}) + P(C=1 \mid \mathbf{x}) = 1$$

Bayes' Rule

prior *likelihood*

posterior

$$P(C \mid \boldsymbol{x}) = \frac{P(C)\,p(\boldsymbol{x} \mid C)}{p(\boldsymbol{x})}$$

evidence

- Bayesian learning: from training data how to estimate P(C) and P(x|C)

Lecture Notes for E Alpaydın 2004 Introduction to Machine Learning © The MIT Press (V1.1)

Bayes' Rule

$$\underbrace{P(C \mid \boldsymbol{x})}_{\text{posterior}} = \frac{\overbrace{P(C)}^{\text{prior}}\,\overbrace{p(\boldsymbol{x} \mid C)}^{\text{likelihood}}}{\underbrace{p(\boldsymbol{x})}_{\text{evidence}}}$$

$$P(C = 0) + P(C = 1) = 1$$
$$p(\boldsymbol{x}) = p(\boldsymbol{x} \mid C = 1)P(C = 1) + p(\boldsymbol{x} \mid C = 0)P(C = 0)$$
$$p(C = 0 \mid \boldsymbol{x}) + P(C = 1 \mid \boldsymbol{x}) = 1$$

Bayes' Rule: K>2 Classes

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

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

- MAP: Maximal a Posteriori class: pick the class that maximizes the posterior probability

$$\text{choose } C_i \text{ if } P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$$
$$\mathrm{C}_{\mathrm{MAP}} \equiv \arg_{C_j} \max P(C_j \mid \mathbf{x})$$

Bayes' Rule: K>2 Classes

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

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

choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

$\mathrm{C_{MAP}} \equiv \arg_{C_j} \max P(C_j \mid \mathbf{x}) = P(\mathbf{x} \mid C_j)P(C_j)$

$C_{ML} \equiv \arg_{C_j} \max P(\mathbf{x} \mid C_j)$

- ML: Maximal Likelihood Class: Special case; assume all class priors $P(C_j)$ are equal

Losses and Risks

- It may be the case that decisions are not equally good or costly.
- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart, 1973)

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

$$\text{choose } \alpha_i \text{ if } R(\alpha_i \mid \boldsymbol{x}) = \min{}_k R(\alpha_k \mid \boldsymbol{x})$$

Losses and Risks: 0/1 Loss

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

$$\begin{aligned} R(\alpha_i \mid \boldsymbol{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \boldsymbol{x}) \\ &= \sum_{k \neq i} P(C_k \mid \boldsymbol{x}) \\ &= 1 - P(C_i \mid \boldsymbol{x}) \end{aligned}$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

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

$$R(\alpha_{K+1} \mid \boldsymbol{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \boldsymbol{x}) = \lambda$$

$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{k \neq i} P(C_k \mid \boldsymbol{x}) = 1 - P(C_i \mid \boldsymbol{x})$$

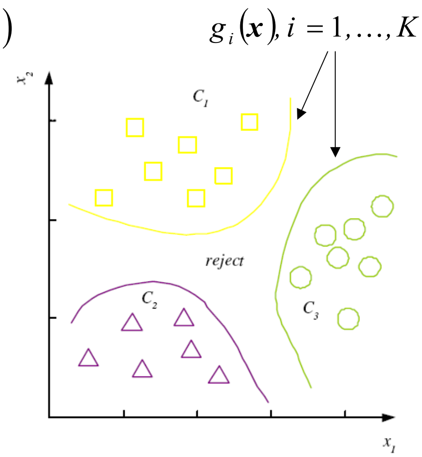
choose C_i if $P(C_i \mid \boldsymbol{x}) > P(C_k \mid \boldsymbol{x}) \ \forall k \neq i$ and $P(C_i \mid \boldsymbol{x}) > 1 - \lambda$

reject otherwise

Discriminant Functions

$$\text{choose } C_i \text{ if } g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x})$$

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



K decision regions $\mathcal{R}_1,\ldots,\mathcal{R}_K$

$$\mathcal{R}_i = \{\boldsymbol{x} \mid g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x})\}$$

K=2 Classes

- Dichotomizer (K=2) vs Polychotomizer (K>2)
- $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

$$\text{choose } \begin{cases} C_1 \text{ if } g(\boldsymbol{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$$

- *Log odds:*

$$\log \frac{P(C_1 \mid \boldsymbol{x})}{P(C_2 \mid \boldsymbol{x})}$$

Utility Theory

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

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

$$\text{Choose } \alpha_i \text{ if } EU(\alpha_i \mid \boldsymbol{x}) = \max_j EU(\alpha_j \mid \boldsymbol{x})$$

Value of Information

- Expected utility using $\boldsymbol{x}$ only

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

- Expected utility using $\boldsymbol{x}$ and new feature z

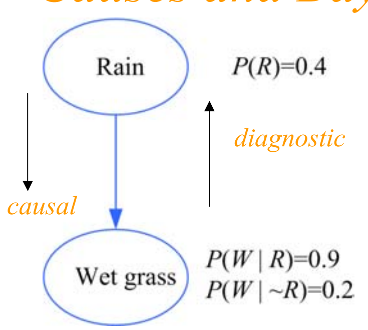
$$EU(\boldsymbol{x}, z) = \max_i \sum_k U_{ik} P(S_k \mid \boldsymbol{x}, z)$$

- z is useful if $EU(\boldsymbol{x}, z) > EU(\boldsymbol{x})$

Bayesian Networks

- graphical models, probabilistic networks, belief networks
- Nodes are hypotheses (random vars) and the prob corresponds to our belief in the truth of the hypothesis
- Arcs are direct 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?

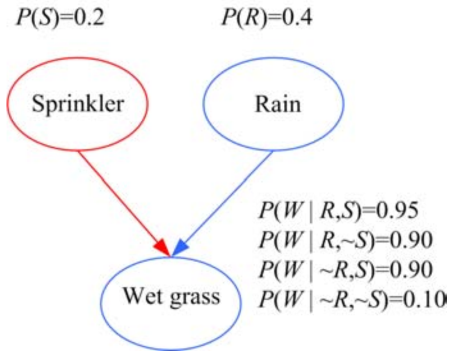
$$P(R \mid W) = \frac{P(W \mid R)P(R)}{P(W)}$$

$$= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)}$$

$$= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75$$

Notice: knowing the grass is wet increases the probability of rain from 0.4 to 0.75

Causal vs Diagnostic Inference



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

$P(W|S) = P(W|R,S)\, P(R|S) + P(W|\sim R,S)\, P(\sim R|S)$

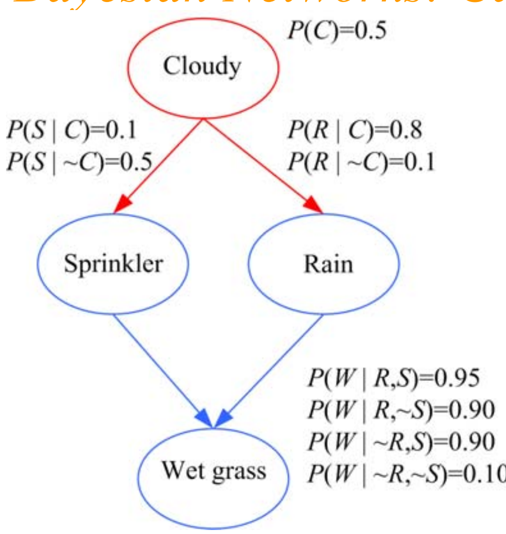
$= P(W|R,S)\, P(R) + P(W|\sim R,S)\, P(\sim R)$

$= 0.95\ 0.4 + 0.9\ 0.6 = 0.92$

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on? $P(S|W) = 0.35 > 0.2\ P(S)$

$P(S|R,W) = 0.21$ *Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

Bayesian Networks: Causes



Causal inference:

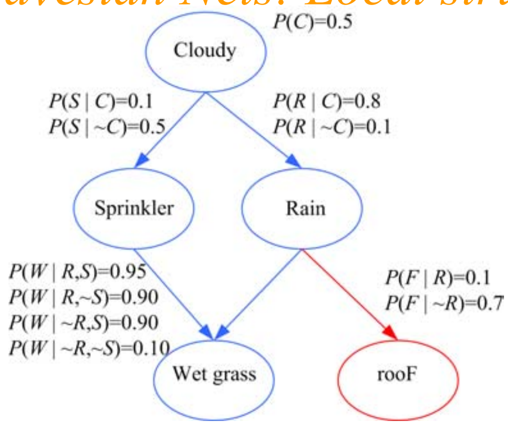
$P(W|C) = P(W|R,S)\, P(R,S|C) + P(W|\sim R,S)\, P(\sim R,S|C) + P(W|R,\sim S)\, P(R,\sim S|C) + P(W|\sim R,\sim S)\, P(\sim R,\sim S|C)$

and use the fact that

$P(R,S|C) = P(R|C)\, P(S|C)$

Diagnostic: $P(C|W) = ?$

Bayesian Nets: Local structure



$P(F \mid C) = ?$

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

$$P(X_1,\ldots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents}\ (X_i))$$

Bayesian Networks: Inference

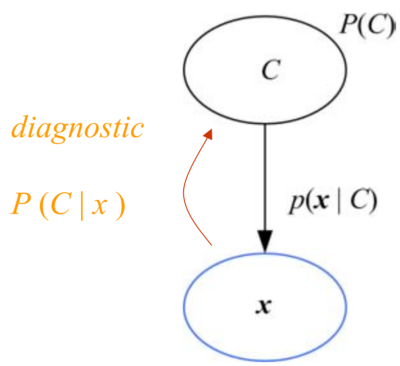
$P\ (C,S,R,W,F) = P\ (C)\ P\ (S|C)\ P\ (R|C)\ P\ (W|R,S)\ P\ (F|R)$

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

$P\ (F|C) = P\ (C,F)\ /\ P(C)$ *Not efficient!*

Belief propagation (Pearl, 1988)
Junction trees (Lauritzen and Spiegelhalter, 1988)

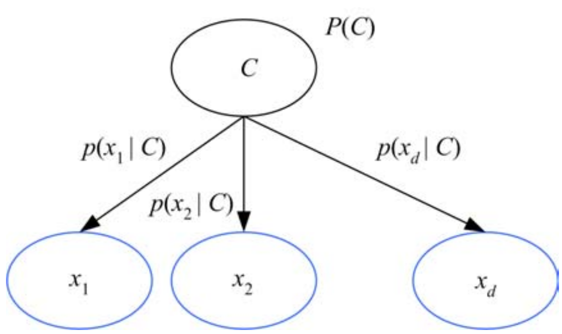
Bayesian Networks: Classification



Bayes' rule inverts the arc:

$$P(C \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C) P(C)}{p(\boldsymbol{x})}$$

Naive Bayes' Classifier



Given C, x_j are independent:

$$p(\boldsymbol{x}|C) = p(x_1|C)\, p(x_2|C) \dots p(x_d|C)$$