CHAPTER 3

Bayesian Decision Theory

Plan

- Discuss probability theory as the framework for making decisions under uncertainty.
- Bayes' rule is used to calculate the probabilities of the classes.
- Rational decisions to minimize expected risk.
- Introduce Bayesian networks to visually and efficiently represent dependencies among random variables.

Probability and Inference

- Result of tossing a coin is $\in$ {Heads,Tails}
- Random var X $\in${1,0}
  - Bernoulli: P (X=1) = ?
    - $p_0^X (1 - p_0)^{(1-X)}$
- How do you predict the result of the next toss?
  - Heads if $p_0 > ½$, Tails otherwise
- We do not know P(X)
  - Sample, instances/examples drawn from distribution p(x): $X = \{x^t\}_{t=1}^N$
  - Build an approximator to it p'(x)
  - Estimation: p'= # {Heads}/#{Tosses} = $\Sigma_t x_t / N$

Classification

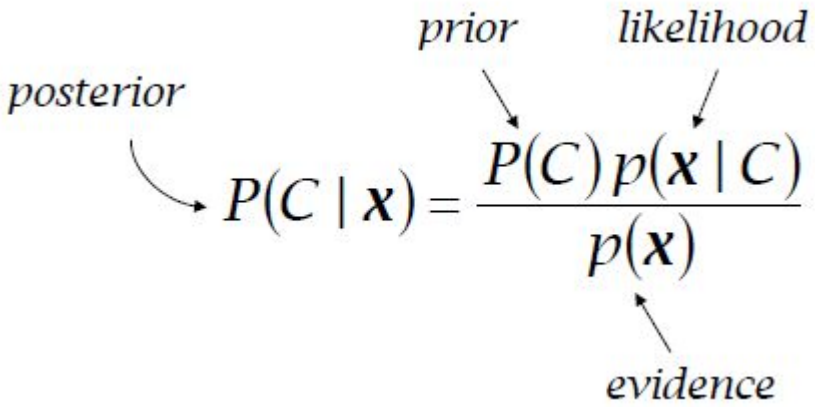
- Credit scoring: Inputs are income and savings.
  - Many other unobserved attributes
  - Output is low-risk vs high-risk
- Input: $\mathbf{x} = [x_1, x_2]^T$, Output: $C \in \{0,1\}$
- The credibility of a customer: a Bernoulli random variable **C** conditioned on the observables $X = [\mathbf{X}_1, \mathbf{X}_2]^T$
- When a new application arrives with $[x_1, x_2]$:

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

or equivalently

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

Bayes' Rule



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

- What is P(C = 1)?
  - probability that a customer is high-risk, regardless of the customer
- $P(x_1,x_2|C=1)$?
  - probability a high-risk customer has x_1,x_2
- $p(x_1,x_2)$:
  - is the marginal probability that an observation x_1,x_2 is seen

Bayes' Rule: K>2 Classes

e.g. digit recognition

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

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

the Bayes' classifier chooses the class with ?

- highest posterior probability:

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

Losses and Risks

- Should we directly choose based on output of the Bayes' classifier? Are decisions always equally good or costly?
- Loss for a high-risk applicant erroneously accepted vs. potential gain for an erroneously rejected low-risk applicant.
- How about medical diagnosis or earthquake prediction?
- Define action α_i: decide on C_i
- λ_{ik} : Loss of taking action α_i (decide C_i) when the state is C_k
- How do you decide which action to take?
- What is the expected risk of taking action α_i $R(\alpha_i \mid \boldsymbol{x})$

Losses and Risks

Define action α_i: decide on C_i

λ_{ik} : Loss of taking action α_i (decide C_i) when the state is C_k

The expected risk of taking action α_i:

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

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

Losses and Risks: 0/1 Loss

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

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

$$= \sum_{k \neq i} P(C_k \mid \boldsymbol{x})$$

$$= 1 - P(C_i \mid \boldsymbol{x})$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

In some applications, wrong decisions - namely misclassifications- may have very high cost, and more complex, manual-decision is made if the automatic system has low certainty of its decision.

New action, reject: $R(\alpha_{K+1} \mid \mathbf{x})$ How can we formulate the loss?

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

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

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

Losses and Risks: Reject. Decision rule

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

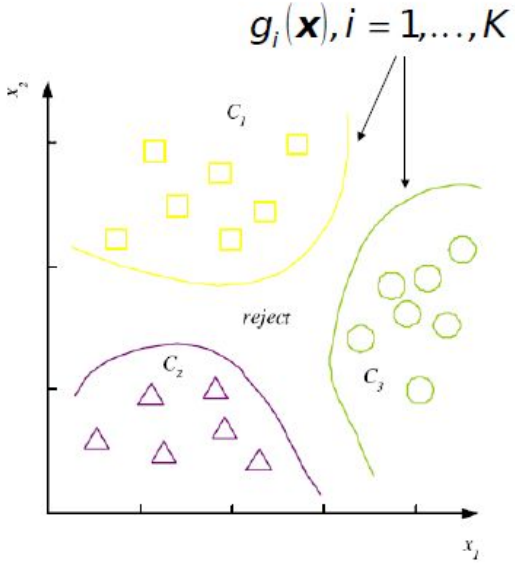
$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{k \neq i} P(C_k \mid \boldsymbol{x}) = 1 - P(C_i \mid \boldsymbol{x})$$

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

Classification can also be seen as implementing a set of discriminant functions, g(x), i = 1, ... , K, such that



We can represent the Bayes' classifier in this way, and the maximum discriminant function corresponds to minimum conditional risk.

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

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

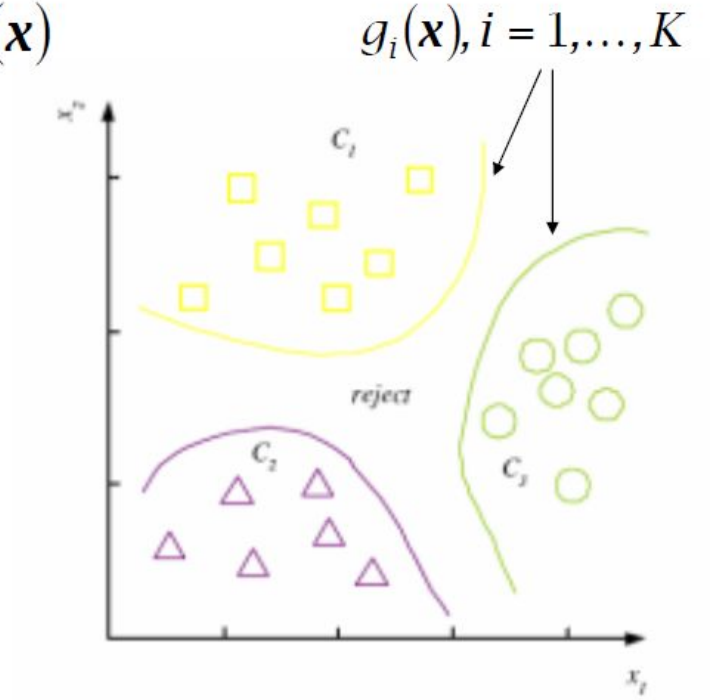
Discriminant functions

choose C_i 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* $\mathcal{R}_1,\ldots,\mathcal{R}_K$

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



Decision boundaries

K=2 classes

- Dichotomizer (K=2) vs Polychotomizer (K>2)

- $g(x) = g_1(x) - g_2(x)$

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

Log odds:

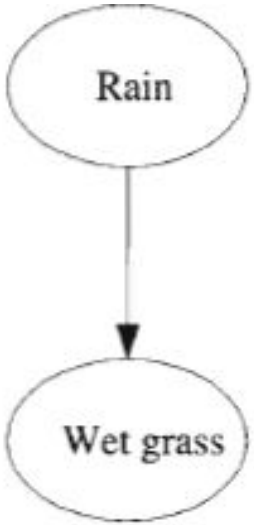
$$\log \frac{P(C_1 \mid \boldsymbol{x})}{P(C_2 \mid \boldsymbol{x})}$$

Bayesian Networks

Bayesian networks, also called belief networks or probabilistic networks, are graphical models for representing the interaction between variables visually.

- Composed of nodes and arcs.
- Each node corresponds to a random variable
- a directed arc from node X to node Y: X has a direct influence on Y.
- Specified by the conditional probability: P(YIX).
- Nodes and arcs: structure
- Conditional probabilities: representation

Bayesian Networks

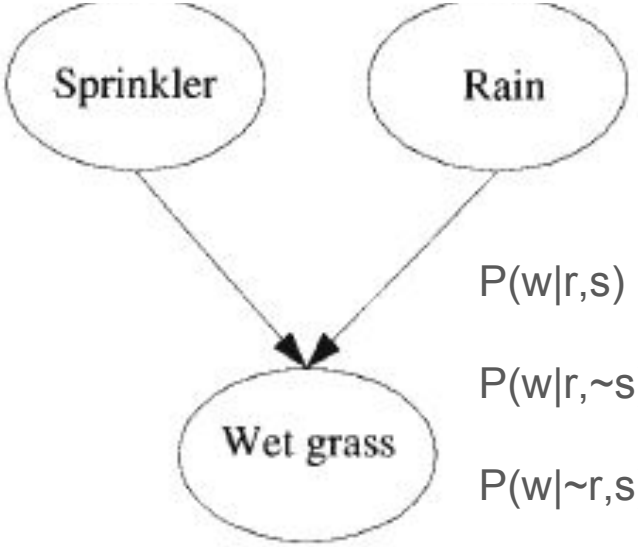


P(r)=0.4

P(w|r) =0.4

P(w|~r)=0.2

P(s)=0.2 P(r)=0.4



P(w|r,s) = 0.95

P(w|r,~s) = 0.90

P(w|~r,s) = 0.90

P(w|~r,~s) = 0.10

Need to compute joint distribution

Diagnosis through causal graph: knowing that the grass is wet, the probability that it rained?

Probability of having wet grass given the sprinkler is on, not knowing whether it rained or not

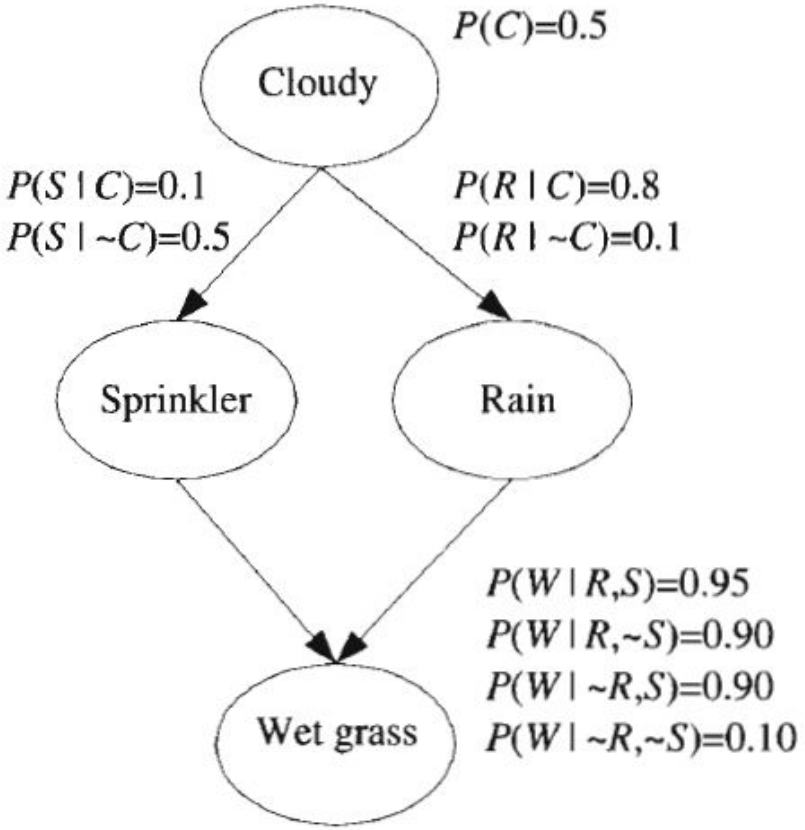
p(s|w)=?

p(s|r,w)=?

Bayesian Networks

we may think that R and S are actually dependent in the presence of another variable: we usually do not turn on the sprinkler if it is likely to rain.

The probability of having wet grass if it is cloudy?

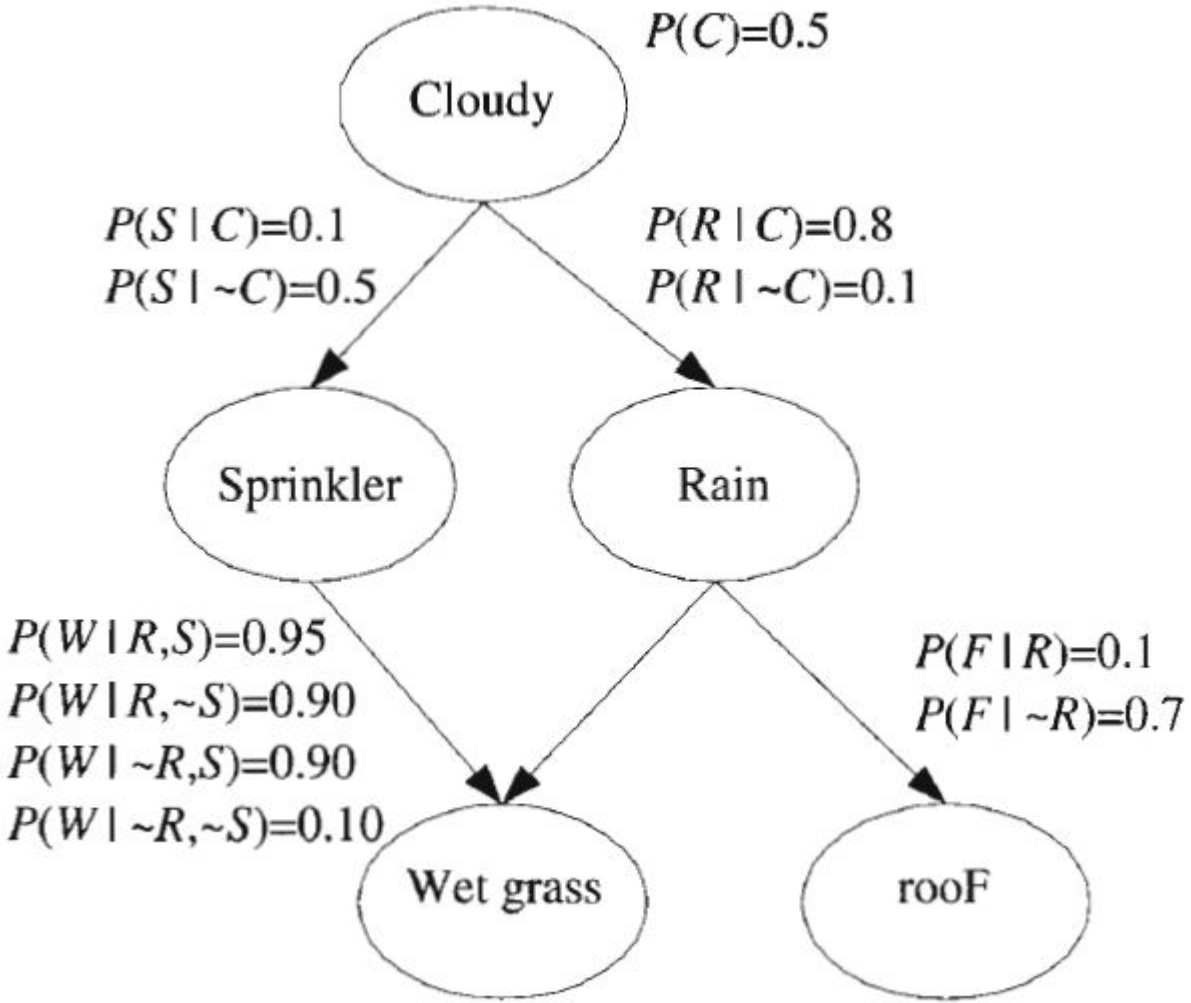


BN

Probability that we hear the cat on the roof given that it is cloudy.

Or

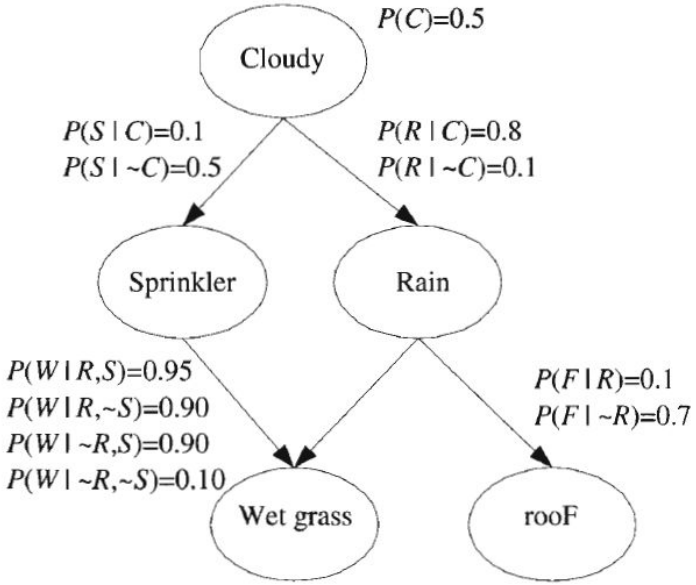
We hear the cat on the roof. The prob. of being cloudy? - Diagnosis



Bayesian Networks

Advantage?

- Local structures
- Eases analysis and computation.
- No input-output designation

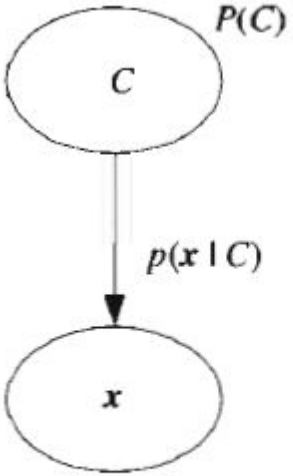


$$P(X_1, \ldots, X_d) = \prod_{i=1}^{a} P(X_i | \text{parents}(X_i))$$

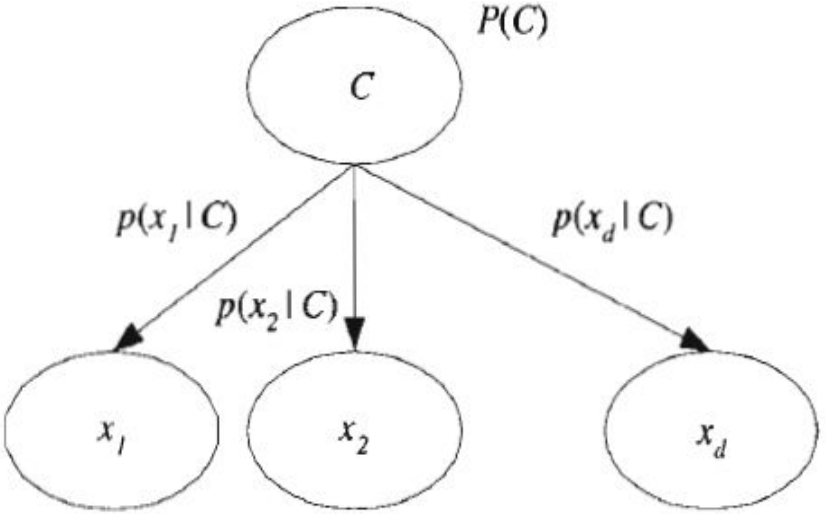
Does not denote causality: Says directly link

Bayesian Networks

Most of the methods discussed can be written down as BNs.



Bayesian network for classification.



Naive Bayes' classifier is a Bayesian network for classification