FUNDAMENTALS OF MACHINE LEARNING

BAYESIAN DECISION THEORY

CSCI3320

Prof. John C.S. Lui, CSE Department, CUHK
Introduction to Machine Learning

Goals: Making Decision Under Uncertainty

- Use Bayes' rule to calculate the probability of the classes
- Make rational decision among multiple actions to minimize expected risk
- Learning association rules from data

Unobservable variables

- Tossing a coin is completely a random process, can't predict the outcome
- Only can talk about the **probabilities** that the outcome of the next toss will be head or tails
- If we have access to extra knowledge (exact composition of the coin, initial position, force etc.) the exact outcome of the toss can be predicted (*assuming we have a model for it*)

Unobservable Variable

- Unobservable variable is the extra knowledge that we don't have access to
- Coin toss: the only observable variable is the outcome of the toss
- $x = f(z)$, z is unobservable, x is observable
- f is a deterministic function

Bernoulli Random Variable

- Result of tossing a coin is ∈ {Heads,Tails}
- Define a random variable $X \in \{1,0\}$
- p_0 the probability of heads
- $P(X = 1) = p_0$ *and* $P(X = 0) = 1 - P(X = 1) = 1 - p_0$
- You are asked to **predict the next toss**
- If know p_0 we would predict heads if $p_0 > \frac{1}{2}$
- Why? We choose a more probable case to minimize the probability of the error *1 -* p_0
- Sample: $\boldsymbol{X} = \{x^t\}_{t=1}^{N}$
- Estimation: $\hat{p}_0 = \#\{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t \frac{x^t}{N}$
- **Prediction of next toss:**

  Heads if $\hat{p}_0 > ½$, Tails otherwise

Parameter Estimation

$$\hat{p}_o = \frac{\#\{\text{tosses with outcome heads}\}}{\#\{\text{tosses}\}}$$

$$\mathcal{X} = \{1, 1, 1, 0, 1, 0, 0, 1, 1\}$$

$$\hat{p}_o = \frac{\sum_{t=1}^{N} x^t}{N} = \frac{6}{9}$$

This is part of the sampling and point estimation that we learnt

From input data, *X*, perform classification !!

Classification

- **Credit scoring**: two classes – **high risk** and **low risk**
- Decide on observable information: (income and saving)
- Have reasons to believe that these two variables give us some idea about the credibility of a customer
- Represent by two **random variable** X_1 and X_2
- Can't observe customer intentions and moral codes
- Can observe credibility of a past customer
- Bernoulli random variable C conditioned on $\boldsymbol{X} = [X_1, X_2]^T$
- From the past data, we know to find $P(C|X_1, X_2)$

Classification

- Assume know $P(C|X_1, X_2)$
- New application comes in: $X_1 = x_1, X_2 = x_2$
- How should we decide? What is the right rule?

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

or equivalently

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

Classification

- The probability of error is
  $$1 - \max\{ P(C = 1|x_1, x_2), P(C = 0|x_1, x_2) \}$$

- Similar to coin toss but C is conditioned on two other observable variables $\boldsymbol{X} = [X_1, X_2]^T$

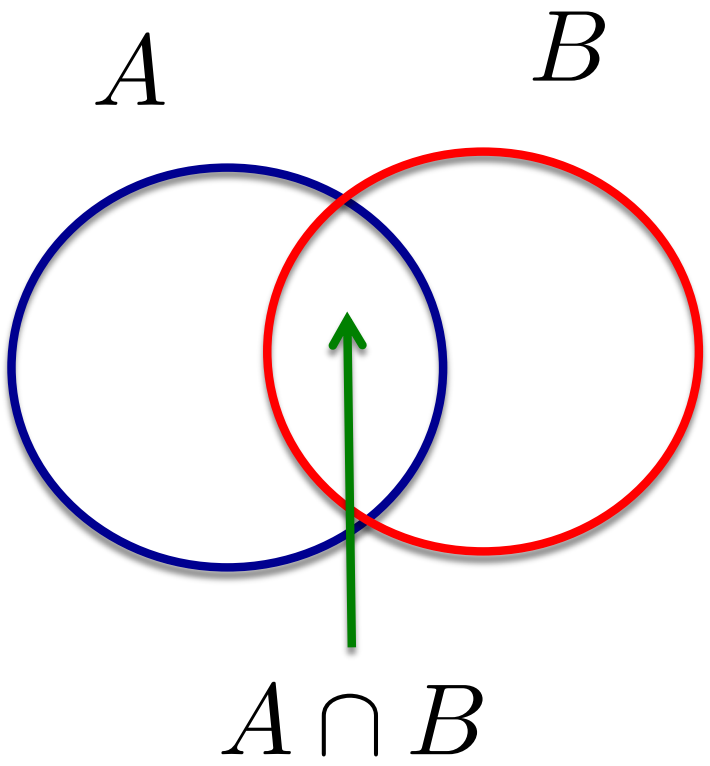
- The problem : *How to calculate* $P(C|\boldsymbol{X})$?

- Use **Bayes' rule**

Conditional Probability

- Probability of A (point will be inside A) if we know that B happens (point is inside B)

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$



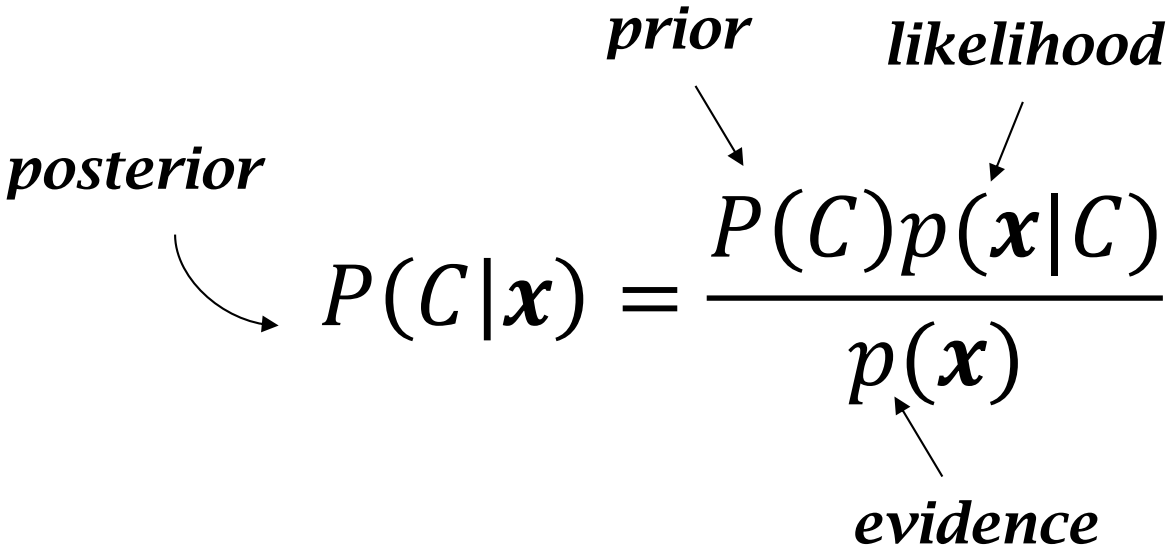
Explain why we have normalization

Bayes' Rule

- First: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Also: $P(B|A) = \frac{P(A \cap B)}{P(A)} \rightarrow P(A \cap B) = P(B|A)P(A)$

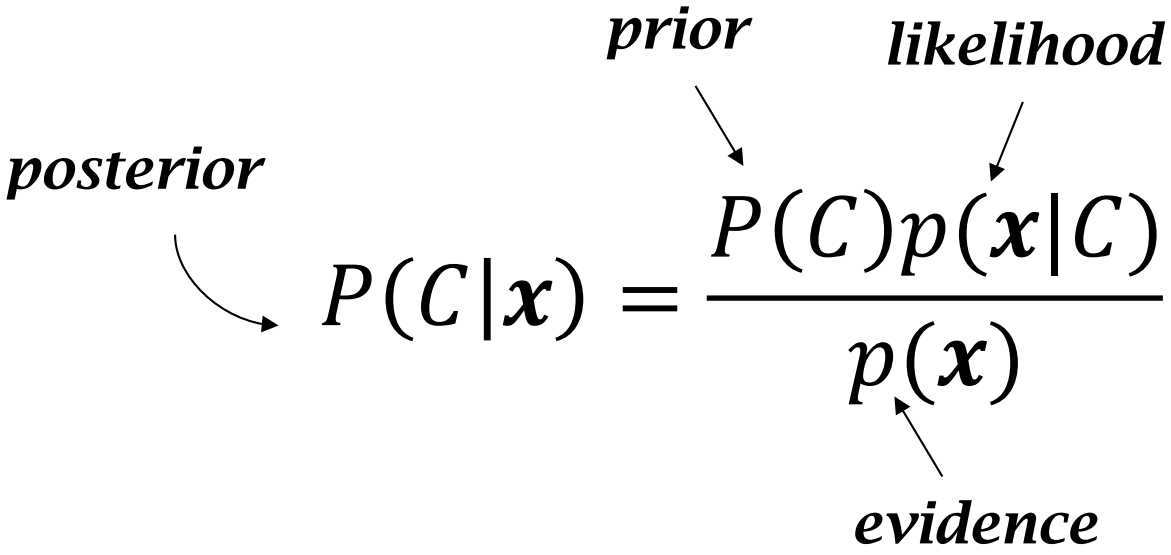
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule



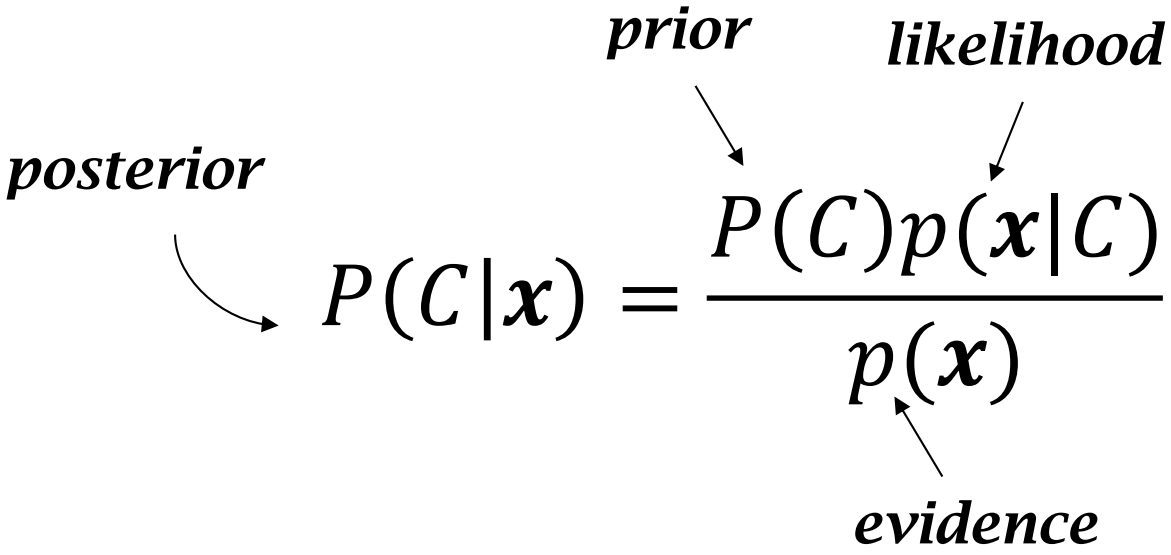
- We need to evaluate both $P(C = 0|\boldsymbol{x})$ *and* $P(C = 1|\boldsymbol{x})$
- **Prior:** probability of a customer is high risk regardless of $\boldsymbol{x}$.
- Knowledge we have as to *the value of* C *before looking at observables* $\boldsymbol{x}$

Bayes' Rule



- **Likelihood:** probability that event in C will have observable $\boldsymbol{x}$
- $P(x_1, x_2|C = 1)$ is the probability that a high-risk customer has his $X_1 = x_1, X_2 = x_2$

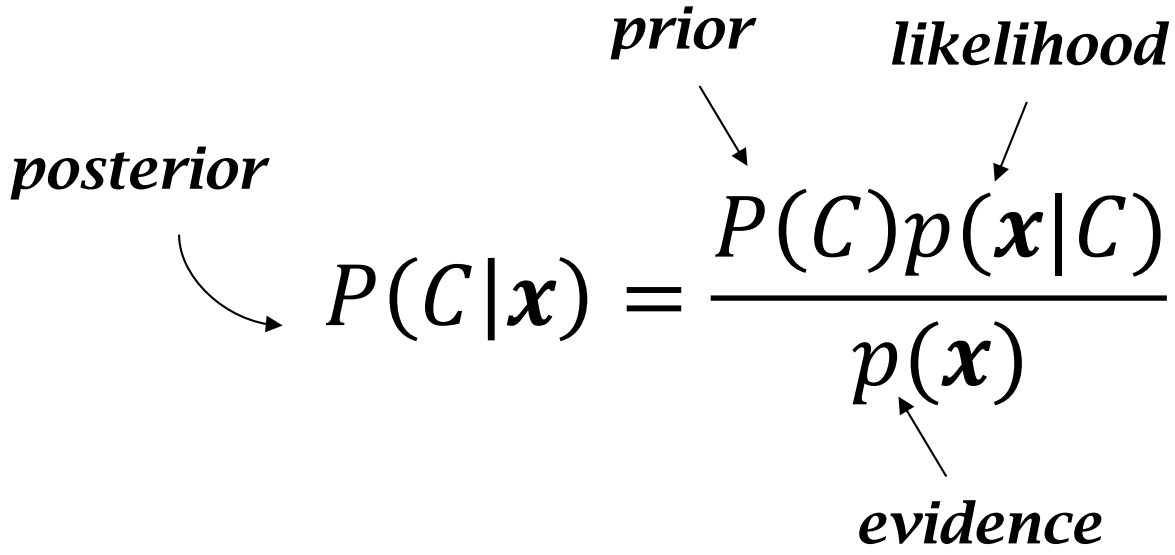
Bayes' Rule



- **Evidence:** $p(\boldsymbol{x})$ probability that observation $\boldsymbol{x}$ is seen regardless if positive or negative

$$p(\boldsymbol{x}) = \sum_{C} p(\boldsymbol{x}, C) = p(\boldsymbol{x}|C = 1)P(C = 1) + p(\boldsymbol{x}|C = 0)P(C = 0)$$

Bayes' Rule



$$P(C=0) + P(C=1) = 1$$

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

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

Once we have the *posterior*, we use "prediction rule" to decide

Summary of Bayes' Rule for classification

- Assume we know : prior, evidence and likelihood
- *Will learn how to estimate them from the data later*
- Plug them in into Bayes formula to obtain $P(C|\boldsymbol{x})$
- **Rule:**

$$\text{If } P(C=1|\boldsymbol{x}) > P(C=0|\boldsymbol{x})$$

$$\text{Choose } C=1$$

$$\text{else Choose } C=0$$

Bayes' Rule: $K > 2$ Classes

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

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

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

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

Bayes' classifier with the minimum error

Bayes' Rule

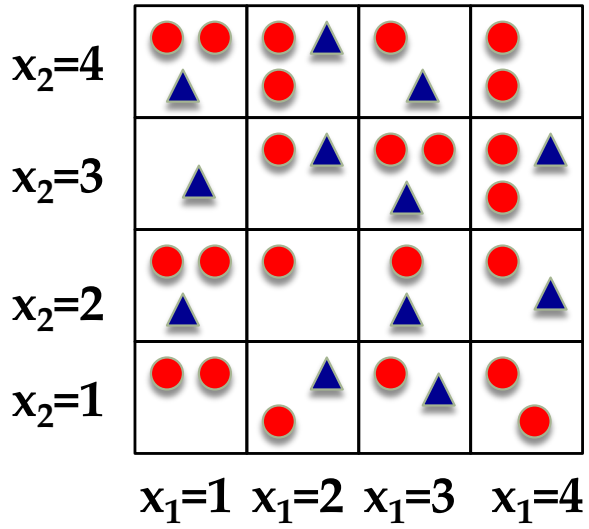
$$P(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i)P(C_i)}{p(\boldsymbol{x})}$$

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

- Deciding on specific input $\boldsymbol{x}$
- $p(\boldsymbol{x})$ is the same for all classes
- Don't need it to compare posterior (*reduce computation*)

Pictorial View of Bayesian Decision

- Consider two classes first. Class 1 Red, Class 2 Blue

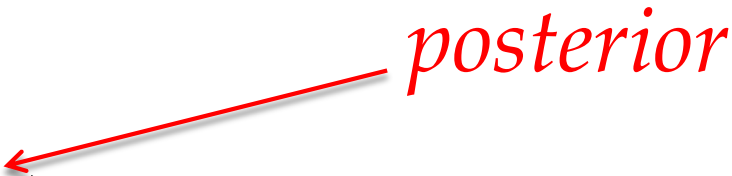


Total red points=23
Total blue points=12

- $P(C_1)$= total red/total points=23/35; $P(C_2)$=total blue/total point=12/35
- $P(x_1=1,x_2=1|C_1) = 2/23$; $P(x_1=1,x_2=3|C_1) = 0/23$;
- $P(x_1=1,x_2=1|C_2) = 0/12$; $P(x_1=1,x_2=3|C_2) = 1/12$;
- $P(C=1|x_1=1,x_2=2)$=? $P(x_1=1,x_2=2|C=1)P(C=1)=2/23*23/35=2/35$
- $P(C=2|x_1=1,x_2=2)$=? $P(x_1=1,x_2=2|C=2)P(C=2)=1/12*12/35=1/35$
- **What is the rule? Why we don't need to normalize?**

Losses and Risks

- Decisions/Errors are not equally good or costly
- **Actions:** α_i is assignment to class i
- **Loss** of α_i when the state is C_k : λ_{ik}
- **Expected risk**:



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

$$\text{Choose } \alpha_i \text{ if } R(\alpha_i|\boldsymbol{x}) = \min_j R(\alpha_j|\boldsymbol{x})$$

Losses and Risks: 0/1 Loss

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

All mistakes are the same cost

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

For minimum risk, ***choose the most probable class***

Losses and Risks: Reject

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

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

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

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

reject otherwise

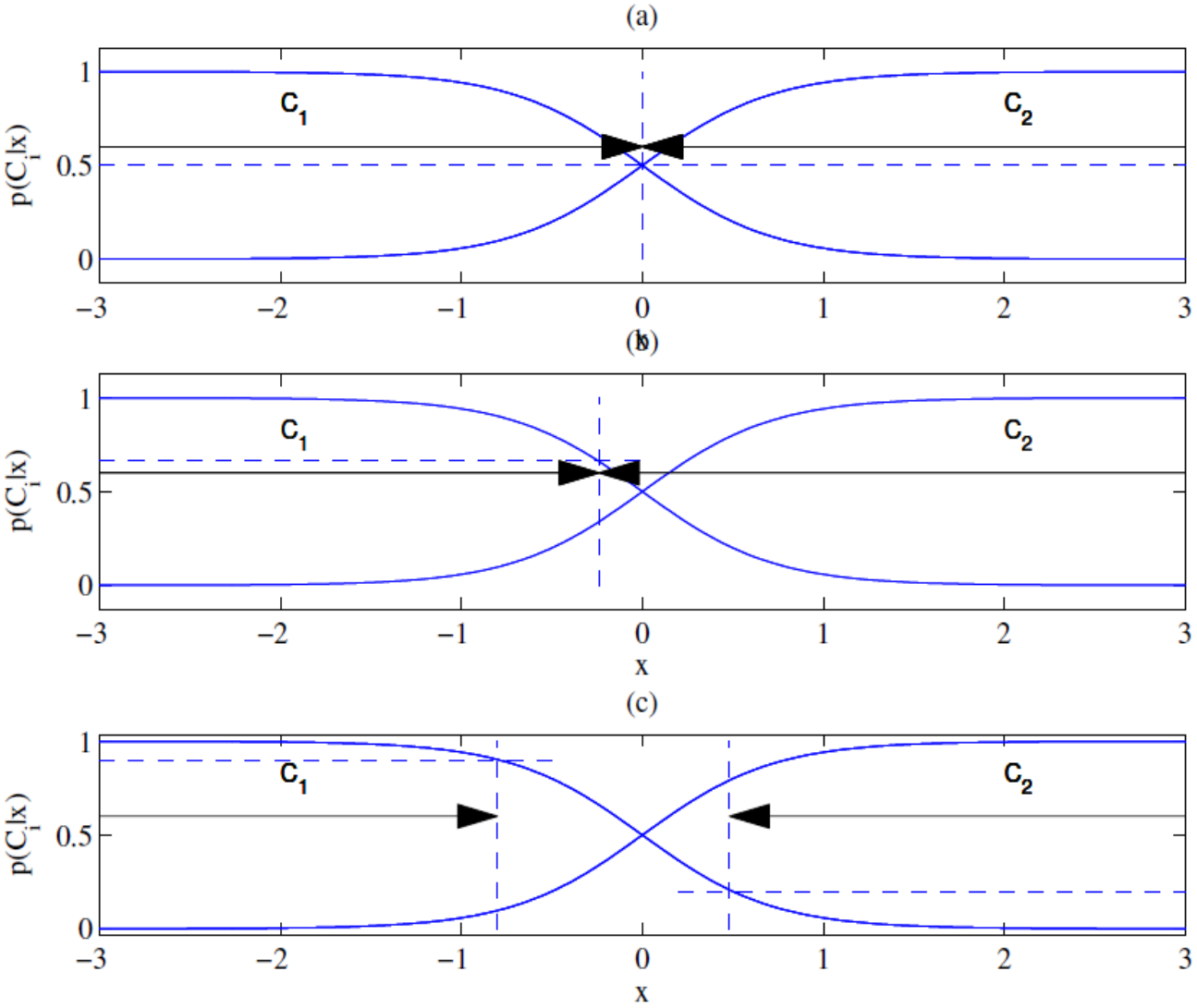
Show derivation

Different Losses and Reject

Equal losses

Unequal losses
$\lambda_{12} > \lambda_{21}$

With reject



Classification via Discriminant Functions

- Define a function $g_i(\boldsymbol{x})$ for each class i ("goodness" of selecting class C_i given observables $\boldsymbol{x}$)

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

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

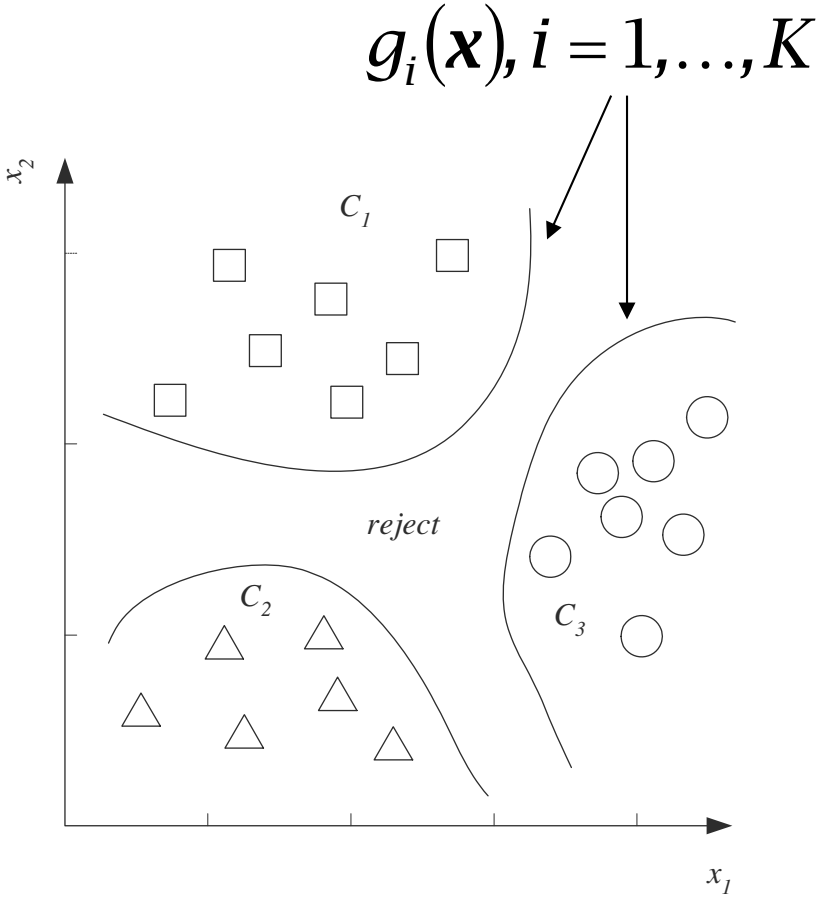
Ignore normalized term $p(x)$

  - **Maximum discriminant** corresponds to **minimum conditional risk**

Decision Regions in Feature Space

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

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



In abstract terms, we use ***evidence*** *(e.g., probability, statistics, …,etc) to compose our* ***discriminant functions****, from these functions we carve out the* ***decision regions***

Utility Theory

- Probability of state k given evidence $\boldsymbol{x}$: $P(S_k|\boldsymbol{x})$
- Utility of action α_i when state is k: U_{ik} (*can be positive or negative*)
- Expected utility:

$$E[U(\alpha_i|\boldsymbol{x})] = \sum_{k=1}^{K} U_{ik} P(S_k|\boldsymbol{x})$$

$$\text{choose } \alpha_i \text{ if } E[U(\alpha_i|\boldsymbol{x})] = \max_j E[U(\alpha_j|\boldsymbol{x})]$$

$$\text{For example, } U_{ii} > 0,\ U_{ij} = -10,\ U_{i,K+1} = -5$$

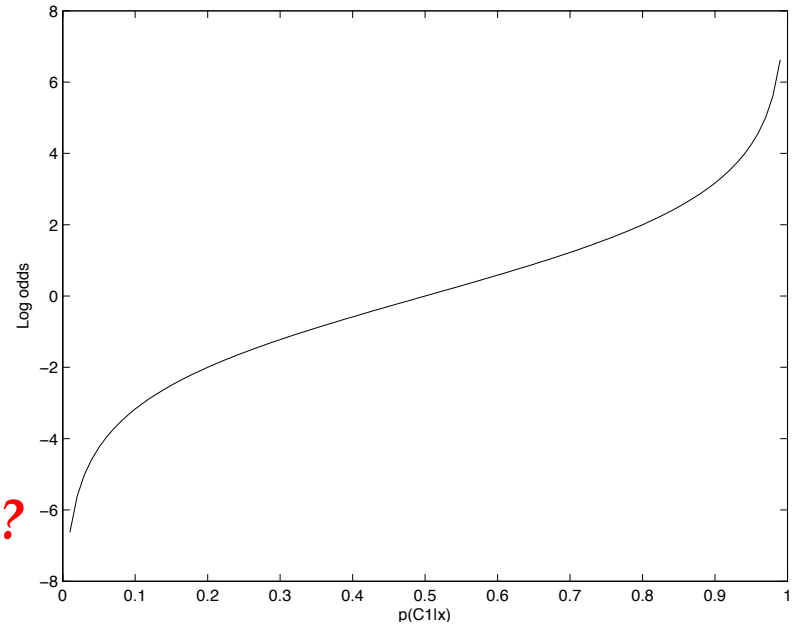
K=2 Classes

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

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

- *Log odds:*

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

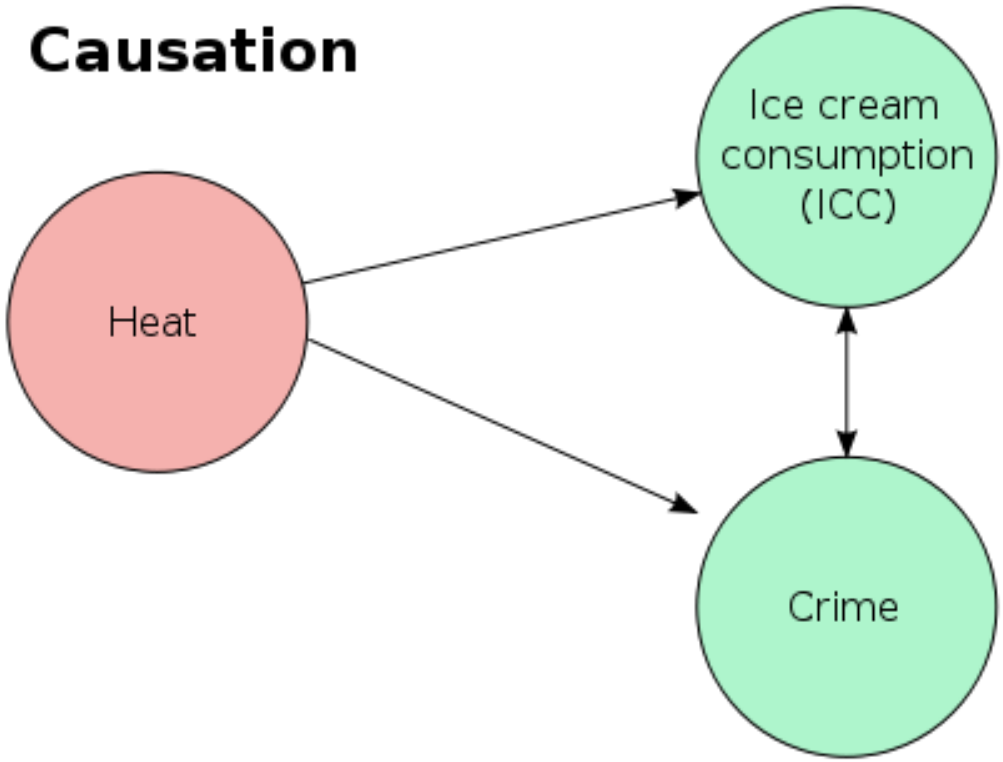


What should the decision be for log odds?

Some food for thought

- Example used as illustration, can be complicated (imagine a continuous *d*-dimensional pdf)
- The theory of Bayes' rule serves as a guiding principle
- Imagine you have given *N* *d*-dimensional points, each point is labeled
- You can easily use Bayes' rule to get all the probabilities. So when a new point **x** comes in, you can do the classification
  - Application: spam email detection (explain in class)
- Some "computer science" challenges:
  - *What happen if I want to include more "attributes"?*
  - *After M new points, how should I update the probabilities?*
  - *Given d-dimensional space, how can I reduce memory/storage requirement?*

Correlation and Causality

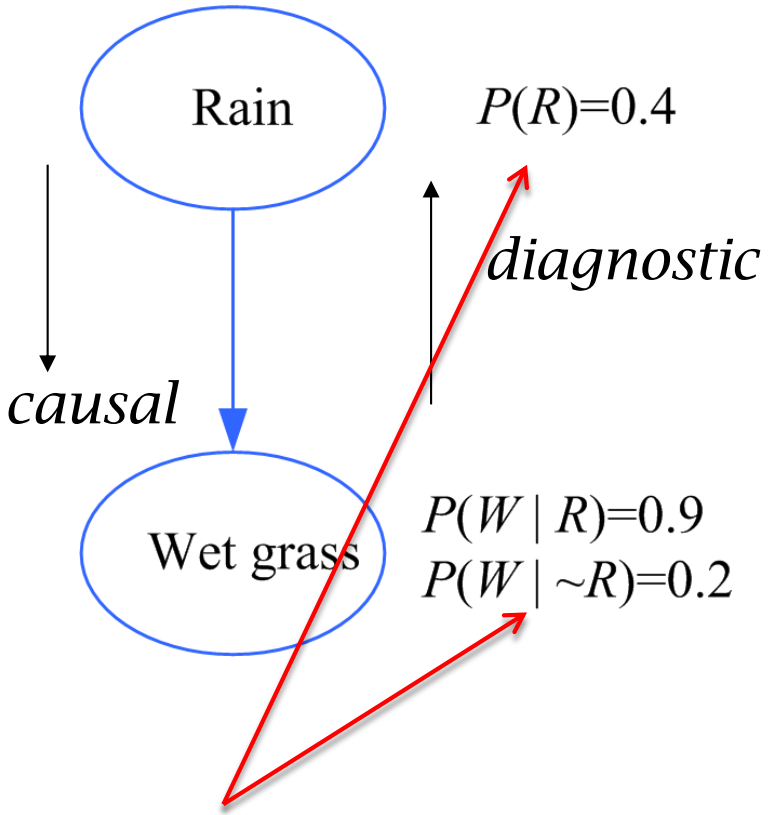


Interaction Between Variables

- Have many variables we want to represent and analyze ?
  - Structure of dependency among them
  - Conditional and marginal probabilities

- Use **graphical representation**
- **Nodes** (events) and **arcs** (dependencies)
- Numbers of nodes and arcs (conditional probabilities)

Causes and Bayes' Rule



Explain notation

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

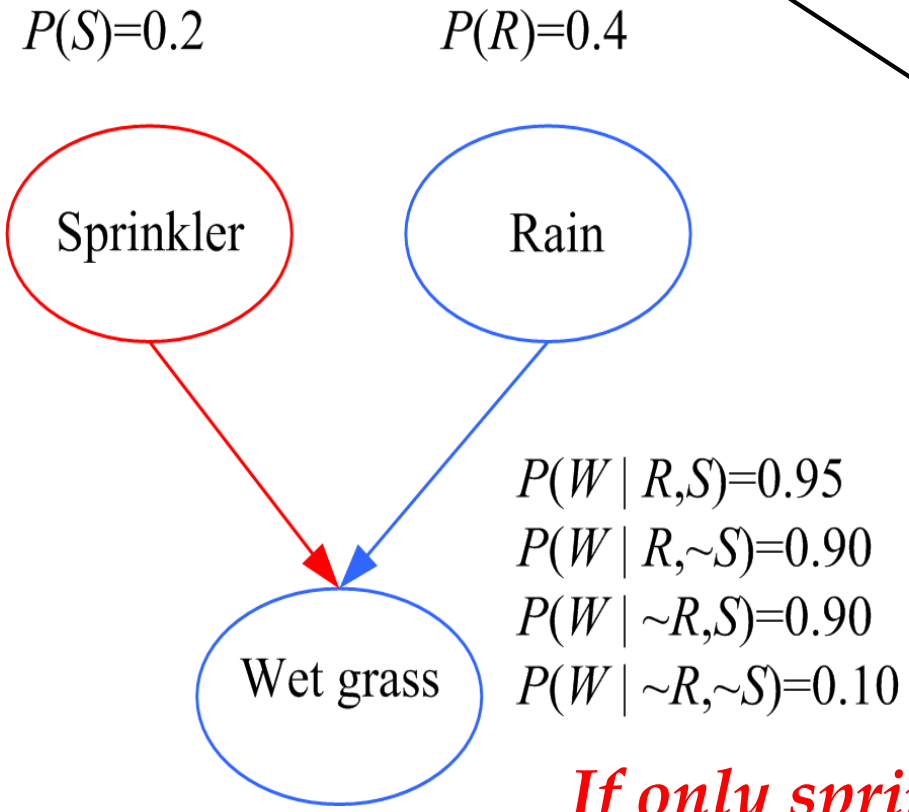
$$P(R|W) = \frac{P(W|R)P(R)}{P(W)}$$

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

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

Causal vs. Diagnostic Inference

$$P(W|S) = \frac{P(WS)}{P(S)} = \frac{P(WRS) + P(W{\sim}RS)}{P(S)}$$
$$= P(W|RS)P(R|S) + P(W|{\sim}RS)P({\sim}R|S)$$



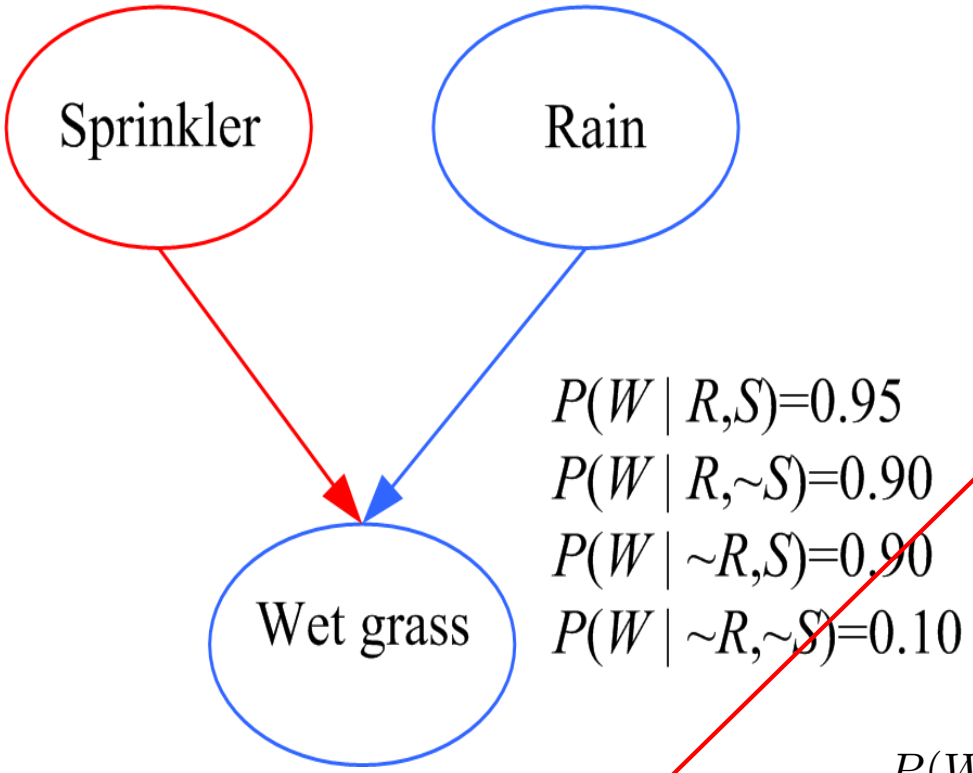
Causal inference: *If the sprinkler is on, what is the probability that the grass is wet?*

$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|{\sim}R,S)\,P({\sim}R|S)$$
$$= P(W|R,S)\,P(R) + P(W|{\sim}R,S)\,P({\sim}R)$$
$$= 0.95*0.4 + 0.9*0.6 = 0.92$$

why ?

If only sprinkler is on, it has a lower probability than when both sprinkler and raining in causing wet glasses.

Causal vs. Diagnostic Inference



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

$P(S|W) = 0.35 > P(S)=0.2$

Show in class

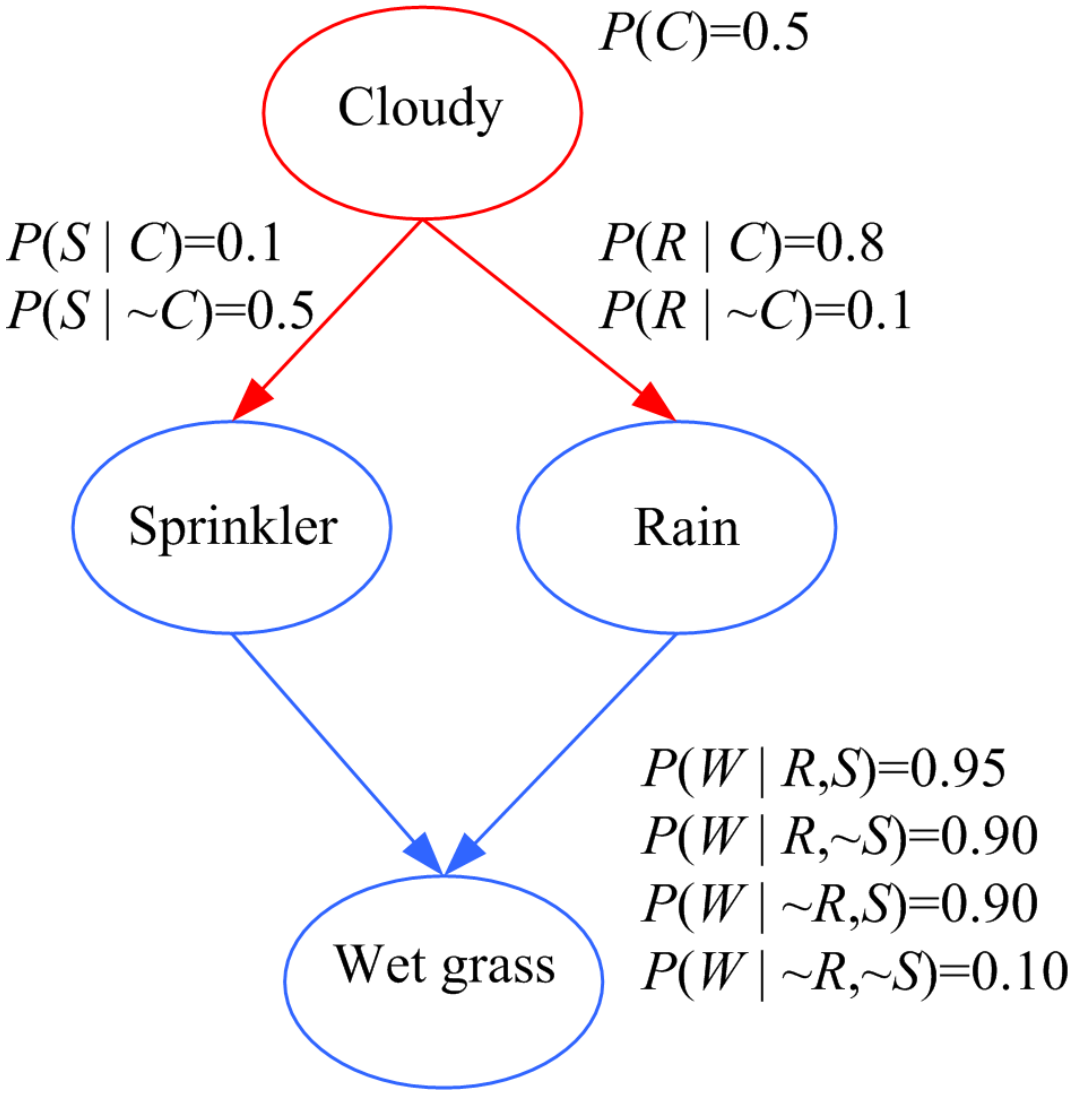
$P(S|R,W) = 0.21$

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

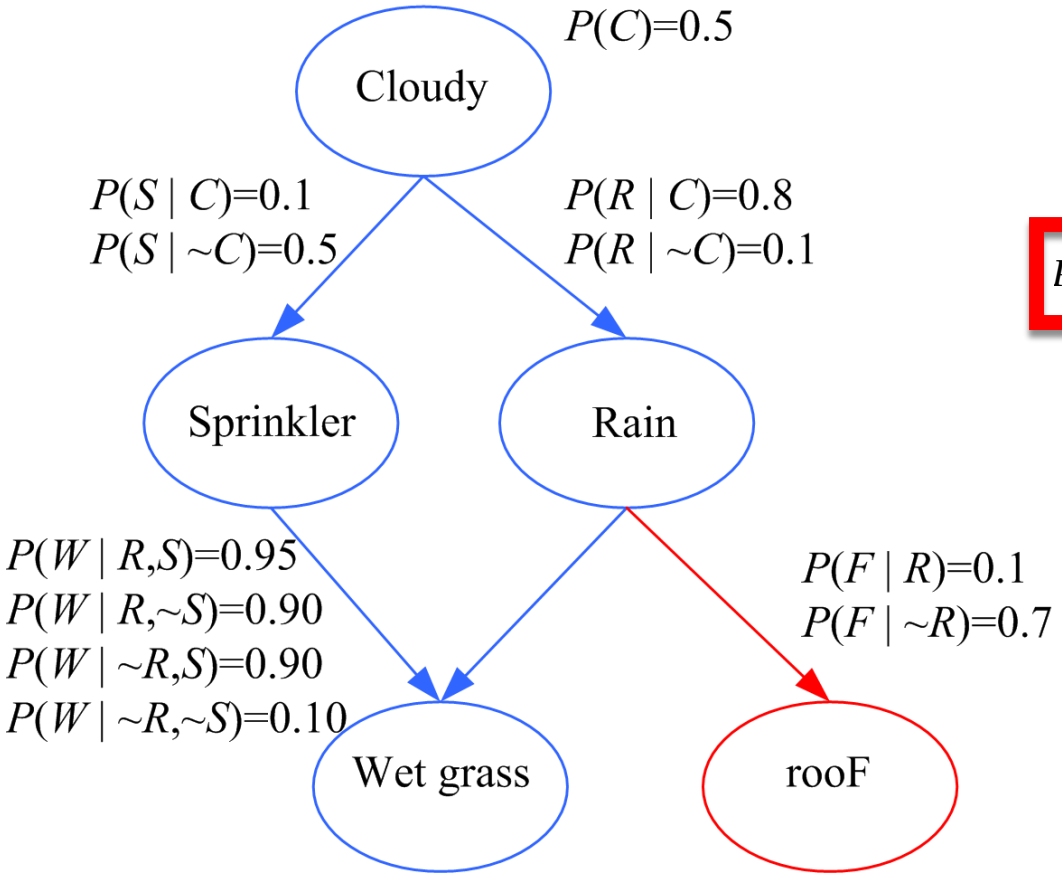
$$P(S|W) = \frac{P(W|S)P(S)}{P(W)} = \frac{0.92 * 0.2}{0.95 * 0.2 * 0.4 + 0.9 * 0.4 * 0.8 + 0.9 * 0.6 * 0.2 + 0.1 * 0.6 * 0.8} = 0.35$$

$$P(S|RW) = \frac{P(WRS)}{P(RW)} = \frac{P(W|RS)P(RS)}{P(W|R)P(R)} = \frac{P(W|RS)P(R)P(S)}{(P(W|SR)P(S|R) + P(W|\sim SR)P(\sim S|R))\,P(R)} = \frac{0.95 * 0.4 * 0.2}{(0.95 * 0.2 + 0.9 * 0.8) * 0.4} = 0.21$$

Bayesian Networks: Causes



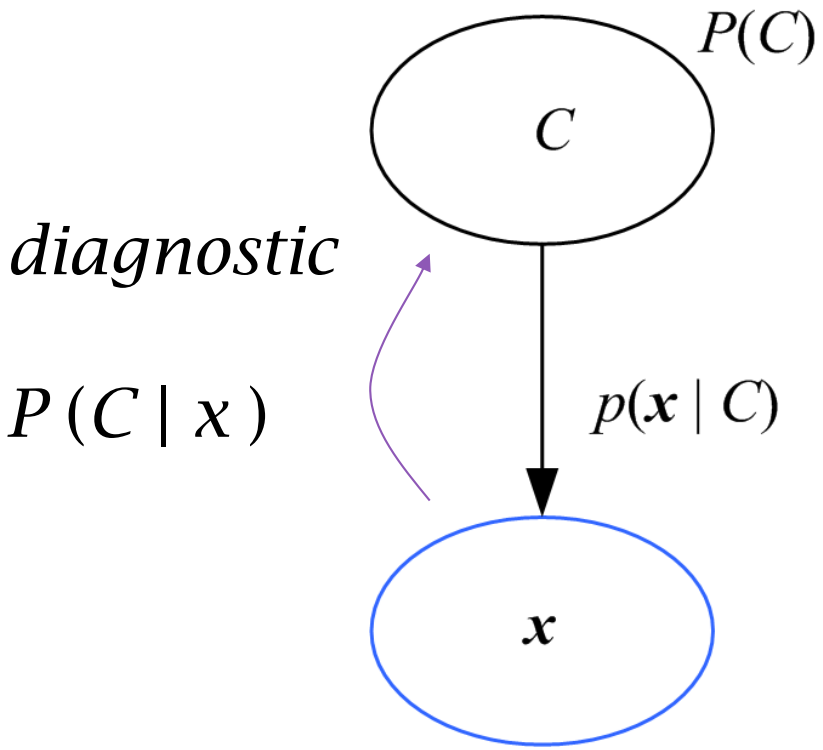
Bayesian Nets: Local structure



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

- Only need to specify interactions between neighboring events

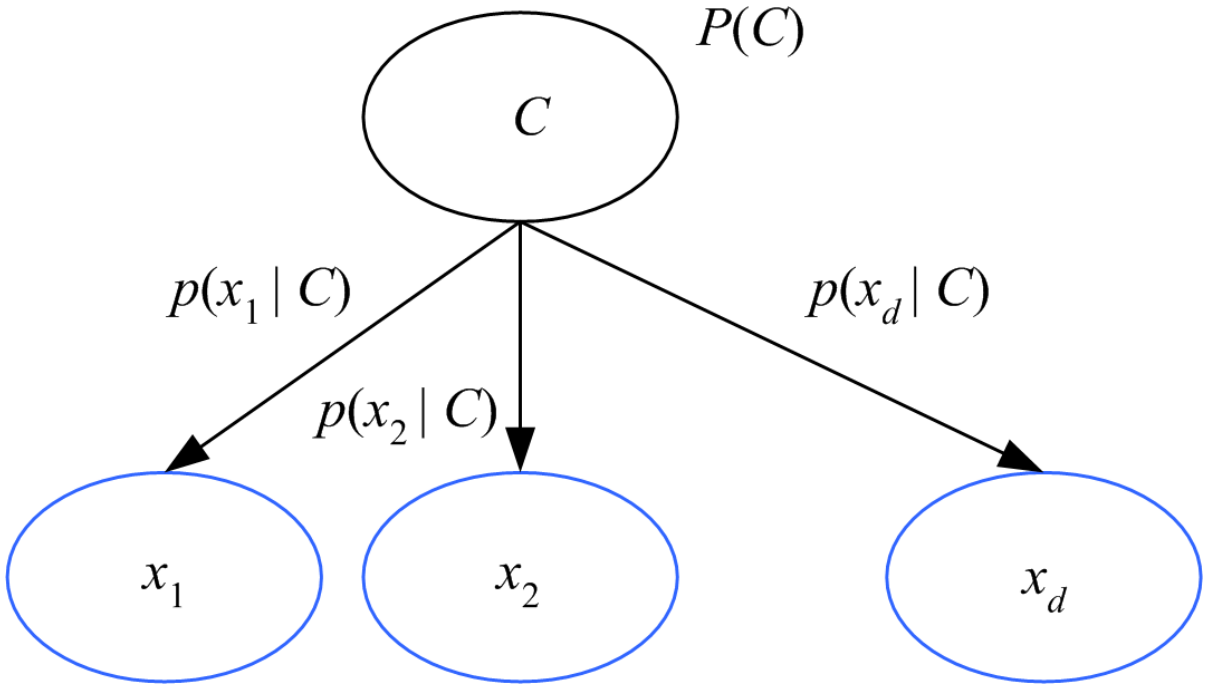
Bayesian Networks: Classification



Bayes' rule inverts the arc:

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

Naive Bayes' Classifier



Given C, x_j are independent:

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

Naïve Bayes' Classifier

- Why naïve?
- Ignores dependency among the inputs
- Dependency among "baby food" and "diapers"
- Hidden variables ("have children")
- Insert node/arcs and estimate their values from data

Association Rules

- Association rule: $X \rightarrow Y$
- *People who buy/click/visit/enjoy X are also likely to buy/click/visit/enjoy Y.*
- *X* is the "antecedent" and *Y* is the "consequent"
- A rule implies association, not necessarily causation.

Metrics for Association Rules

- Association rule: $X \rightarrow Y$
- **Support** ($X \rightarrow Y$): **Show statistical significance**

$$\text{Support}(X, Y) \equiv P(X, Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- **Confidence** ($X \rightarrow Y$): **Want large P(Y|X) and LARGER than P(Y)**
  Why?

$$\text{Confidence}(X \rightarrow Y) \equiv P(Y|X) = \frac{P(X, Y)}{P(X)}$$

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

Metrics for Association Rules

- **Lift:** $\text{Lift}(X \rightarrow Y) = \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)}$
  - If **X** and **Y** are *independent: Lift is (or close to) 1*
  - If "Lift" is more than 1:
    - X and Y are *dependent* and
    - X makes Y more likely
  - If "Lift" is less than 1:
    - X and Y are *dependent* and
    - X makes Y less likely
- Generalization to more than 2 variables, e.g.,
  - Given (X, Y, Z), find $X, Z \rightarrow Y$, that is, $P(Y|X,Z)$

Association Rule

- Only one customer bought chips
- **Same** customer bought beer
- $P(C|B) = 1$
- But support, or $P(C, B)$ is **SMALL**
- In other words, we need *Support* to show statistical significance

Examples of Association Measures

- **Support ($X \rightarrow Y$):**

$$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- **Confidence ($X \rightarrow Y$):**

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

- **Lift ($X \rightarrow Y$):**

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

$$= \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)}$$

Example

Transaction	*Items in basket*
1	*milk, bananas, chocolate*
2	*milk, chocolate*
3	*milk, bananas*
4	*chocolate*
5	*chocolate*
6	*milk, chocolate*

The association rules and their support and confidence values are as follows:

$$\text{milk} \rightarrow \text{bananas} : \text{Support} = 2/6, \text{Confidence} = 2/4 \quad \text{Lift}(X \rightarrow Y) = \frac{2/4}{2/6} = 1.5$$

$$\text{bananas} \rightarrow \text{milk} : \text{Support} = 2/6, \text{Confidence} = 2/2 \quad \text{Lift}(X \rightarrow Y) = \frac{1}{4/6} = 1.5$$

$$\text{milk} \rightarrow \text{chocolate} : \text{Support} = 3/6, \text{Confidence} = 3/4 \quad \text{Lift}(X \rightarrow Y) = \frac{3/4}{5/6} = 0.9$$

$$\text{chocolate} \rightarrow \text{milk} : \text{Support} = 3/6, \text{Confidence} = 3/5 \quad \text{Lift}(X \rightarrow Y) = \frac{3/5}{4/6} = 0.9$$

discuss

Apriori algorithm (Agrawal et al., 1996)

- For (X, Y, Z), a 3-item set, to be frequent (have enough support), (X, Y), (X, Z), and (Y, Z) should be frequent.
- If (X, Y) is not frequent, none of its supersets can be frequent.
- Once we find the frequent *k*-item sets, we convert them to rules: X, Y $\rightarrow$ Z, ...

  and X $\rightarrow$ Y, Z, ...

Apriori algorithm (Agrawal et al., 1996)

- **Step 1:**

  We start by finding the frequent one-item sets and at each step, inductively, from frequent k-item sets, we generate candidate $k+1$-item sets and then do a pass over the data to check if they have enough support.

Apriori algorithm (Agrawal et al., 1996)

- **Step 2:**
  - From step 1, we have frequent *k*-item sets.
  - Split *k* items into (*k-1*) antecedent and 1 consequent; check whether this association has enough confidence (remove if it does not).
  - Then check whether we can move one item from antecedent to consequent
    - To have rules with two items in the consequent with enough confidence, each of the two rules with single consequent by itself should have enough confidence;
  - Repeat

Conclusion

- Conditional probability & Bayes' rule
- Objective is ***posterior probability***, we need to get the (1) ***prior probability,*** (2) ***likelihood*** and (3) ***evidence***
- Bayesian modeling via the ***Graphical Model***
- **Casual** and **Diagnostic inference**
- Frequent items generation