CS 461: Machine Learning
Lecture 5

Dr. Kiri Wagstaff
kiri.wagstaff@calstatela.edu

Plan for Today

- Midterm Exam
- Notes
  - Room change for 2/16: E&T A129
  - Sign up for post-midterm conferences
  - Questions on Homework 3?
- Probability
  - Axioms
- Bayesian Learning
  - Classification
  - Bayes's Rule
  - Bayesian Networks
  - Naïve Bayes Classifier
  - Association Rules

Review from Lecture 4

- Support Vector Machines
  - Non-separable data: the Kernel Trick
  - Regression
- Neural Networks
  - Perceptrons
  - Multilayer Perceptrons
  - Backpropagation

Probability

Appendix A

Background and Axioms of Probability

- Random variable: X
- Probability: fraction of possible worlds where X is true

- Axioms
  - Positivity
  - Excluded Middle
  - Conjunction ("and")
  - Disjunction ("or")
- Conditional probabilities

Bayesian Learning

Chapter 3

Making Inferences with Probability

- Result of tossing a coin is $\in$ {Heads,Tails}
- Random var *Coin* $\in$ {Heads,Tails}
  - Bernoulli: $P\{Coin=Heads\} = p_o$, $P\{Coin=Tails\} = 1-p_o$
- Sample: $\boldsymbol{X} = \{x^t\}_{t=1}^N$
  - Estimation: p_o = #{Heads}/#{Tosses}
- 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: $C \in \{0,1\}$
- Prediction:

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

or equivalently

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

Bayes's Rule

prior *likelihood*

posterior

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

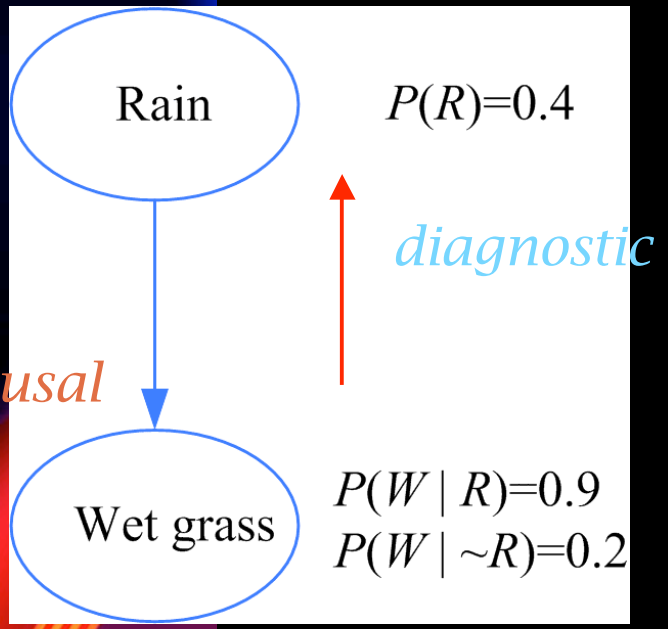
evidence

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

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

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

Causes and Bayes's Rule



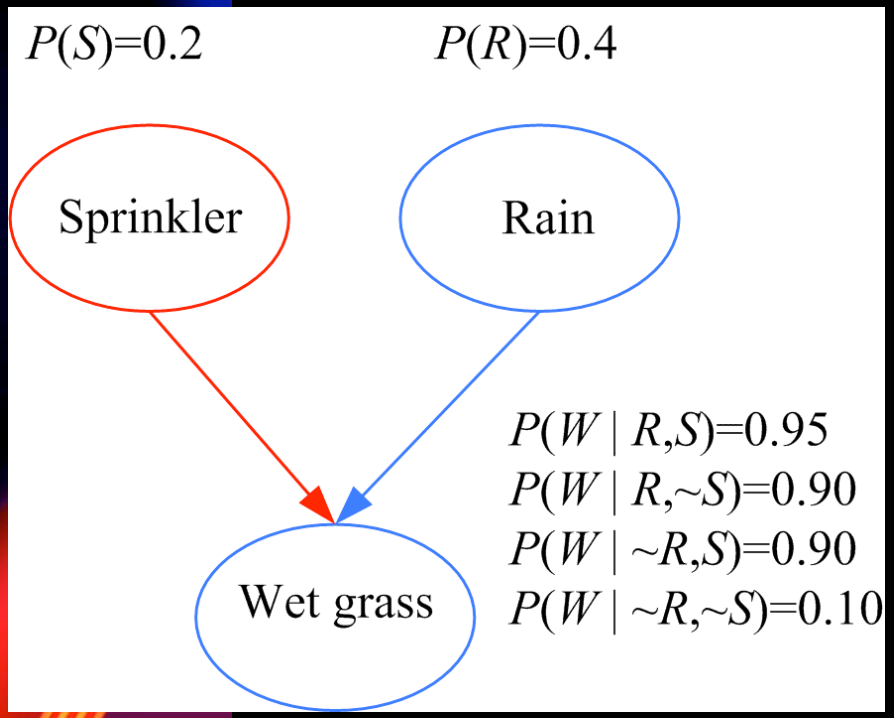
Diagnostic inference:
Knowing that the grass is wet, what is the probability that rain is the cause?

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

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

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

Causal vs. Diagnostic Inference

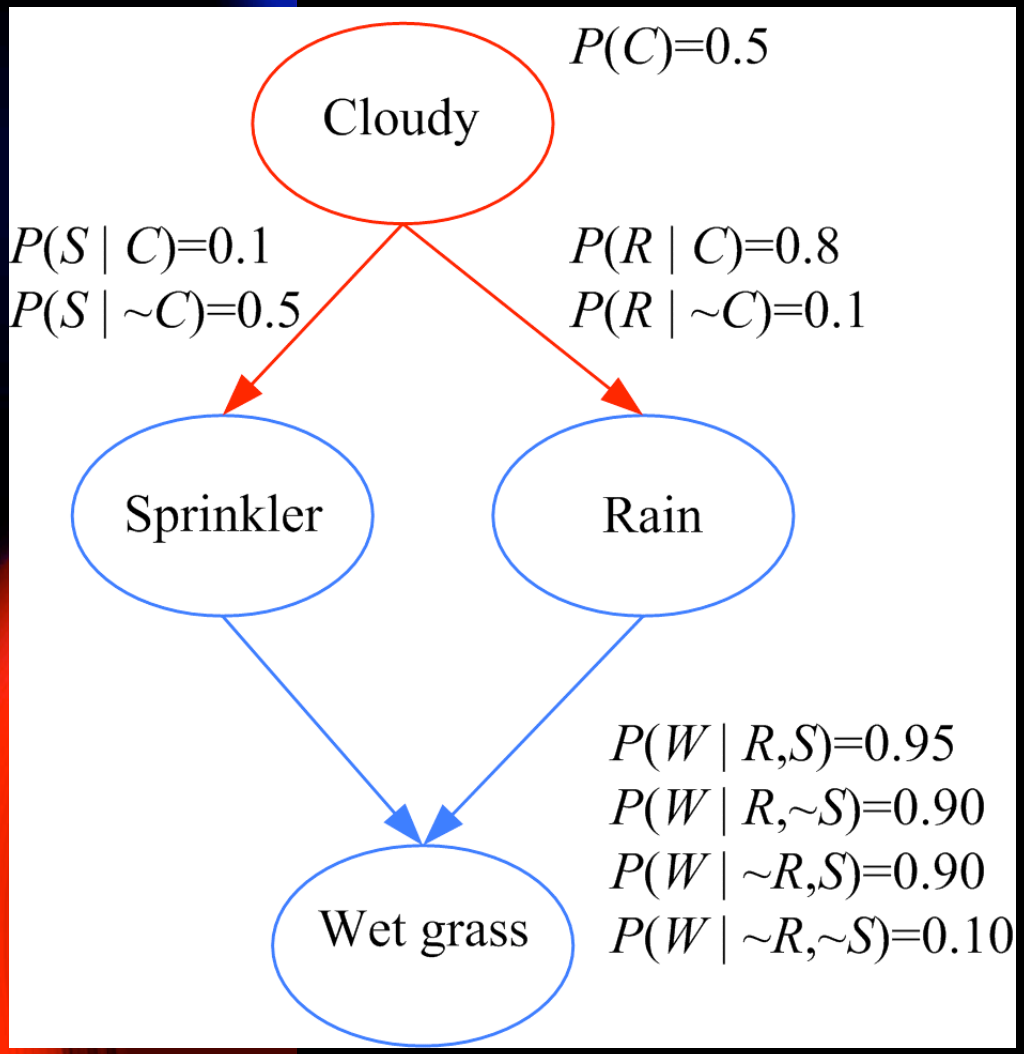


Causal inference:
If the sprinkler is on, what is the probability that the grass is wet?

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

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on?
$P(S|W) = 0.35 > 0.2$ $\qquad$ $P(S|R,W) = 0.21$
Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.

Bayesian Networks: Causes



Causal inference:

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

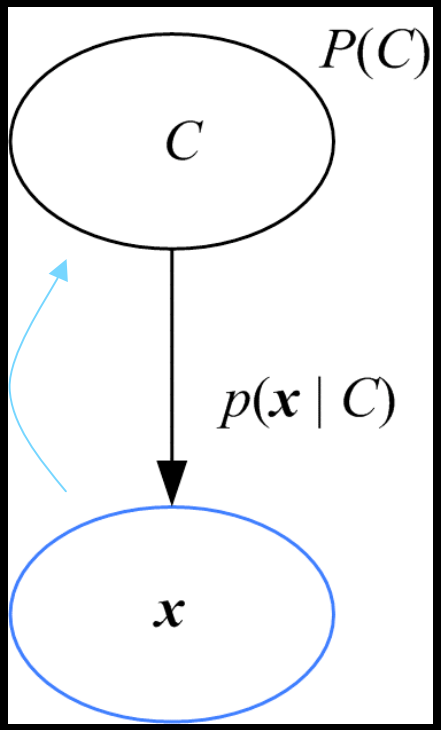
and use the fact that

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

Bayesian Networks: Classification

diagnostic

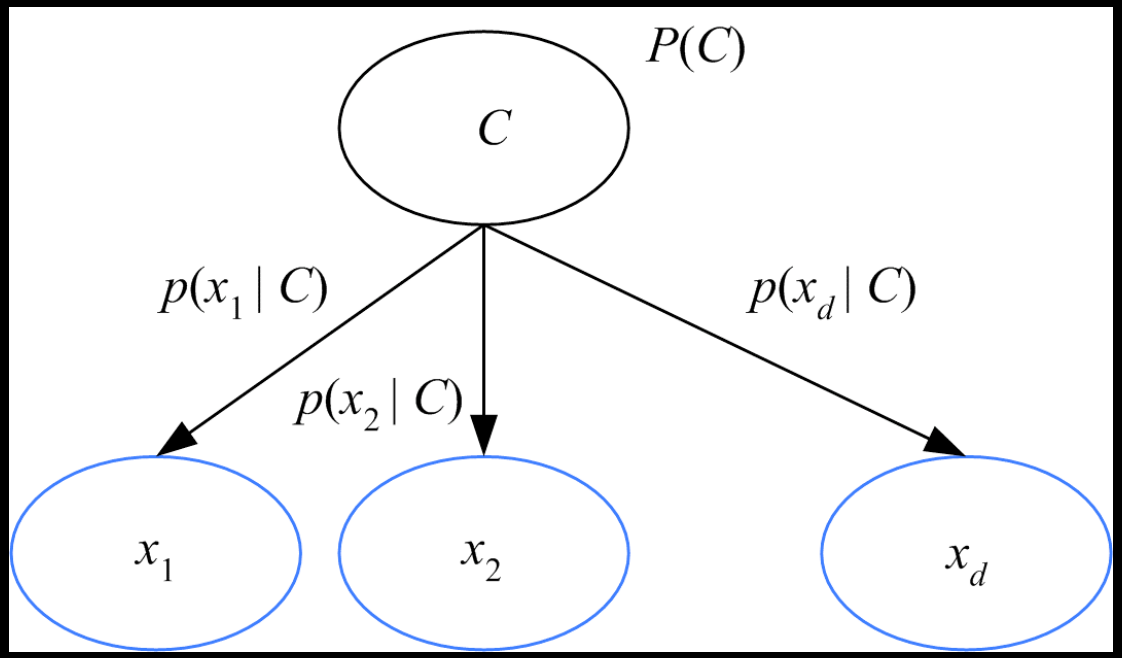
$P(C \mid x)$



Bayes rule inverts the arc:

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

Naïve Bayes… why "naïve"?



Given C, x_j are independent:

$$p(\mathbf{x}|C) = p(x_1|C)\ p(x_2|C)\ \ldots\ p(x_d|C)$$

Naïve Bayes Classifier

- The book's formulation = maximum likelihood estimator (MLE)

$$C^{\text{predict}} = \operatorname*{argmax}_{c} P(X_1 = u_1 \cdots X_m = u_m \mid C = c)$$

- More useful: maximum a-posteriori (MAP) classifier

$$C^{\text{predict}} = \operatorname*{argmax}_{c} P(C = c \mid X_1 = u_1 \cdots X_m = u_m)$$

MAP broken down

$$
\begin{aligned}
& P(C = c \mid X_1 = u_1 \cdots X_m = u_m) \\
= & \frac{P(X_1 = u_1 \cdots X_m = u_m \mid C = c) P(C = c)}{P(X_1 = u_1 \cdots X_m = u_m)} \\
= & \frac{P(X_1 = u_1 \cdots X_m = u_m \mid C = c) P(C = c)}{\sum_{j=1}^{k} P(X_1 = u_1 \cdots X_m = u_m \mid C = c_j) P(C = c_j)}
\end{aligned}
$$

Naïve Bayes in action

$$= \frac{P(X_1 = u_1 \cdots X_m = u_m \mid C = c)P(C = c)}{\sum_{j=1}^{k} P(X_1 = u_1 \cdots X_m = u_m \mid C = c_j)P(C = c_j)}$$

Train

Outlook	Temperature	Humidity	Wind	PlayTennis
Training Sets				
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Test

Outlook	Temperature	Humidity	Wind	PlayTennis
Testing Examples				
Sunny	Mild	Normal	Weak	Yes
Overcast	Hot	High	Strong	Yes
Sunny	Mild	High	Strong	No
Rain	Hot	High	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	High	Strong	Yes

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 ($X \rightarrow Y$):

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

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

Summary: Key Points for Today

- Probability
  - Axioms
- Bayesian Learning
  - Classification
  - Bayes's Rule
  - Bayesian Networks
  - Naïve Bayes Classifier
  - Association Rules

Next Time

- Parametric Methods
  (read Ch. 4.1-4.5, Mitchell p. 177-179)

- Questions to answer from the reading
  - Posted on the website (calendar)
    … if you find them useful?