CS 461: Machine Learning
Lecture 5

Dr. Kiri Wagstaff

wkiri@wkiri.com

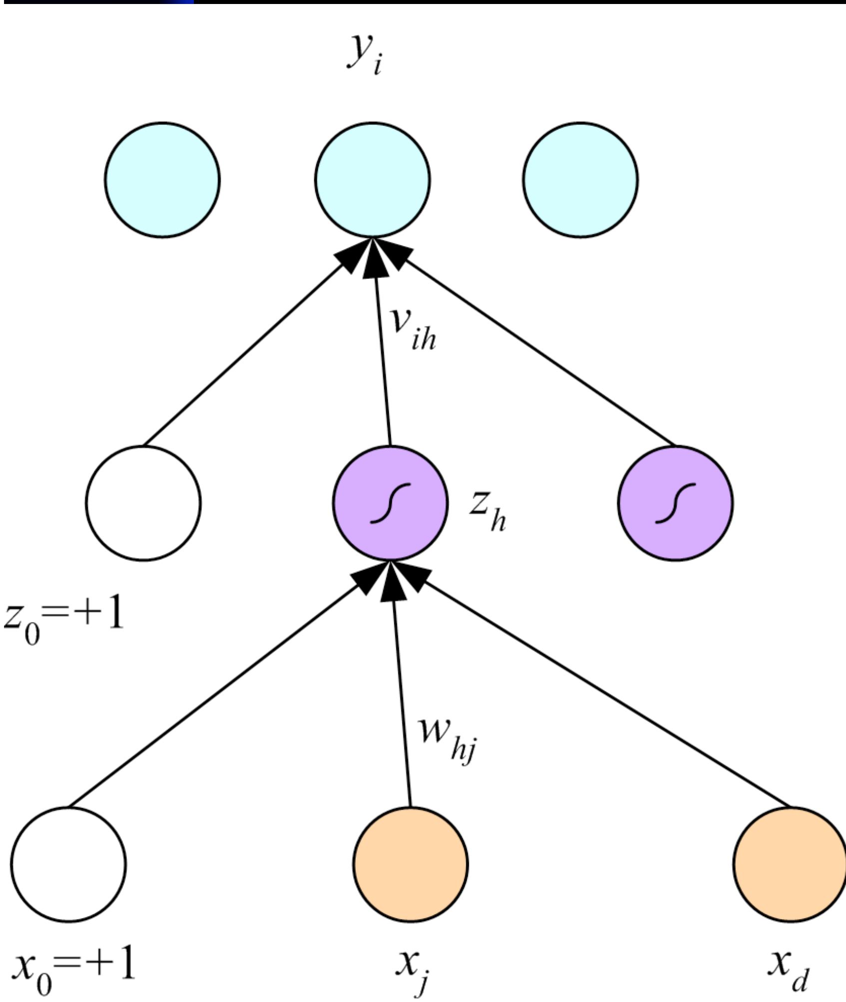
Plan for Today

- Midterm Exam
- Notes
  - Room change for 2/14: E&T A331
  - Reminder: post-midterm conferences (2/14)
  - Questions on Homework 3?
- MLP learning: Backpropagation
- Probability
  - Axioms
- Bayesian Learning
  - Bayes's Rule
  - Bayesian Networks
  - Naïve Bayes Classifier
  - Association Rules

Review from Lecture 4

- Neural Networks
  - Perceptrons
  - Multilayer Perceptrons

Backpropagation: MLP training



$$y_i = \mathbf{v}_i^T \mathbf{z} = \sum_{h=1}^{H} v_{ih} z_h + v_{i0}$$

$$z_h = \text{sigmoid}\left(\mathbf{w}_h^T \mathbf{x}\right)$$

$$= \frac{1}{1 + \exp\left[-\left(\sum_{j=1}^{d} w_{hj} x_j + w_{h0}\right)\right]}$$

$$\frac{\partial E}{\partial w_{hj}} = \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial z_h} \frac{\partial z_h}{\partial w_{hj}}$$

Backpropagation: Regression

$$E(\mathbf{W},\mathbf{v} \mid \mathcal{X}) = \frac{1}{2}\sum_t \left(y^t - \hat{y}^t\right)^2$$

$$y^t = \sum_{h=1}^{H} v_h z_h^t + v_0$$

$$\Delta v_h = \eta \sum_t \left(y^t - \hat{y}^t\right) z_h^t$$

Backward

Forward

$$z_h = \text{sigmoid}\left(\boldsymbol{w}_h^T \boldsymbol{x}\right)$$

$$\boldsymbol{x}$$

$$\begin{aligned}
\Delta w_{hj} &= -\eta \frac{\partial E}{\partial w_{hj}} \\
&= -\eta \sum_t \frac{\partial E}{\partial \hat{y}^t} \frac{\partial \hat{y}^t}{\partial z_h^t} \frac{\partial z_h^t}{\partial w_{hj}} \\
&= -\eta \sum_t -\left(y^t - \hat{y}^t\right) v_h z_h^t \left(1 - z_h^t\right) x_j^t \\
&= \eta \sum_t \left(y^t - \hat{y}^t\right) v_h z_h^t \left(1 - z_h^t\right) x_j^t
\end{aligned}$$

Backpropagation Algorithm

Initialize all v_{ih} and w_{hj} to rand$(-0.01, 0.01)$
Repeat
    For all $(\boldsymbol{x}^t, r^t) \in \mathcal{X}$ in random order
        For $h = 1, \ldots, H$
            $z_h \leftarrow \text{sigmoid}(\boldsymbol{w}_h^T \boldsymbol{x}^t)$
        For $i = 1, \ldots, K$
            $y_i = \boldsymbol{v}_i^T \boldsymbol{z}$
        For $i = 1, \ldots, K$
            $\Delta \boldsymbol{v}_i = \eta(r_i^t - y_i^t)\boldsymbol{z}$
        For $h = 1, \ldots, H$
            $\Delta \boldsymbol{w}_h = \eta(\sum_i (r_i^t - y_i^t) v_{ih}) z_h (1 - z_h) \boldsymbol{x}^t$
        For $i = 1, \ldots, K$
            $\boldsymbol{v}_i \leftarrow \boldsymbol{v}_i + \Delta \boldsymbol{v}_i$
        For $h = 1, \ldots, H$
            $\boldsymbol{w}_h \leftarrow \boldsymbol{w}_h + \Delta \boldsymbol{w}_h$
Until convergence

Probability

Appendix A

Background and Axioms of Probability

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

- Axioms
  - Positivity
  - Conjunction (“and”)
  - Disjunction (“or”)
- Conditional probabilities

Bayesian Learning

Chapter 3

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

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

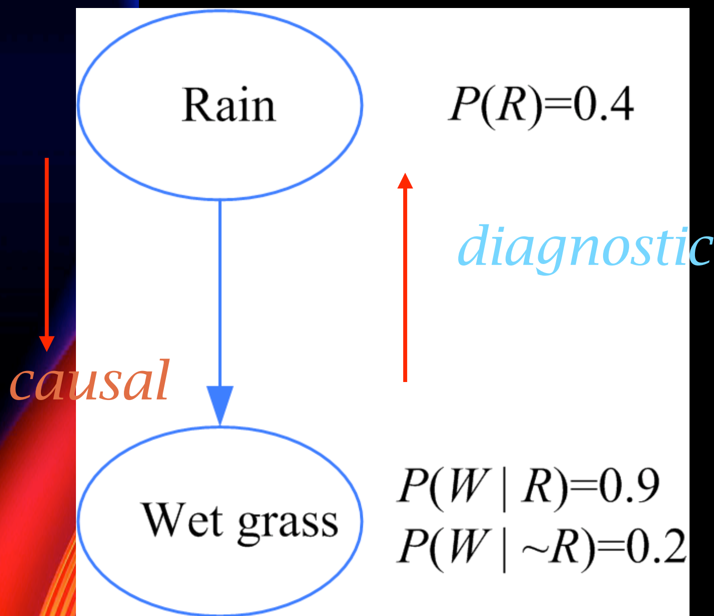
evidence

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

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

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

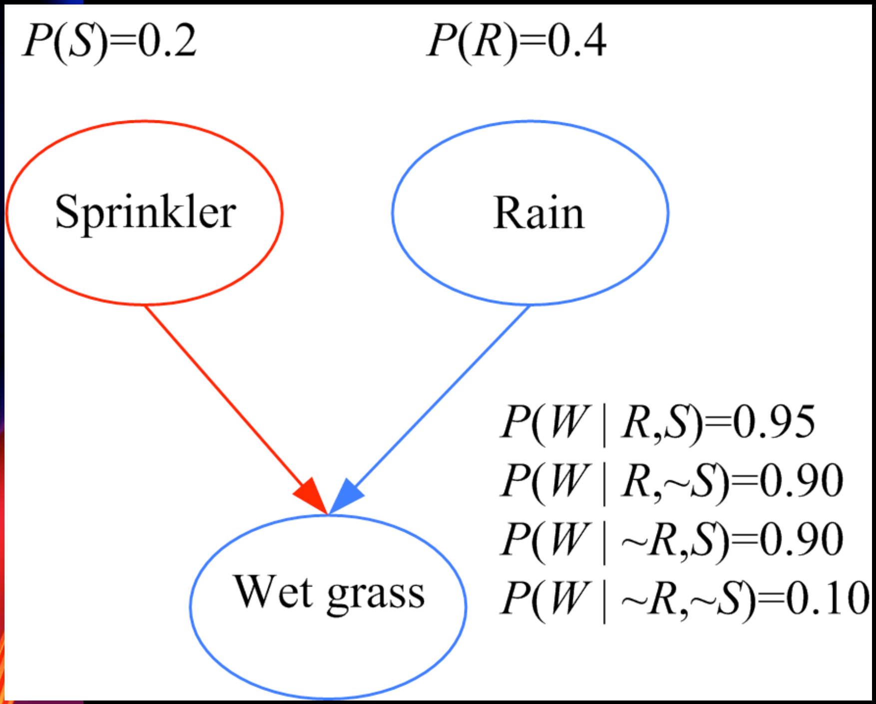
Causes and Bayes's Rule



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

$$
\begin{aligned}
P(R \mid W) &= \frac{P(W \mid R)P(R)}{P(W)} \\
&= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\end{aligned}
$$

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?

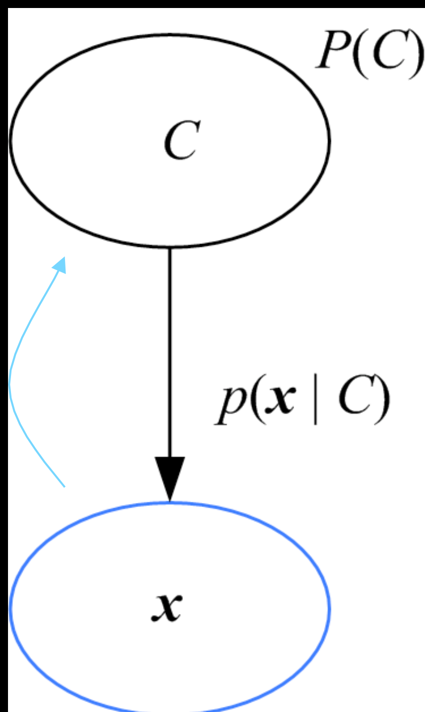
$P(S|W) = 0.35 > 0.2$ $\qquad$ $P(S|R,W) = 0.21$

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

Bayesian Networks: Classification

diagnostic

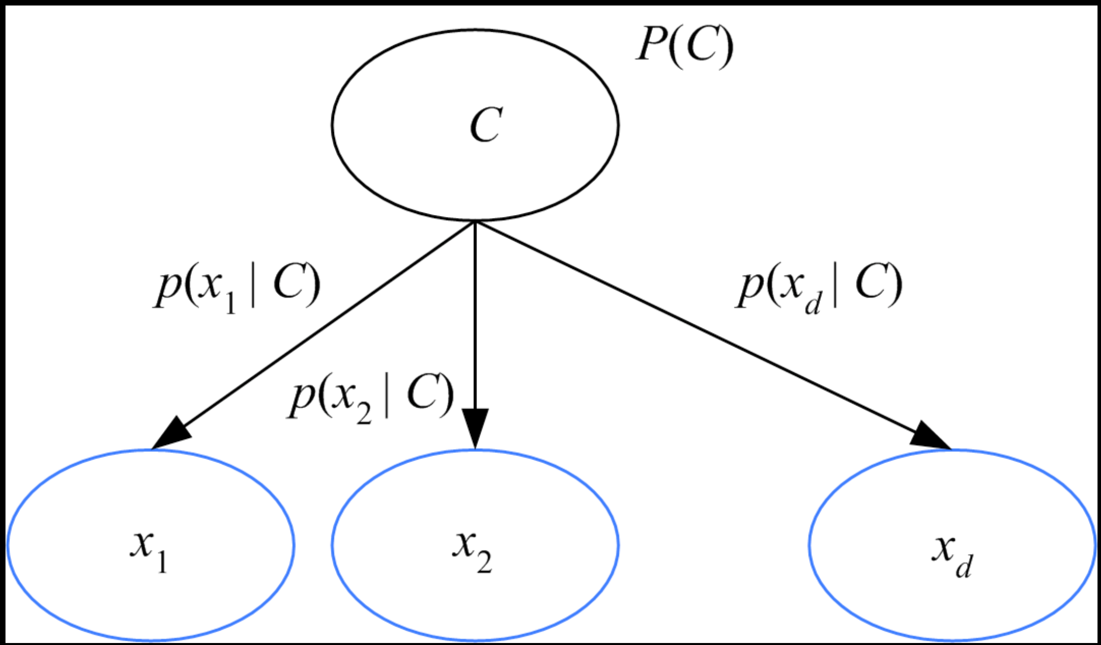
$P(C \mid x)$



Bayes rule inverts the arc:

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

Naïve Bayes… why "naïve"?



Given C, x_j are independent:

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

 [Alpaydin 2004 © The MIT Press]

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

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

Next Time

- Reading: Probability and Bayesian Learning (read Appendix A, Ch. 3.1, 3.2, 3.7, 3.9)
- Questions to answer from the reading
  - Volunteers: Herman, Sam, Sassja
- Class will be in E&T A331