Lecture 8: Graphical Models II
Machine Learning

Andrew Rosenberg

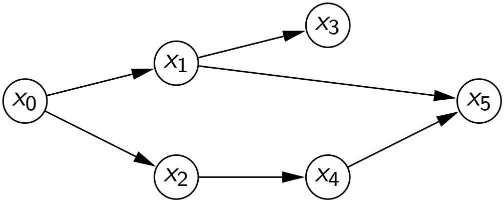
March 5, 2010

Today

- Graphical Models
  - Naive Bayes classification
  - Conditional Probability Tables (CPTs)
  - Inference in Graphical Models and Belief Propagation

Graphical Models

- Graphical representation of the dependency relationships between random variables.



Topological Graph

Graphical models factorize probabilities

$$p(x_0, \ldots, x_{n-1}) = \prod_{i=0}^{n-1} p(x_i|pa_i) = \prod_{i=0}^{n-1} p(x_i|\pi_i)$$

Nodes are generally topologically ordered so that parents, π come before children.

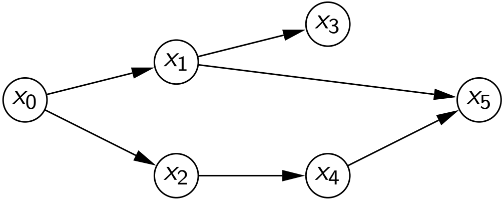
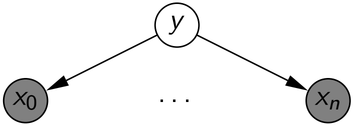
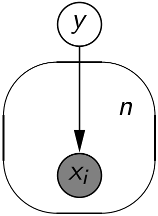


Plate Notation of a Graphical Model

Recall the Naive Bayes Graphical Model

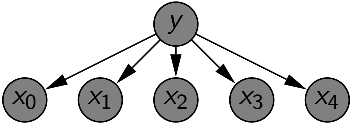


There can be many variables x_i.
Plate notation gives a compact representation of models like this:



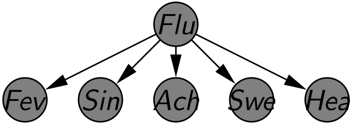
Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	Y	N	Y	Y
Y	M	Y	N	N	Y



Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y

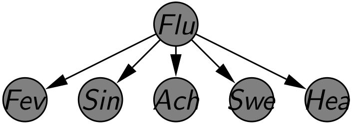


$p(flu) =$

Y	N
.75	.25

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y



$p(flu) =$

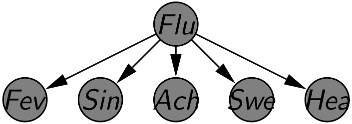
Y	N
.75	.25

$p(fev|flu) =$

	L	M	H
Y	.33	.33	.33
N	0	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y



$p(flu) =$

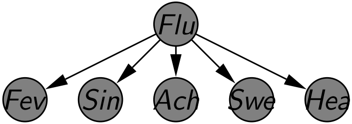
Y	N
.75	.25

$p(sinus|flu) =$

	Y	N
Y	.67	.33
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y



$p(flu) =$

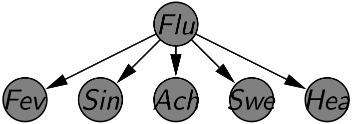
Y	N
.75	.25

$p(ache|flu) =$

	Y	N
Y	.33	.67
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y



$p(flu) =$

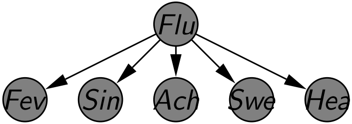
Y	N
.75	.25

$p(swell|flu) =$

	Y	N
Y	.67	.33
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y



$p(flu) =$

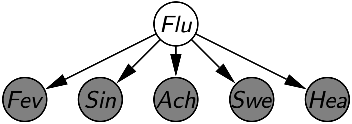
Y	N
.75	.25

$p(head|flu) =$

	Y	N
Y	.67	.33
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



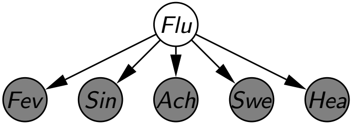
Find $p(flu|fever, sinus, ache, swell, head)$.
$p(flu = Y)p(fev = M|flu = Y)p(sin = N|flu = Y)p(ach = N|flu = Y)p(swe = N|flu = Y)p(head = N|flu = Y)$

$p(flu) =$

Y	N
.75	.25

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



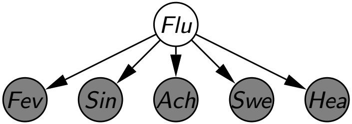
Find $p(flu|fever, sinus, ache, swell, head)$.
$.75 * p(fev = M|flu = Y)p(sin = N|flu = Y)p(ach = N|flu = Y)p(swe = N|flu = Y)p(head = N|flu = Y)$

$p(fev|flu) =$

	L	M	H
Y	.33	.33	.33
N	0	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



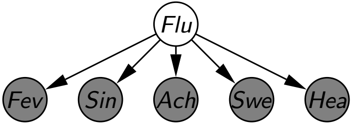
Find $p(flu|fever, sinus, ache, swell, head)$.
$.75 * .33 * p(sin = N|flu = Y)p(ach = N|flu = Y)p(swe = N|flu = Y)p(head = N|flu = Y)$

$p(sinus|flu) =$

	Y	N
Y	.67	.33
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



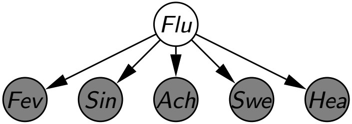
Find $p(flu|fever, sinus, ache, swell, head)$.
$.75 * .33 * .33 * p(ach = N|flu = Y)p(swe = N|flu = Y)p(head = N|flu = Y)$

$p(ache|flu) =$

	Y	N
Y	.33	.67
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



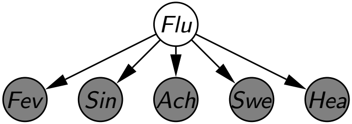
Find $p(flu|fever, sinus, ache, swell, head)$.
$.75 * .33 * .33 * .67 * p(swe = N|flu = Y)p(head = N|flu = Y)$

$p(swell|flu) =$

	Y	N
Y	.67	.33
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



Find $p(flu|fever, sinus, ache, swell, head)$.

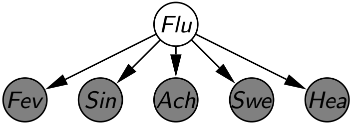
$.75 * .33 * .33 * .67 * .33 * p(head = N|flu = Y)$

$p(head|flu) =$

	Y	N
Y	.67	.33
N	1	0

Naive Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y
?	M	N	N	N	N



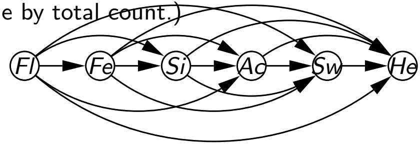
Find $p(flu|fever, sinus, ache, swell, head)$.
$.75 * .33 * .33 * .67 * .33 * .33 = 0.0060$

Suppose we have observations for every node.

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	N	N	Y	Y
Y	M	Y	N	N	Y

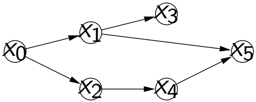
In the simplest – least general – graph, assume each independent. Train 6 separate models.

2nd simplest graph – most general – assume no independence. Build a 6-dimensional table. (Divide by total count.)



Maximum Likelihood Conditional Probability Tables

Consider this Graphical Model



- Each node has a conditional probability table θ_i.
- Given the table, we have a pdf

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=0}^{M-1} p(x_i|\pi_i, \theta_i)$$

- We have m variables in $\mathbf{x}$, and N data points, $\mathbf{X}$.
- Maximum (log) Likelihood

$$\begin{aligned} \theta^* &= \operatorname*{argmax}_\theta \ln p(\mathbf{X}|\theta) \\ &= \operatorname*{argmax}_\theta \sum_{n=0}^{N-1} \ln p(\mathbf{X}_n|\theta) \end{aligned}$$

$$\begin{aligned} &= \operatorname*{argmax}_\theta \sum_{n=0}^{N-1} \ln \prod_{i=0}^{M-1} p(x_{in}|\theta_i) \\ &= \operatorname*{argmax}_\theta \sum_{n=0}^{N-1} \sum_{i=0}^{M-1} \ln p(x_{in}|\theta_i) \end{aligned}$$

Maximum Likelihood CPTs

First, Kronecker's delta function.

$$\delta(x_n, x_m) = \begin{cases} 1 & \text{if } x_n = x_m \\ 0 & \text{otherwise} \end{cases}$$

Counts: the number of times something appears in the data

$$\begin{aligned} m(x_i) &= \sum_{n=0}^{N-1} \delta(x_i, x_{in}) \\ m(\mathbf{X}) &= \sum_{n=0}^{N-1} \delta(\mathbf{X}, \mathbf{X}_n) \end{aligned}$$

$$N = \sum_{x_1} m(x_1) = \sum_{x_1} \left(\sum_{x_2} \delta(x_1, x_2) \right) = \sum_{x_1} \left(\sum_{x_2} \left(\sum_{x_3} \delta(x_1, x_2, x_3) \right) \right) \ldots$$

Maximum likelihood CPTs

$$
\begin{aligned}
l(\theta) &= \sum_{n=0}^{N-1} \ln p(\mathbf{X}_n|\theta) \\
&= \sum_{n=0}^{N-1} \ln \prod_{\mathbf{X}} p(\mathbf{X}|\theta)^{\delta(\mathbf{x}_n,\mathbf{X})} \\
&= \sum_{n=0}^{N-1} \sum_{\mathbf{X}} \delta(\mathbf{x}_n, \mathbf{X}) \ln p(\mathbf{X}|\theta) \\
&= \sum_{x_n} m(\mathbf{X}) \ln p(\mathbf{X}|\theta)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x_n} m(\mathbf{X}) \ln \prod_{i=0}^{M-1} p(x_i|\pi_i, \theta_i) \\
&= \sum_{x_n} \sum_{i=0}^{M-1} m(\mathbf{X}) \ln p(x_i|\pi_i, \theta_i) \\
&= \sum_{i=0}^{M-1} \sum_{x_i,\pi_i} \sum_{\mathbf{X}\setminus x_i \setminus \pi_i} m(\mathbf{X}) \ln p(x_i|\pi_i, \theta_i) \\
&= \sum_{i=0}^{M-1} \sum_{x_i,\pi_i} m(x_i, \pi_i) \ln p(x_i|\pi_i, \theta_i)
\end{aligned}
$$

Define a function:

$$\theta(x_i, \pi_i) = p(x_i|\pi_i, \theta_i)$$

Constraint:

$$\sum_{x_i} \theta(x_i, \pi_i) = 1$$

Lagrange Multipliers

To maximize $f(x, y)$subject to $g(x, y) = c$.
Maximize $f(x, y) - \lambda(g(x, y) - c)$

$$
\begin{aligned}
l(\theta) &= \sum_{i=0}^{M-1} \sum_{x_i} \sum_{\pi_i} m(x_i, \pi_i) \ln \theta(x_i, \pi_i) - \sum_{i=0}^{M-1} \sum_{\pi_i} \lambda_{\pi_i} \left(\sum_{x_i} \theta(x_i, \pi_i) - 1 \right) \\
\frac{\partial l(\theta)}{\partial \theta(x_i, \pi_i)} &= \frac{m(x_i, \pi_i)}{\theta(x_i, \pi_i)} - \lambda_{\pi_i} = 0 \\
\theta(x_i, \pi_i) &= \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} \\
\sum_{x_i} \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} &= 1 \text{ – the constraint} \\
\lambda_{\pi_i} &= \sum_{x_i} m(x_i, \pi_i) = m(\pi_i) \\
\theta(x_i, \pi_i) &= \frac{m(x_i, \pi_i)}{m(\pi_i)} \text{ – counts!}
\end{aligned}
$$

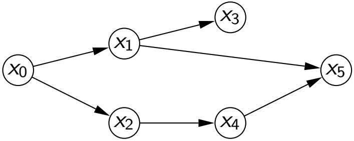
Maximum A Posteriori CPT training

For the Bayesians, MAP leads to:

$$\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i) + \epsilon}{m(\pi_i) + \epsilon|x_i|}$$

Example of maximum likelihood.

Flu (x_0)	Fever (x_1)	Sinus (x_2)	Ache (x_3)	Swell (x_4)	Head (x_5)
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	Y	N	Y	Y
Y	M	Y	N	N	Y



$$\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{m(\pi_i)}$$

Conditional Dependence Test.

- We also what to be able to check conditional independencies in a graphical model.
- i.e. "Is achiness (x_3) independent of flu (x_0) given fever (x_1)?"
- i.e. "Is achiness (x_3) independent of sinus infection (x_2) given fever (x_1)?"

$$p(x) = p(x_0)p(x_1|x_0)p(x_2|x_0)p(x_3|x_1)p(x_4|x_2)p(x_5|x_1,x_4)$$

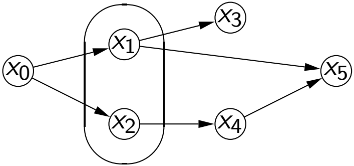
$$\begin{aligned}
p(x_3|x_0,x_1,x_2) &= \frac{p(x_0,x_1,x_2,x_3)}{p(x_0,x_1,x_2)} \\
&= \frac{p(x_0)p(x_1|x_0)p(x_2|x_0)p(x_3|x_1)}{p(x_0)p(x_1|x_0)p(x_2|x_0)} \\
&= p(x_3|x_1)
\end{aligned}$$

$$x_3 \perp\!\!\!\perp x_0, x_2 | x_1$$

No problem, right?

What about $x_0 \perp\!\!\!\perp x_5 | x_1, x_2$?

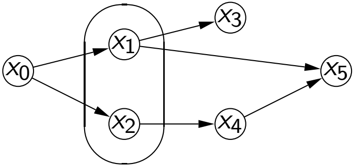
D-separation and Bayes Ball



Intuition: nodes are **separated**, or **blocked** by sets of nodes

- Example: nodes x_1 and x_2, "block" the path from x_0 to x_5 , then $x_0 \perp\!\!\!\perp x_5 | x_2, x_3$

D-separation and Bayes Ball



Intuition: nodes are **separated**, or **blocked** by sets of nodes

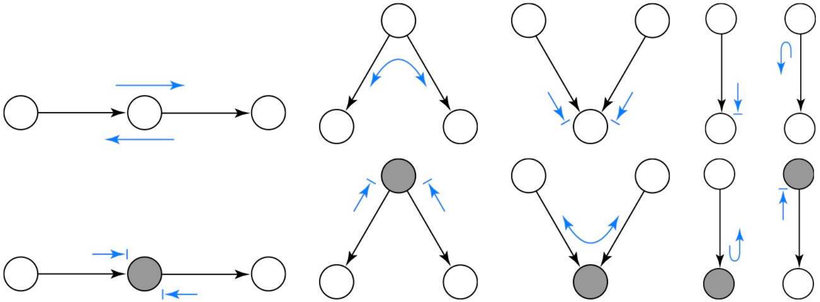
- Example: nodes x_1 and x_2, "block" the path from x_0 to x_5 , then $x_0 \perp\!\!\!\perp x_5 | x_2, x_3$
- While this is true in **undirected** graphs, it is not in directed graphs.
- We need more than simple **Separation**
- We need directed separation – **D-Separation**
- the D-separation is computed using the **Bayes Ball** algorithm.
- Allows us to prove general statements $x_a \perp\!\!\!\perp x_b | x_c$.

Bayes Ball Algorithm

$$x_a \perp\!\!\!\perp x_b | x_c$$

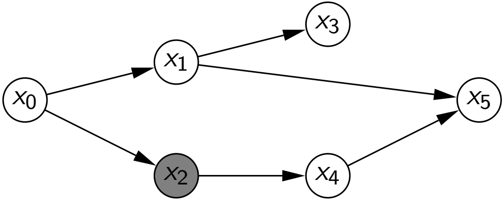
- Shade nodes x_c
- Place a "ball" at each node in x_a
- Bounce balls around the graph according to some rules
- If no balls read x_b, then $x_a \perp\!\!\!\perp x_b | x_c$, else false.
- Balls can travel along/against edges
- Pick any path
- Test to see if the ball goes through or bounces back.

Ten Rules of Bayes Ball



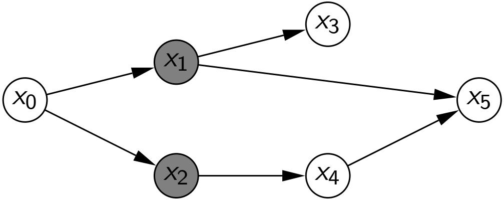
Bayes Ball Example - I

$$x_0 \perp\!\!\!\perp x_4 | x_2?$$

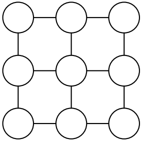


Bayes Ball Example - II

$$x_0 \perp\!\!\!\perp x_5 | x_1, x_2 ?$$

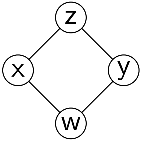


Undirected Graphs



- What if we allow undirected graphs?
- What do they correspond to?
- It's not cause/effect, or trigger/response, rather, general dependence.
- Example: Image pixels, where each pixel is a bernouli.
- Can have a probability over all pixels $p(x_{11}, x_{1M}, x_{M1}, x_{MM})$
- Bright pixels have bright neighbors.
- No parents, just probabilities.
- Grid models are called **Markov Random Fields**.

Undirected Graphs



$$x \perp\!\!\!\perp y | \{w, x\}$$

- Undirected separation is easy.
- To check $x_a \perp\!\!\!\perp x_b | x_c$, check Graph reachability of x_a and x_b without going through nodes in x_c.

Bye

- Next
  - Representing probabilities in Undirected Graphs.