Bayesian Networks

David Rosenberg

New York University

June 9, 2015
Probabilistic Reasoning

- **Represent** system of interest by a set of random variables
  \[(X_1, \ldots, X_d)\].

- Suppose by research or machine learning, we get a joint probability distribution
  \[p(x_1, \ldots, x_d)\].

- We’d like to be able to do **inference** on this model – essentially, answer queries:
  1. What is the most likely value of \(X_1\)?
  2. What is the most likely value of \(X_1\), given we’ve observed \(X_2 = 1\)?
  3. Distribution of \((X_1, X_2)\) given observation of \((X_3 = x_3, \ldots, X_d = x_d)\)?
Example: Medical Diagnosis

- **Variables for each symptom**
  - fever, cough, fast breathing, shaking, nausea, vomiting

- **Variables for each disease**
  - pneumonia, flu, common cold, bronchitis, tuberculosis

- Diagnosis is performed by **inference** in the model:

\[ p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0) \]

- The QMR-DT (Quick Medical Reference - Decision Theoretic) has
  - 600 diseases
  - 4000 symptoms

---

Example from David Sontag’s *Inference and Representation*, Lecture 1.
Some Notation

- This lecture we’ll only be considering discrete random variables.
- Capital letters $X_1, \ldots, X_d, Y$, etc. denote random variables.
- Lower case letters $x_1, \ldots, x_n, y$ denote the values taken.
- Probability that $X_1 = x_1$ and $X_2 = x_2$ will be denoted
  \[ P(X_1 = x_1, X_2 = x_2). \]
- We’ll generally write things in terms of the probability mass function:
  \[ p(x_1, x_2, \ldots, x_d) := P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d) \]
Let’s consider the case of discrete random variables. Conceptually, everything can be represented with probability tables.

Variables

- Temperature $T \in \{\text{hot, cold}\}$
- Weather $W \in \{\text{sun, rain}\}$

<table>
<thead>
<tr>
<th>$t$</th>
<th>$p(t)$</th>
<th>$w$</th>
<th>$p(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>0.5</td>
<td>sun</td>
<td>0.6</td>
</tr>
<tr>
<td>cold</td>
<td>0.5</td>
<td>rain</td>
<td>0.4</td>
</tr>
</tbody>
</table>

These are the marginal probability distributions.

To do reasoning, we need the joint probability distribution.

---

Based on David Sontag’s *DS-GA 1003 Lectures, Spring 2014*, Lecture 10.
A joint probability distribution for $T$ and $W$ is given by

<table>
<thead>
<tr>
<th>$t$</th>
<th>$w$</th>
<th>$p(t,w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

A valid probability distribution if

1. $\forall t, w: p(t, w) \geq 0$
2. $\sum_{t, w} p(t, w) = 1$. 

Based on David Sontag’s DS-GA 1003 Lectures, Spring 2014, Lecture 10.
Conditional Distributions From the Joint Distribution

- We observe $T = \text{hot}$. What’s the conditional distribution of $W$?

$$p(w \mid T = \text{hot}) = ?$$

- Method:
  1. Find entries in joint distribution table where $T = \text{hot}$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$w$</th>
<th>$p(t, w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
</tbody>
</table>

  2. Renormalize to get conditional probability.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$w$</th>
<th>$p(t, w)$</th>
<th>$p(w \mid T = \text{hot})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td>0.4/0.5 = 0.8</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td>0.1/0.5 = 0.2</td>
</tr>
</tbody>
</table>
Conditional Distributions From the Joint Distribution

Definition

The **conditional probability** for \( w \) given \( t \) is

\[
p(w \mid t) = \frac{p(w, t)}{p(t)}.
\]

<table>
<thead>
<tr>
<th>( t )</th>
<th>( w )</th>
<th>( p(t, w) )</th>
<th>( p(w \mid T = \text{hot}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td>( 0.4/0.5 = 0.8 )</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td>( 0.1/0.5 = 0.2 )</td>
</tr>
</tbody>
</table>
Consider random variables $X_1, \ldots, X_d \in \{0, 1\}$.

How many parameters do we need to represent the joint distribution?

Joint probability table has $2^d$ rows.

For QMR-DT, that’s $2^{4600} > 10^{1000}$ rows.

That’s not going to happen.

Having exponentially many parameters is a problem for

- storage
- computation (inference is summing over exponentially many rows)
- statistical estimation / learning
  - (Estimating $10^{1000}$ parameters? Nope.)
How to Restrict the Complexity?

- Restrict the space of probability distributions
- We will make various independence assumptions.
- Extreme assumption: $X_1, \ldots, X_d$ are mutually independent.

**Definition**

Discrete random variables $X_1, \ldots, X_d$ are **mutually independent** if their joint probability mass function (PMF) factorizes as

$$p(x_1, x_2, \ldots, x_d) = p(x_1)p(x_2)\cdots p(x_d).$$

- Note: We usually just write independent for “mutually independent”.
- How many parameters to represent the joint distribution, assuming independence?
Assume Full Independence

- How many parameters to represent the joint distribution?
- Say $p(X_i = 1) = \theta_i$, for $i = 1, \ldots, d$.
- **Clever representation**: Since $x_i \in \{0, 1\}$, we can write
  \[
P(X_i = x_i) = \theta_i^{x_i} (1 - \theta_i)^{1-x_i}.
\]
- Then by independence,
  \[
p(x_1, \ldots, x_d) = \prod_{i=1}^{d} \theta_i^{x_i} (1 - \theta_i)^{1-x_i}
\]
- How many parameters?
- $d$ parameters needed to represent the joint.
Suppose $X$ and $Y$ are independent, then

$$p(x \mid y) = p(x).$$

Proof:

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x).$$

With full independence, we have no relationships among variables.
Information about one variable says nothing about any other variable.
Would mean diseases don’t have symptoms.
Consider 3 events:

1. $W = \{\text{The grass is wet}\}$
2. $S = \{\text{The road is slippery}\}$
3. $R = \{\text{It’s raining}\}$

These events are certainly not independent.

- Raining ($R$) $\implies$ Grass is wet AND The road is slippery ($W \cap S$)
- Grass is wet ($W$) $\implies$ More likely that the road is slippery ($S$)

Suppose we know that it’s raining.

- Then, we learn that the grass is wet.
- Does this tell us anything new about whether the road is slippery?

Once we know $R$, then $W$ and $S$ become independent.

This is called conditional independence, and we’ll denote it as

$$W \perp S \mid R.$$
Conditional Independence

Definition

We say $W$ and $S$ are **conditionally independent** given $R$, denoted

$$W \perp S \mid R,$$

if the conditional joint factorizes as

$$p(w, s \mid r) = p(w \mid r)p(s \mid r).$$

Also holds when $W$, $S$, and $R$ represent **sets of random variables**.
Example: Rainy, Slippery, Wet

- Consider 3 events:
  1. \( W = \{ \text{The grass is wet} \} \)
  2. \( S = \{ \text{The road is slippery} \} \)
  3. \( R = \{ \text{It’s raining} \} \)

- Represent joint distribution as

\[
p(w, s, r) = p(w, s | r)p(r) \quad \text{(no assumptions so far)}
\]
\[
= p(w | r)p(s | r)p(r) \quad \text{(assuming } W \perp S | R)\]

- How many parameters to specify the joint?
  - \( p(w | r) \) requires two parameters: one for \( r = 1 \) and one for \( r = 0 \).
  - \( p(s | r) \) requires two.
  - \( p(r) \) requires one parameter,

Bayesian Networks: Introduction

- Bayesian Networks are
  - used to specify joint probability distributions that
  - have a particular factorization.

\[
p(c, h, a, i) = p(c)p(a) \times p(h | c, a)p(i | a)
\]

- With practice, one can read conditional independence relationships directly from the graph.

From Percy Liang's "Lecture 14: Bayesian networks II" slides from Stanford’s CS221, Autumn 2014.
Directed Graphs

A directed graph is a pair $G = (\mathcal{V}, \mathcal{E})$, where

- $\mathcal{V} = \{1, \ldots, d\}$ is a set of nodes and
- $\mathcal{E} = \{(s, t) \mid s, t \in \mathcal{V}\}$ is a set of directed edges.

Parents(5) = \{3\}
Parents(4) = \{2, 3\}
Children(3) = \{4, 5\}
Descendants(1) = \{2, 3, 4, 5\}
NonDescendants(3) = \{1, 2\}

KPM Figure 10.2(a).
A DAG is a directed graph with no directed cycles.

Every DAG has a topological ordering, in which parents have lower numbers than their children.

http://www.geeksforgeeks.org/wp-content/uploads/SCC1.png and KPM Figure 10.2(a).
Bayesian Networks

Definition

A **Bayesian network** is a

- DAG $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, d\}$, and
- a corresponding set of random variables $X = \{X_1, \ldots, X_d\}$

where

- the joint probability distribution over $X$ factorizes as

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | x_{\text{Parents}(i)}).$$

Bayesian networks are also known as

- **directed graphical models**, and
- **belief networks**.
Bayesian Networks: Example

Consider the Bayesian network depicted below:

\[
p(x_1, x_2, x_3, x_4, y) = p(y)p(x_1 | y)p(x_2 | x_1, y)p(x_3 | x_1, y)p(x_4 | x_3, y)
\]

KPM Figure 10.2(b).
Bayesian Networks: “A Common Cause”

\[
p(a, b, c) = p(c)p(a | c)p(b | c)
\]

Are \(a\) and \(b\) independent? (\(c=\text{Rain}, a=\text{Slippery}, b=\text{Wet}\)?)

\[
p(a, b) = \sum_c p(c)p(a | c)p(b | c),
\]

which in general will not be equal to \(p(a)p(b)\).

From Bishop’s *Pattern recognition and machine learning*, Figure 8.15.
Bayesian Networks: “A Common Cause”

Are $a$ and $b$ independent, conditioned on observing $c$? ($c$=Rain, $a$=Slippery, $b$=Wet?)

\[
p(a, b \mid c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a \mid c)p(b \mid c)}{p(c)}
\]

So $a \perp b \mid c$.

From Bishop’s *Pattern recognition and machine learning*, Figure 8.16.
Bayesian Networks: “An Indirect Effect”

\[ p(a, b, c) = p(a)p(c | a)p(b | c) \]

Are \( a \) and \( b \) independent? (Note: This is a Markov chain) (e.g. \( a=\text{raining}, \ c=\text{wet ground}, \ b=\text{mud on shoes} \))

\[
\begin{align*}
p(a, b) &= \sum_c p(a, b, c) \\
&= p(a) \sum_c p(c | a)p(b | c)
\end{align*}
\]

So doesn’t factorize, thus not independent, in general.

From Bishop’s *Pattern recognition and machine learning*, Figure 8.17.
Bayesian Networks: “An Indirect Effect”

\[
p(a, b, c) = p(a)p(c | a)p(b | c)
\]

Are \( a \) and \( b \) independent after observing \( c \)? (e.g. \( a \)=raining, \( c \)=wet ground, \( b \)=mud on shoes)

\[
p(a, b | c) = \frac{p(a, b, c)}{p(c)}
\]

\[
= \frac{p(a)p(c | a)p(b | c)}{p(c)}
\]

\[
= p(a | c)p(b | c)
\]

So \( a \perp b | c \).

---

From Bishop’s *Pattern recognition and machine learning*, Figure 8.18.
Bayesian Networks: “A Common Effect”

\[ p(a, b, c) = p(a)p(b)p(c \mid a, b) \]

Are \( a \) and \( b \) independent? (\( a \)=course difficulty, \( b \)=knowledge, \( c \)= grade)

\[
\begin{align*}
p(a, b) &= \sum_c p(a)p(b)p(c \mid a, b) \\
&= p(a)p(b)\sum_c p(c \mid a, b) \\
&= p(a)p(b)
\end{align*}
\]

So \( a \perp b \).

---

From Bishop’s *Pattern recognition and machine learning*, Figure 8.19.
Bayesian Networks: “A Common Effect” or “V-Structure”

Are \( a \) and \( b \) independent, given observation of \( c \)? (\( a=\)course difficulty, \( b=\)knowledge, \( c=\) grade)

\[
p(a, b | c) = p(a)p(b)p(c | a, b) / p(c)
\]

which does not factorize into \( p(a | c)p(b | c) \), in general.

From Bishop’s *Pattern recognition and machine learning*, Figure 8.20.
In general, given 3 sets of nodes \( A, B, \) and \( C \)

How can we determine whether

\[
A \perp B \mid C
\]

There is a purely graph-theoretic notion of “\textit{d-separation}” that is equivalent to conditional independence.

Suppose we have observed \( C \) and we want to do inference on \( A \).

We could ignore any evidence collected about \( B \), where \( A \perp B \mid C \).

See KPM Section 10.5.1 for details.
Markov Blanket

- Suppose we have a very large Bayesian network.
- We’re interested in a single variable $A$, which we cannot observe.
- To get maximal information about $A$, do we have to observe all other variables?
- No! We only need to observe the Markov blanket of $A$:

$$p(A \mid \text{all other nodes}) = p(A \mid \text{MarkovBlanket}(A)).$$

- In a Bayesian network, the Markov blanket of $A$ consists of
  - the parents of $A$
  - the children of $A$
  - the “co-parents” of $A$, i.e. the parents of the children of $A$

(See KPM Sec. 10.5.3 for details.)
Markov Blanket

Markov Blanket of $A$ in a Bayesian Network:

From http://en.wikipedia.org/wiki/Markov_blanket: "Diagram of a Markov blanket" by Laughsinthestocks - Licensed under CC0 via Wikimedia Commons
Bayesian Networks are great when

- you know something about the relationships between your variables, or
- you will routinely need to make inferences with incomplete data.

Challenges:

- The naive approach to inference doesn’t work beyond small scale.
- Need more sophisticated algorithm:
  - exact inference
  - approximate inference
Naive Bayes: A Generative Model for Classification

- \( \mathcal{X} = \left\{ (x_1, x_2, x_3, x_4) \in \{0, 1\}^4 \right\} \) \( \mathcal{Y} = \{0, 1\} \) be a class label.
- Consider the Bayesian network depicted below:

  ![Bayesian Network Diagram](image)

  - BN structure implies joint distribution factors as:
    
    \[
    p(x_1, x_2, x_3, x_4, y) = p(y)p(x_1 \mid y)p(x_2 \mid y)p(x_3 \mid y)p(x_4 \mid y)
    \]

  - Features \( X_1, \ldots, X_4 \) are independent given the class label \( Y \).

KPM Figure 10.2(a).
Parameters for Naive Bayes

- Generalize to $d$ features.
- Knowing the joint distribution means we need to know $p(y), p(x_1 \mid y), \ldots, p(x_d \mid y)$.

- We could parameterize as:

$$
\begin{align*}
P(Y = 1) &= \theta_y \\
P(X_i = 1 \mid Y = 1) &= \theta_{i1} \\
P(X_i = 1 \mid Y = 0) &= \theta_{i0}
\end{align*}
$$

$\implies$ 1 + 2$d$ parameters to characterize the joint distribution
Parameterized Expression for Joint

- **Parameters:**

\[ P(Y = 1) = \theta_y \quad P(X_i = 1 \mid Y = 1) = \theta_{i1} \quad P(X_i = 1 \mid Y = 0) = \theta_{i0} \]

- **Joint distribution is**

\[
p(x_1, \ldots, x_d, y) \\
= p(y) \prod_{i=1}^{n} p(x_i \mid y) \\
= (\theta_y)^y (1 - \theta_y)^{1-y} \\
\times \prod_{i=1}^{n} (\theta_{i1})^{y x_i} (1 - \theta_{i1})^{(1-y) x_i} (\theta_{i0})^{(1-y) (1-x_i)} (1 - \theta_{i0})^{(1-y) (1-x_i)}
\]
Suppose we know all conditional distributions:

\[ p(y), \, p(x_1 \mid y), \ldots, \, p(x_d \mid y) \]

We observe \( X = (X_1, \ldots, X_d) \). What’s the prediction for \( Y \)?

We have a full probability model

\[
p(y, x_1, \ldots, x_d) = p(y)p(x_1, \ldots, x_d \mid y) \quad \text{(no assumptions)}
\]

\[
= p(y) \prod_{i=1}^{d} p(x_i \mid y) \quad \text{(conditional independence)}
\]

We can use Bayes rule to compute anything we want...
Posterior Class Probability

- Let \( x = (x_1, \ldots, x_d) \), and apply Bayes rule:

\[
p(y \mid x) = \frac{p(y, x)}{p(x)} = \frac{p(y) \prod_{i=1}^{d} p(x_i \mid y)}{p(x)}
\]

- We know everything except \( p(x) \).
- We can compute it explicitly:

\[
p(x) = \sum_{y \in \{0, 1\}} p(x, y) = \sum_{y \in \{0, 1\}} p(x \mid y) p(y)
\]

- So final predicted probability distribution is

\[
p(y \mid x) = \frac{p(y) \prod_{i=1}^{d} p(x_i \mid y)}{\sum_{y \in \{0, 1\}} p(x \mid y) p(y)}
\]
Consider $p(y \mid x)$ as a distribution over $y$, for fixed $x$.

$$p(y \mid x) = \frac{p(y, x)}{p(x)}.$$ 

With $x$ fixed, $p(x)$ is a constant – let’s write it as $k$ to make it clear:

$$p(y \mid x) = k^{-1}p(y, x)$$

$\implies p(y \mid x) \propto p(y, x)$

How to recover value of $k$? $p(y \mid x)$ must be a distribution on $y$:

$$\sum_{y \in \{0, 1\}} p(y \mid x) = k^{-1} \sum_{y \in \{0, 1\}} p(y, x) = 1$$

$\implies k = \sum_{y \in \{0, 1\}} p(y, x)$

So we can always recover the normalizing constant whenever we want.

Often no need to keep track of it.
Recall the logistic regression prediction function is of the form

\[ x \mapsto p(Y = 1 \mid x) = \frac{1}{1 + \exp(-w^T x)}, \]

for some parameter vector \( w \in \mathbb{R}^d \).

**Theorem**

If \( p(y, x) \) is any Naive Bayes model with binary \( x \) and \( y \), the prediction function

\[ x \mapsto p(Y = 1 \mid x) \]

corresponds to logistic regression, for some \( w \in \mathbb{R}^d \).

**Proof**: Homework.
Naive Bayes vs Logistic Regression

- Naive Bayes is a model for the joint distribution \( p(y, x) \).
  - We can sample \((x, y)\) pairs from this distribution.
  - Models of the joint distribution are called **generative models**.
- Logistic regression is directly modeling the conditional distribution
  \[ p(y \mid x). \]
  - No model for the features \( x = (x_1, \ldots, x_d) \).
  - Conditional probability models are called **discriminative models**.
- Logistic regression is a specialist in the conditional distribution.
- Naive Bayes is doing more!
Naive Bayes vs Logistic Regression

- **Missing data** is no problem for Naive Bayes.
- Suppose we’re missing $X_1$ and $X_2$ from the input vector.
- Just predict with

$$P(y \mid x_3, \ldots, x_d) \propto p(y, x_3, \ldots, x_d)$$

$$= \sum_{x_1, x_2 \in \{0, 1\}} p(y, x)$$

- For logistic regression? No natural way to predict with missing features.
Naive Bayes vs Logistic Regression

- Logistic regression handles binary or continuous features seamlessly.
- For naive Bayes, you need a different family of conditional distributions, e.g.
  \[ p(x_i \mid y) = \mathcal{N}(x_i \mid \mu_{iy}, \sigma_{iy}^2) \]
- Wasted effort to model all features if you only care about \( p(y \mid x) \)?
- Suppose we’re missing \( X_1 \) and \( X_2 \) from the input vector.
- Just predict with
  \[
  \mathbb{P}(y \mid x_3, \ldots, x_d) \propto p(y, x_3, \ldots, x_d) \\
  = \sum_{x_1, x_2 \in \{0,1\}} p(y, x)
  \]
- No natural method for missing features with logistic regression.
Easy Estimators for Naive Bayes

- Training set $\mathcal{D} = \{(x^1, y^1), \ldots, (x^n, y^n)\}$.
- There are obvious “plug-in” estimators for the Naive Bayes model:

$$P(Y = 1) \approx \hat{\theta}_y = \frac{1}{n} \sum_{i=1}^{n} 1(y^i = 1)$$

$$P(X_i = 1 \mid Y = 1) \approx \hat{\theta}_{i1} = \frac{\sum_{j=1}^{n} 1(y^j = 1 \text{ and } x^i_j = 1)}{\sum_{j=1}^{n} 1(y^j = 1)}$$

$$P(X_i = 1 \mid Y = 0) = \hat{\theta}_{i0} = \frac{\sum_{j=1}^{n} 1(y^j = 0 \text{ and } x^i_j = 1)}{\sum_{j=1}^{n} 1(y^j = 0)}$$
Training set \( \mathcal{D} = \{(x^1, y^1), \ldots, (x^n, y^n)\} \).

More principled: find the MLE for the Naive Bayes model.

The log-likelihood objective function is

\[
J(\theta) = \sum_{i=1}^{n} \log p(y^i, x^i),
\]

where we found the likelihood for a single point \((x, y)\) is

\[
p(x, y) = (\theta_y)^y (1-\theta_y)^{1-y} \\
\times \prod_{i=1}^{n} (\theta_{i1})^{y x_i} (1-\theta_{i1})^{y(1-x_i)} \\
\times \prod_{i=1}^{n} (\theta_{i0})^{(1-y)x_i} (1-\theta_{i0})^{(1-y)(1-x_i)}
\]

**Theorem:** MLE is exactly the plug-in estimator.

**Proof:** Optional Homework.
Class Prediction

- If we want to predict a single class, we would use

\[ y^* = \arg \max_y p(y \mid x). \]

- One approach to this is to write

\[
\frac{p(Y = 1 \mid x)}{p(Y = 0 \mid x)} = \frac{p(Y = 1, x)/p(x)}{p(Y = 0, x)/p(x)} = \frac{p(Y = 1, x)}{p(Y = 0, x)}
\]

\[
= \frac{p(Y = 1) \prod_{i=1}^{d} p(x_i \mid Y = 1)}{p(Y = 0) \prod_{i=1}^{d} p(x_i \mid Y = 0)}
\]

\[
= \frac{p(Y = 1)}{p(Y = 0)} \prod_{i=1}^{d} \frac{p(x_i \mid Y = 1)}{p(x_i \mid Y = 0)}
\]

- Compare ratio to 1 to get prediction.
A Markov chain model has structure:

\[ p(x_1, x_2, x_3, \ldots) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2) \ldots \]

- Conditional distributions \( p(x_i \mid x_{i-1}) \) is called the transition model.
- When conditional distribution independent of \( i \), called time-homogeneous.
- 4-state transition model for \( X_i \in \{S_1, S_2, S_3, S_4\} \):

KPM Figure 10.3(a) and Koller and Friedman’s *Probabilistic Graphical Models* Figure 6.04.
Hidden Markov Model

- A hidden Markov model (HMM) has structure:

\[ p(z_1, z_2, z_3, \ldots) = p(z_1) \prod_{t=2}^{T} p(z_t | z_{t-1}) \prod_{t=1}^{T} p(x_t | z_t) \]

\( \text{Transition Model} \quad \text{Observation Model} \)

- At deployment time, we typically only observe \( X_1, \ldots, X_T \).
- Want to infer \( Z_1, \ldots, Z_T \).
- e.g. Want to most likely sequence \( (Z_1, \ldots, Z_T) \). (Use Viterbi algorithm.)
A maximum entropy Markov model (MEMM) has structure:

\[
p(y_1 \ldots, y_5 \mid x) = p(y_0) \prod_{t=1}^{5} p(y_t \mid y_{t-1}, x)
\]

At deployment time, we only observe \(X_1, \ldots, X_T\).

This is a conditional model. (And not a generative model).

---

Koller and Friedman’s *Probabilistic Graphical Models* Figure 20.A.1.
The MEMM transition model takes the following form:

\[
p(y_i|y_{i-1}, x) \propto \exp\left( \sum_k \lambda_k f_k(y_{i-1}, y_i) + \sum_r \mu_r g_r(y_i, x) \right)
\]

The functions \( f_k \) and \( g_r \) are **feature functions**.

Suppose \( Y \)'s represent parts-of-speech; \( X \)'s represent words.

Could have

\[
g_r(y_i, x) = \begin{cases} 
1 & \text{if } y_i = "\text{NOUN}" \text{ and } x_i = "\text{apple}" \\
0 & \text{otherwise}
\end{cases}
\]

For the "transition features", typical would be

\[
f_k(y_{i-1}, y_i) = \begin{cases} 
1 & \text{if } (y_{i-1}, y_i) = (\text{ADJ}, \text{NOUN}) \\
0 & \text{otherwise.}
\end{cases}
\]