Announcement 1: Assignment 3

Due Wednesday April 11th, 11:59 AM

Any questions?
Announcement 2: Progress Report on Project

Due Monday April 16\textsuperscript{th}, 11:59 AM

Build on the proposal:
- Update to address comments
- Discuss the progress you’ve made
- Discuss what remains to be done
- Discuss any new blocks you’ve experienced (or anticipate experiencing)

Any questions?
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
   Naïve Bayes
   Hidden Markov Models

Message Passing: Directed Graphical Model Inference
   Most likely sequence
   Total (marginal) probability
   EM in D-PGMs
Recap from last time...
Expectation Maximization (EM): E-step

0. Assume some value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

\[ p(z_i) \xrightarrow{\text{count}} \text{count}(z_i, w_i) \]

2. M-step: maximize log-likelihood, assuming these uncertain counts

\[ p^{(t)}(z) \xrightarrow{\text{estimated counts}} p^{(t+1)}(z) \]
EM Math

**E-step:** count under uncertainty

\[
\max_\theta \mathbb{E}_{z \sim p_{\theta(t)}(\cdot|w)} \left[ \log p_{\theta}(z, w) \right]
\]

**M-step:** maximize log-likelihood

\[
\mathcal{M}(\theta) = \text{marginal log-likelihood of observed data } X
\]
\[
\mathcal{C}(\theta) = \text{log-likelihood of complete data } (X,Y)
\]
\[
\mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y
\]

\[
\mathcal{M}(\theta) = \mathbb{E}_{Y \sim \theta(t)} [\mathcal{C}(\theta)|X] - \mathbb{E}_{Y \sim \theta(t)} [\mathcal{P}(\theta)|X]
\]

EM does not decrease the marginal log-likelihood
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
   Naïve Bayes
   Hidden Markov Models

Message Passing: Directed Graphical Model Inference
   Most likely sequence
   Total (marginal) probability
   EM in D-PGMs
Lagrange multipliers

Assume an original optimization problem

$$\min_x x^2$$

$$st. \quad x \geq b$$
Lagrange multipliers

Assume an original optimization problem

$$\min_x x^2$$

$$st. \quad x \geq b$$

We convert it to a new optimization problem:

$$\min_x \max_\alpha L(x, \alpha)$$

$$st. \quad \alpha \geq 0$$

$$L(x, \alpha) = x^2 - \alpha(x - b)$$
Lagrange multipliers: an equivalent problem?

\[
\begin{align*}
\min_x x^2 \\
st. & \quad x \geq b
\end{align*}
\]

\[
\begin{align*}
\min_x \max_\alpha L(x, \alpha) \\
st. & \quad \alpha \geq 0 \\
L(x, \alpha) &= x^2 - \alpha(x - b)
\end{align*}
\]

\[
x < b \quad \Rightarrow \quad (x - b) < 0 \quad \Rightarrow \quad \max_\alpha -\alpha(x - b) = \inf
\]
Lagrange multipliers: an equivalent problem?

\[
\begin{align*}
\min_{x} & \quad x^2 \\
\text{st.} & \quad x \geq b
\end{align*}
\quad \longrightarrow \quad
\begin{align*}
\min_{x} \max_{\alpha} & \quad L(x, \alpha) \\
\text{st.} & \quad \alpha \geq 0 \\
L(x, \alpha) &= x^2 - \alpha(x - b)
\end{align*}
\]

\[
\begin{align*}
x < b & \quad \Rightarrow \quad (x - b) < 0 \quad \Rightarrow \quad \max_{\alpha} -\alpha(x - b) = \inf \\
x > b & \quad \Rightarrow \quad (x - b) > 0 \quad \Rightarrow \quad \max_{\alpha} -\alpha(x - b) = 0, \alpha = 0 \quad \Rightarrow \quad L(x, \alpha) = x^2
\end{align*}
\]
Lagrange multipliers: an equivalent problem?

\[
\begin{align*}
\min_x & \quad x^2 \\
\text{st.} & \quad x \geq b
\end{align*}
\]

\[
\begin{align*}
\min \max_{x, \alpha} & \quad L(x, \alpha) \\
\text{st.} & \quad \alpha \geq 0 \\
L(x, \alpha) = & \quad x^2 - \alpha(x - b)
\end{align*}
\]

\[
\begin{align*}
x < b & \quad \Rightarrow \quad (x - b) < 0 \quad \Rightarrow \quad \max_{\alpha} -\alpha(x - b) = \inf \\
x > b & \quad \Rightarrow \quad (x - b) > 0 \quad \Rightarrow \quad \max_{\alpha} -\alpha(x - b) = 0, \alpha = 0 \quad \Rightarrow \quad L(x, \alpha) = x^2 \\
x = b & \quad \Rightarrow \quad (x - b) = 0 \quad \Rightarrow \quad \alpha = \text{any thing} \quad \Rightarrow \quad L(x, \alpha) = x^2
\end{align*}
\]
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
  Naïve Bayes
  Hidden Markov Models

Message Passing: Directed Graphical Model Inference
  Most likely sequence
  Total (marginal) probability
  EM in D-PGMs
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

N different (independent) rolls

\[ w_1 = 1 \]

\[ w_2 = 5 \]

\[ w_3 = 4 \]

\[ \cdots \]

Generative Story

for roll \( i = 1 \) to \( N \):

\[ w_i \sim \text{Cat}(\theta) \]

a probability distribution over 6 sides of the die

\[ \sum_{k=1}^{6} \theta_k = 1 \]

\[ 0 \leq \theta_k \leq 1, \forall k \]
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

N different (independent) rolls

\[ w_1 = 1 \]
\[ w_2 = 5 \]
\[ w_3 = 4 \]
\[ \cdots \]

Generative Story

for roll \( i = 1 \) to \( N \):
\[ w_i \sim \text{Cat}(\theta) \]

Maximize Log-likelihood

\[ \mathcal{L}(\theta) = \sum_i \log p_\theta(w_i) = \sum_i \log \theta_{w_i} \]
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Generative Story

for roll \( i = 1 \) to \( N \):

\( w_i \sim \text{Cat}(\theta) \)

Maximize Log-likelihood

\[ \mathcal{L}(\theta) = \sum_i \log \theta_{w_i} \]

Q: What’s an easy way to maximize this, as written \textit{exactly} (even without calculus)?
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Generative Story
for roll \( i = 1 \) to \( N \):
\( w_i \sim \text{Cat}(\theta) \)

Maximize Log-likelihood
\[ \mathcal{L}(\theta) = \sum_i \log \theta_{w_i} \]

Q: What’s an easy way to maximize this, as written exactly (even without calculus)?

A: Just keep increasing \( \theta_k \) (we know \( \theta \) must be a distribution, but it’s not specified)
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

N different (independent) rolls

Maximize Log-likelihood (with distribution constraints)

\[ \mathcal{L}(\theta) = \sum_i \log \theta_{w_i} \quad \text{s.t.} \quad \sum_{k=1}^{6} \theta_k = 1 \]

(we can include the inequality constraints \(0 \leq \theta_k\), but it complicates the problem and, right now, is not needed)

solve using Lagrange multipliers
Probabilistic Estimation of Rolling a Die

\[
p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i)
\]

Maximize Log-likelihood (with distribution constraints)

\[
F(\theta) = \sum_i \log \theta_{w_i} - \lambda \left( \sum_{k=1}^{6} \theta_k - 1 \right)
\]

\[
\frac{\partial F(\theta)}{\partial \theta_k} = \sum_{i:w_i=k} \frac{1}{\theta_{w_i}} - \lambda \quad \frac{\partial F(\theta)}{\partial \lambda} = -\sum_{k=1}^{6} \theta_k + 1
\]

(we can include the inequality constraints \(0 \leq \theta_k\), but it complicates the problem and, right now, is not needed)
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

N different (independent) rolls

Maximize Log-likelihood (with distribution constraints)

\[ F(\theta) = \sum_i \log \theta_{w_i} - \lambda \left( \sum_{k=1}^{6} \theta_k - 1 \right) \]

\[ \theta_k = \frac{\sum_{i:w_i=k} 1}{\lambda} \]

optimal \( \lambda \) when \( \sum_{k=1}^{6} \theta_k = 1 \)

(we can include the inequality constraints \( 0 \leq \theta_k \), but it complicates the problem and, right now, is not needed)
Probabilistic Estimation of Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Maximize Log-likelihood (with distribution constraints)

\[ \mathcal{F}(\theta) = \sum_i \log \theta_{w_i} - \lambda \left( \sum_{k=1}^{6} \theta_k - 1 \right) \]

\[ \theta_k = \frac{\sum_{i:w_i=k} 1}{\sum_k \sum_{i:w_i=k} 1} = \frac{N_k}{N} \]

optimal \( \lambda \) when \( \sum_{k=1}^{6} \theta_k = 1 \)
Example: Conditionally Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i) \]

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N) = \prod_{i} p(w_i|z_i)p(z_i) \]

\[ p\text{(heads)} = \lambda \]
\[ p\text{(tails)} = 1 - \lambda \]

\[ p\text{(heads)} = \gamma \]
\[ p\text{(tails)} = 1 - \gamma \]

\[ p\text{(heads)} = \psi \]
\[ p\text{(tails)} = 1 - \psi \]
Example: Conditionally Rolling a Die

\[
p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i)
\]

\[
p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N)
\]

\[
= \prod_i p(w_i|z_i) p(z_i)
\]

\[
p(\text{heads}) = \lambda
\]

\[
p(\text{tails}) = 1 - \lambda
\]

\[
p(\text{heads}) = \gamma
\]

\[
p(\text{tails}) = 1 - \gamma
\]

**Generative Story**

\[
\lambda = \text{distribution over penny}
\]

\[
\gamma = \text{distribution for dollar coin}
\]

\[
\psi = \text{distribution over dime}
\]

for item \(i = 1\) to \(N\):

\[
z_i \sim \text{Bernoulli}(\lambda)
\]

if \(z_i = H\): \(w_i \sim \text{Bernoulli}(\gamma)
\]

else: \(w_i \sim \text{Bernoulli}(\psi)\)
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
  Naïve Bayes
  Hidden Markov Models

Message Passing: Directed Graphical Model Inference
  Most likely sequence
  Total (marginal) probability
  EM in D-PGMs
Classify with Bayes Rule

\[ \text{argmax}_Y p(Y | X) \]

\[ \text{argmax}_Y \log p(X | Y) + \log p(Y) \]

likelihood \hspace{1cm} prior
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

Adapted from Jurafsky & Martin (draft)
### Bag of Words Representation

The table below shows the bag of words representation of a text sample. Each row represents a word and its frequency count.

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen</td>
<td>2</td>
</tr>
<tr>
<td>sweet</td>
<td>1</td>
</tr>
<tr>
<td>whimsical</td>
<td>1</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

This representation is used in natural language processing, particularly in text classification tasks where a classifier is used to predict the class of the text based on the bag of words.

Adapted from Jurafsky & Martin (draft)
Naïve Bayes: A Generative Story

Generative Story

$$\phi = \text{distribution over } K \text{ labels}$$

for label $$k = 1 \text{ to } K$$:

$$\theta_k = \text{generate parameters}$$
Naïve Bayes: A Generative Story

Generative Story

\( \phi = \) distribution over \( K \) labels

for label \( k = 1 \) to \( K \):

\( \theta_k = \) generate parameters

for item \( i = 1 \) to \( N \):

\( y_i \sim \text{Cat}(\phi) \)
Naïve Bayes: A Generative Story

**Generative Story**

$\phi = \text{distribution over } K \text{ labels}$

For label $k = 1$ to $K$:

$\theta_k = \text{generate parameters}$

For item $i = 1$ to $N$:

$y_i \sim \text{Cat}(\phi)$

For each feature $j$:

$x_{ij} \sim F_j(\theta_{y_i})$
Naïve Bayes: A Generative Story

Generative Story

\( \phi = \text{distribution over } K \text{ labels} \)

for label \( k = 1 \) to \( K \):

\( \theta_k = \text{generate parameters} \)

for item \( i = 1 \) to \( N \):

\( y_i \sim \text{Cat}(\phi) \)

for each feature \( j \)

\( x_{ij} \sim F_j(\theta_{y_i}) \)
Naïve Bayes: A Generative Story

Generative Story

\[ \phi = \text{distribution over } K \text{ labels} \]

for label \( k = 1 \) to \( K \):

\[ \theta_k = \text{generate parameters} \]

for item \( i = 1 \) to \( N \):

\[ y_i \sim \text{Cat}(\phi) \]

for each feature \( j \)

\[ x_{ij} \sim F_j(\theta_{y_i}) \]

Maximize Log-likelihood

\[ \mathcal{L}(\theta) = \sum_i \sum_j \log F_{y_i}(x_{ij}; \theta_{y_i}) + \sum_i \log \phi_{y_i} \quad \text{s.t.} \]

\[ \sum_k \phi_k = 1 \]

\[ \theta_k \text{ is valid for } F_k \]
Multinomial Naïve Bayes: A Generative Story

Generative Story

\(\phi = \) distribution over \(K\) labels

for label \(k = 1\) to \(K\):

\(\theta_k = \) distribution over \(J\) feature values

for item \(i = 1\) to \(N\):

\(y_i \sim \text{Cat}(\phi)\)

for each feature \(j\)

\(x_{ij} \sim \text{Cat}(\theta_{y_i,j})\)

Maximize Log-likelihood

\[
\mathcal{L}(\theta) = \sum_i \sum_j \log \theta_{y_i,x_{i,j}} + \sum_i \log \phi_{y_i} \quad \text{s.t.}
\]

\[
\sum_k \phi_k = 1 \quad \sum_j \theta_{kj} = 1 \quad \forall k
\]
Multinomial Naïve Bayes: A Generative Story

**Generative Story**

\( \phi = \text{distribution over } K \text{ labels} \)

for label \( k = 1 \) to \( K \):

\( \theta_k = \text{distribution over } J \text{ feature values} \)

for item \( i = 1 \) to \( N \):

\( y_i \sim \text{Cat}(\phi) \)

for each feature \( j \)

\( x_{ij} \sim \text{Cat}(\theta_{y_i,j}) \)

Maximize Log-likelihood via Lagrange Multipliers

\[ \mathcal{L}(\theta) = \sum_i \sum_j \log \theta_{y_i,x_i,j} + \sum_i \log \phi_{y_i} - \mu \left( \sum_k \phi_k - 1 \right) - \sum_k \lambda_k \left( \sum_j \theta_{k,j} - 1 \right) \]
Multinomial Naïve Bayes: Learning

**Calculate class priors**
For each $k$:

\[ \text{items}_k = \text{all items with class} = k \]

\[ p(k) = \frac{|\text{items}_k|}{\# \text{ items}} \]

**Calculate feature generation terms**
For each $k$:

\[ \text{obs}_k = \text{single object containing all items labeled as} \ k \]

For each feature $j$

\[ n_{kj} = \# \text{ of occurrences of} \ j \text{ in} \ \text{obs}_k \]

\[ p(j|k) = \frac{n_{kj}}{\sum_{j'} n_{kj'}} \]
Brill and Banko (2001)
With enough data, the classifier may not matter

Adapted from Jurafsky & Martin (draft)
Summary: Naïve Bayes is Not So Naïve, but not without issue

**Pro**

- Very Fast, low storage requirements
- Robust to Irrelevant Features
- Very good in domains with many equally important features
- Optimal if the independence assumptions hold
- Dependable baseline for text classification (but often not the best)

**Con**

- Model the posterior in one go? (e.g., use conditional maxent)
- Are the features really uncorrelated?
- Are plain counts always appropriate?
- Are there “better” ways of handling missing/noisy data? (automated, more principled)

Adapted from Jurafsky & Martin (draft)
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
  Naïve Bayes
  **Hidden Markov Models**

Message Passing: Directed Graphical Model Inference
  Most likely sequence
  Total (marginal) probability
  EM in D-PGMs
Hidden Markov Models

(i): Adjective $\rightarrow$ Noun $\rightarrow$ Verb $\rightarrow$ Prep $\rightarrow$ Noun $\rightarrow$ Noun

(ii): Noun $\rightarrow$ Verb $\rightarrow$ Noun $\rightarrow$ Prep $\rightarrow$ Noun $\rightarrow$ Noun

$p(\text{British Left Waffles on Falkland Islands})$

Class-based model

$p(w_i|z_i)$

Bigram model of the classes

$p(z_i|z_{i-1})$

Model all class sequences

$\sum_{z_1,\ldots,z_N} p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N)$
Hidden Markov Model

\[
p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) \\
= \prod_i p(w_i | z_i) p(z_i | z_{i-1})
\]

Goal: maximize (log-)likelihood

In practice: we don’t actually observe these \( z \) values; we just see the words \( w \)
Hidden Markov Model

\[
p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) = \prod_i p(w_i | z_i) p(z_i | z_{i-1})
\]

Goal: maximize (log-likelihood)

In practice: we don’t actually observe these \( z \) values; we just see the words \( w \)

if we \textit{did} observe \( z \), estimating the probability parameters would be easy...
but we don’t! :(  

if we \textit{knew} the probability parameters then we could estimate \( z \) and evaluate likelihood... but we don’t! :(  

Hidden Markov Model Terminology

\[
p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N)
= \prod_i p(w_i | z_i) p(z_i | z_{i-1})
\]

Each \( z_i \) can take the value of one of \( K \) latent states
Hidden Markov Model Terminology

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) = \prod_i p(w_i | z_i) p(z_i | z_{i-1}) \]

Each \( z_i \) can take the value of one of \( K \) latent states
Hidden Markov Model Terminology

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) \]

\[ = \prod_i p(w_i | z_i) p(z_i | z_{i-1}) \]

Each \(z_i\) can take the value of one of \(K\) latent states
Hidden Markov Model Terminology

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) = \prod_i p(w_i | z_i) p(z_i | z_{i-1}) \]

Each \( z_i \) can take the value of one of \( K \) latent states

Transition and emission distributions do not change
Hidden Markov Model Terminology

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) \]

\[ = \prod_{i} p(w_i | z_i) p(z_i | z_{i-1}) \]

- emission probabilities/parameters
- transition probabilities/parameters

Each \( z_i \) can take the value of one of \( K \) latent states

Transition and emission distributions do not change

Q: How many different probability values are there with \( K \) states and \( V \) vocab items?
Hidden Markov Model Terminology

\[
p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N)
= \prod_{i} p(w_i | z_i) \ p(z_i | z_{i-1})
\]

Each \(z_i\) can take the value of one of \(K\) latent states

Transition and emission distributions do not change

**Q:** How many different probability values are there with \(K\) states and \(V\) vocab items?

**A:** \(VK\) emission values and \(K^2\) transition values
Hidden Markov Model Representation

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) \]

represent the probabilities and independence assumptions in a graph
Hidden Markov Model Representation

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1 | z_0)p(w_1 | z_1) \cdots p(z_N | z_{N-1})p(w_N | z_N) \]

\[ = \prod_i p(w_i | z_i)p(z_i | z_{i-1}) \]

Initial starting distribution ("__SEQ_SYM__")

\[ p(z_1 | z_0) \]

Transition and emission distributions do not change

Each \( z_i \) can take the value of one of \( K \) latent states

Transition and emission distributions do not change
Example: 2-state Hidden Markov Model as a Lattice
Example: 2-state Hidden Markov Model as a Lattice

\begin{align*}
\text{z}_1 &= V \\
\text{z}_2 &= V \\
\text{z}_3 &= V \\
\text{z}_4 &= V
\end{align*}

\begin{align*}
p(w_1 | V) &\quad p(w_2 | V) &\quad p(w_3 | V) &\quad p(w_4 | V) \\
p(w_1 | N) &\quad p(w_2 | N) &\quad p(w_3 | N) &\quad p(w_4 | N)
\end{align*}

\begin{array}{cccc}
\text{w}_1 & \text{w}_2 & \text{w}_3 & \text{w}_4 \\
.7 & .2 & .05 & .05 \\
.15 & .8 & .05 & \\
.6 & .35 & .05 &
\end{array}
Example: 2-state Hidden Markov Model as a Lattice
Example: 2-state Hidden Markov Model as a Lattice

\[
\begin{align*}
& z_1 = N, \quad z_2 = V, \quad z_3 = V, \quad z_4 = N, \\
& p(w_1 | N) \quad p(w_2 | N) \quad p(w_3 | N) \quad p(w_4 | N), \\
& \text{start} \quad \text{end} \\
& p(V | \text{start}) \quad p(V | V) \quad p(V | V) \quad p(V | V), \\
& p(N | V) \quad p(N | V) \quad p(N | V) \quad p(N | V), \\
& p(w_1 | V) \quad p(w_2 | V) \quad p(w_3 | V) \quad p(w_4 | V), \\
& w_1 \quad w_2 \quad w_3 \quad w_4
\end{align*}
\]
A Latent Sequence is a Path through the Graph

Q: What's the probability of \((N, w_1), (V, w_2), (V, w_3), (N, w_4)\)?

A: \((.7 *.7) * (.8*.6) * (.35*.1) * (.6*.05) = 0.0002822\)
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
   Naïve Bayes
   Hidden Markov Models

Message Passing: Directed Graphical Model Inference
   Most likely sequence
   Total (marginal) probability
   EM in D-PGMs
Message Passing: Count the Soldiers

If you are the front soldier in the line, say the number ‘one’ to the soldier behind you.

If you are the rearmost soldier in the line, say the number ‘one’ to the soldier in front of you.

If a soldier ahead of or behind you says a number to you, add one to it, and say the new number to the soldier on the other side.
Message Passing: Count the Soldiers

If you are the front soldier in the line, say the number ‘one’ to the soldier behind you.

If you are the rearmost soldier in the line, say the number ‘one’ to the soldier in front of you.

If a soldier ahead of or behind you says a number to you, add one to it, and say the new number to the soldier on the other side.
What’s the Maximum *Weighted Path*?
What’s the Maximum *Weighted Path*?
What’s the Maximum Weighted Path?
What’s the Maximum *Weighted Path*?
What’s the Maximum Weighted Path?
What’s the Maximum *Weighted Path*?
What’s the Maximum Value?

Consider “any shared path ending with B (AB, BB, or CB) → B” maximize across the previous hidden state values $v(i, B) = \max_{s'} v(i - 1, s') \ast p(B \mid s') \ast p(\text{obs at } i \mid B)$
What’s the Maximum Value?

Consider “any shared path ending with B (AB, BB, or CB) → B” maximize across the previous hidden state values

\[ v(i, B) = \max_{s'} v(i - 1, s') \times p(B \mid s') \times p(\text{obs at } i \mid B) \]
What’s the Maximum Value?

computing $v$ at time $i-1$ will correctly incorporate (maximize over) paths through time $i-2$: we correctly obey the Markov property

consider “any shared path ending with $B$ (AB, BB, or CB) → $B$”
maximize across the previous hidden state values

$$v(i, B) = \max_{s'} v(i - 1, s') * p(B | s') * p(\text{obs at } i | B)$$

Viterbi algorithm
Viterbi Algorithm

\[ v = \text{double}[N+2][K^*] \]
\[ b = \text{int}[N+2][K^*] \]

\[ v[0][*] = 0 \]

backpointers/
book-keeping

\begin{verbatim}
for(i = 1; i \leq N+1; ++i) {
    for(state = 0; state < K^*; ++state) {
        p_{obs} = p_{emission}(obs_i | state)
        for(old = 0; old < K^*; ++old) {
            p_{move} = p_{transition}(state | old)
            if(\(v[i-1][old] \times p_{obs} \times p_{move} > v[i][state]\)) {
                v[i][state] = v[i-1][old] \times p_{obs} \times p_{move}
                b[i][state] = old
            }
        }
    }
}
\end{verbatim}
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
   Naïve Bayes
   Hidden Markov Models

Message Passing: Directed Graphical Model Inference
   Most likely sequence
   **Total (marginal) probability**
   EM in D-PGMs
\[ \alpha(i, B) \] is the total probability of all paths to that state B from the beginning.
Forward Probability

\[
\alpha(i, B) = \sum_{s'} \alpha(i - 1, s') \times p(B \mid s') \times p(\text{obs at } i \mid B)
\]

\(\alpha(i, B)\) is the total probability of all paths to that state B from the beginning

computing \(\alpha\) at time \(i-1\) will correctly incorporate paths through time \(i-2\): we correctly obey the Markov property
Forward Probability

$$\alpha(i, s) = \sum_{s'} \alpha(i - 1, s') \times p(s \mid s') \times p(\text{obs at } i \mid s)$$

$\alpha(i, s)$ is the total probability of all paths:

1. that start from the beginning
2. that end (currently) in $s$ at step $i$
3. that emit the observation $\text{obs at } i$
Forward Probability

\[ \alpha(i, s) = \sum_{s'} \alpha(i-1, s') \times p(s \mid s') \times p(\text{obs at } i \mid s) \]

- what are the *immediate* ways to get into state \( s \)?
- what’s the total probability *up until* now?
- how likely is it to get into state \( s \) this way?

\( \alpha(i, s) \) is the total probability of all paths:

1. that start from the beginning
2. that end (currently) in \( s \) at step \( i \)
3. that emit the observation \( \text{obs at } i \)
Forward Probability

$$\alpha(i, s) = \sum_{s'} \alpha(i - 1, s') \times p(s \mid s') \times p(\text{obs at } i \mid s)$$

what are the immediate ways to get into state $s$?
what’s the total probability up until now?
how likely is it to get into state $s$ this way?

$\alpha(i, s)$ is the total probability of all paths:

1. that start from the beginning
2. that end (currently) in $s$ at step $i$
3. that emit the observation obs at $i$

Q: What do we return? (How do we return the likelihood of the sequence?)
A: $\alpha[N+1][\text{end}]$
Outline

Recap of EM

Math: Lagrange Multipliers for constrained optimization

Probabilistic Modeling Example: Die Rolling

Directed Graphical Models
  Naïve Bayes
  Hidden Markov Models

Message Passing: Directed Graphical Model Inference
  Most likely sequence
  Total (marginal) probability

EM in D-PGMs
Forward & Backward Message Passing

\( \alpha(i, s) \) is the total probability of all paths:
1. that start from the beginning
2. that end (currently) in \( s \) at step \( i \)
3. that emit the observation \( obs \) at \( i \)

\( \beta(i, s) \) is the total probability of all paths:
1. that start at step \( i \) at state \( s \)
2. that terminate at the end
3. (that emit the observation \( obs \) at \( i+1 \))
Forward & Backward Message Passing

\[ \alpha(i, B) \] \hspace{1cm} \beta(i, B) \]

\[ z_{i-1} = A \]
\[ z_{i-1} = B \]
\[ z_{i-1} = C \]

\[ z_i = A \]
\[ z_i = B \]
\[ z_i = C \]

\[ z_{i+1} = A \]
\[ z_{i+1} = B \]
\[ z_{i+1} = C \]

\( \alpha(i, s) \) is the total probability of all paths:
1. that start from the beginning
2. that end (currently) in \( s \) at step \( i \)
3. that emit the observation \( \text{obs} \) at \( i \)

\( \beta(i, s) \) is the total probability of all paths:
1. that start at step \( i \) at state \( s \)
2. that terminate at the end
3. (that emit the observation \( \text{obs} \) at \( i+1 \))
\[ \alpha(i, B) \times \beta(i, B) = \text{total probability of paths through state } B \text{ at step } i \]

\[ \alpha(i, B) \] is the total probability of all paths:
1. that start from the beginning
2. that end (currently) in \( s \) at step \( i \)
3. that emit the observation \( \text{obs} \) at \( i \)

\[ \beta(i, B) \] is the total probability of all paths:
1. that start at step \( i \) at state \( s \)
2. that terminate at the end
3. (that emit the observation \( \text{obs at } i+1 \))
Forward & Backward Message Passing

$\alpha(i, B)$ is the total probability of all paths:
1. that start from the beginning
2. that end (currently) in $s$ at step $i$
3. that emit the observation $obs$ at $i$

$\beta(i+1, s)$ is the total probability of all paths:
1. that start at step $i$ at state $s$
2. that terminate at the end
3. (that emit the observation $obs$ at $i+1$)
Forward & Backward Message Passing

\[ \alpha(i, B) \] is the total probability of all paths:
1. that start from the beginning
2. that end (currently) in \( s \) at step \( i \)
3. that emit the observation \( \text{obs at } i \)

\[ \beta(i+1, s') \] is the total probability of all paths:
1. that start at step \( i \) at state \( s \)
2. that terminate at the end
3. (that emit the observation \( \text{obs at } i+1 \))

\[ \alpha(i, B) \cdot p(s' \mid B) \cdot p(\text{obs at } i+1 \mid s') \cdot \beta(i+1, s') = \text{total probability of paths through the } B \rightarrow s' \text{ arc (at time } i) \]
With Both Forward and Backward Values

\[ \alpha(i, s) \times \beta(i, s) = \text{total probability of paths through state } s \text{ at step } i \]

\[ p(z_i = s \mid w_1, \ldots, w_N) = \frac{\alpha(i, s) \times \beta(i, s)}{\alpha(N + 1, \text{END})} \]

\[ \alpha(i, s) \times p(s' \mid B) \times p(\text{obs at } i+1 \mid s') \times \beta(i + 1, s') = \text{total probability of paths through the } s \rightarrow s' \text{ arc (at time } i) \]

\[ p(z_i = s, z_{i+1} = s' \mid w_1, \ldots, w_N) = \frac{\alpha(i, s) \times p(s' \mid s) \times p(\text{obs}_{i+1} \mid s') \times \beta(i + 1, s')}{\alpha(N + 1, \text{END})} \]
Expectation Maximization (EM)

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

\[ p(z_i) \quad \rightarrow \quad \text{count}(z_i, w_i) \]

2. M-step: maximize log-likelihood, assuming these uncertain counts

\[ p^{(t)}(z) \quad \leftrightarrow \quad \text{estimated counts} \quad \rightarrow \quad p^{(t+1)}(z) \]
Expectation Maximization (EM)

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

\[ p(z_i) \rightarrow \text{count}(z_i, w_i) \]

2. M-step: maximize log-likelihood, assuming these uncertain counts

\[ p^{(t)}(z) \leftrightarrow \text{estimated counts} \rightarrow p^{(t+1)}(z) \]
Expectation Maximization (EM)

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

\[
p^*(z_i = s | w_1, \cdots, w_N) = \frac{\alpha(i, s) \beta(i, s)}{\alpha(N + 1, \text{END})}
\]

2. M-step: maximize log-likelihood, assuming these uncertain counts

\[
p^*(z_i = s, z_{i+1} = s' | w_1, \cdots, w_N) = \frac{\alpha(i, s) p(s' | s) p(\text{obs}_{i+1} | s') \beta(i + 1, s')}{\alpha(N + 1, \text{END})}
\]
M-Step

“maximize log-likelihood, assuming these uncertain counts”

\[
p_{\text{new}}^{s'}(s) = \frac{c(s \rightarrow s')}{\sum_x c(s \rightarrow x)}
\]

if we observed the hidden transitions...
M-Step

“maximize log-likelihood, assuming these uncertain counts”

\[ p_{\text{new}}(s'|s) = \frac{\mathbb{E}_{s \rightarrow s'}[c(s \rightarrow s')]}{\sum_x \mathbb{E}_{s \rightarrow x}[c(s \rightarrow x)]} \]

we don’t the hidden transitions, but we can approximately count
M-Step

“maximize log-likelihood, assuming these uncertain counts”

\[ p^{\text{new}}(s'|s) = \frac{\mathbb{E}_{s \rightarrow s'}[c(s \rightarrow s')] \sum_x \mathbb{E}_{s \rightarrow x}[c(s \rightarrow x)]}{\sum_x \mathbb{E}_{s \rightarrow x}[c(s \rightarrow x)]} \]

we don’t the hidden transitions, but we can approximately count

we compute these in the E-step, with our \( \alpha \) and \( \beta \) values
\( \alpha = \text{computeForwards()} \)
\( \beta = \text{computeBackwards()} \)

\( L = \alpha[N+1][\text{END}] \)

for \( i = N; i \geq 0; --i \) {
    for (next = 0; next < K*; ++next) {
        \( c_{\text{obs}}(\text{obs}_{i+1} | \text{next}) += \alpha[i+1][\text{next}] \times \beta[i+1][\text{next}]/L \)
    }
    for (state = 0; state < K*; ++state) {
        \( u = p_{\text{obs}}(\text{obs}_{i+1} | \text{next}) \times p_{\text{trans}}(\text{next} | \text{state}) \)
        \( c_{\text{trans}}(\text{next} | \text{state}) += \alpha[i][\text{state}] \times u \times \beta[i+1][\text{next}]/L \)
    }
}