Due this Wednesday 10/11 at 11:59 AM (< 2 days)

Use project id paper1 for the submit utility
Course Announcement: Assignment 2

Due next Wednesday, 10/18 (~9 days)

Any questions?
Course Announcement: Midterm

Monday, 10/30 (3 weeks)

Format:

In-class (75 minutes)

You may bring any notes you created yourself; they must be turned in with the exam (photocopies are OK)

Some practice questions will be out next Wednesday (10/18)
Recap from last time...
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

   \[ p^{(t)}(z) \rightarrow \text{estimated counts} \rightarrow p^{(t+1)}(z) \]
Counting Requires Marginalizing

E-step: count under uncertainty, assuming these parameters

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

E-step: count under uncertainty, assuming these parameters break into 4 disjoint pieces

\[ p(w) = p(z_1, w) + p(z_2, w) + p(z_3, w) + p(z_4, w) = \sum_{z=1}^{4} p(z_i, w) \]
Imagine three coins

Flip 1st coin (penny)

If heads: flip 2nd coin (dollar coin)

If tails: flip 3rd coin (dime)

observed: $a, b, e$, etc.
We run the code, vs.

unobserved: vowel or consonant? part of speech?

The run failed
EM Example 2: Machine Translation Alignment

Want: $P(f|e)$
But don’t know how to train this directly...

Solution: Use $P(a, f|e)$, where $a$ is an alignment

Remember:

$$P(f|e) = \sum_a P(a, f|e)$$

marginalizing across all possible alignments
IBM Model 1 (1993)

$f$: vector of French words

(visualization of alignment)

e: vector of English words

$a$: vector of alignment indices

Le chat est sur la chaise verte

The cat is on the green chair

$$P(a, f | e) = \prod_{j=1}^{m} t(f_j | e_{a_j}) = t(f_1 | e_{a_1}) \cdots t(f_m | e_{a_m})$$

$t(f_j | e_i)$: translation probability of the word $f_j$ given the word $e_i$
Learning the Alignments through EM

0. Assume some value for $t(f_j | e_i)$ and compute other parameter values

Two step, iterative algorithm

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

   \[
   \begin{align*}
   t(f_j | e_i) & \quad P(a, f | e) \\
   P(a | e, f) & \quad \text{estimated counts}
   \end{align*}
   \]

2. M-step: maximize log-likelihood (update parameters), using uncertain counts
Follow up: IBM Model 1 Parameters

For IBM model 1, we can compute all parameters given translation parameters:

\[ t(f_j | e_i) \]

How many of these are there?

\[ |French\ vocabulary| \times |English\ vocabulary| \]

From Rebecca: See Sec. 31 of the Knight tutorial for more about space considerations
Alignment: Output and Complexities

Component of machine translation systems
Produce a translation lexicon automatically
Cross-lingual projection/extraction of information
Supervision for training other models (for example, neural MT systems)

http://www.cis.upenn.edu/~ccb/figures/research-statement/pivoting.jpg
Any Questions on What We’ve Seen of EM So Far?
Hidden Markov Models
Agenda

HMM Motivation (Part of Speech) and Brief Definition

What is Part of Speech?

HMM Detailed Definition

HMM Tasks
Hidden Markov Models

Class-based Model
Use different distributions to explain groupings of observations

Sequence Model

Bigram model of the *classes*, not the observations

Implicitly model all possible class sequences

Algorithms for finding best sequence, and for the marginal likelihood
Hidden Markov Models: Part of Speech

(i): Adjective Noun Verb Prep Noun Noun

(ii): Noun Verb Noun Prep Noun Noun

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

Class-based model
Bigram model of the classes
Model all class sequences
Hidden Markov Models: Part of Speech

(i): Adjective → Noun → Verb → Prep → Noun → Noun
(ii): Noun → Verb → Noun → Prep → Noun → Noun

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

Class-based model
Bigram model of the classes
Model all class sequences
Hidden Markov Models: Part of Speech

(i): Adjective → Noun → Verb → Prep → Noun → Noun

(ii): Noun → Verb → Noun → Prep → Noun → Noun

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

Class-based model

Bigram model of the classes

Model all class sequences
Hidden Markov Models: Part of Speech

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

(i): Adjective → Noun → Verb → Prep → Noun → Noun

(ii): Noun → Verb → Noun → Prep → Noun → Noun

Class-based model
Bigram model of the classes
Model all class sequences
1. Explain this sentence as a sequence of (likely?) latent (unseen) tags (labels)

2. Produce a tag sequence for this sentence

\[ p(\text{British Left Waffles on Falkland Islands}) \]
Agenda

HMM Motivation (Part of Speech) and Brief Definition

What is Part of Speech?

HMM Detailed Definition

HMM Tasks
Brief Aside: Parts of Speech

Classes of words that behave like one another in similar syntactic contexts
Parts of Speech

Classes of words that behave like one another in similar syntactic contexts

Pronunciation (stress) can differ: object (noun: OB-ject) vs. object (verb: ob-JECT)

It can help improve the inputs to other systems (text-to-speech, syntactic parsing)
It was *I* who allowed the alliance to know the location of the shield generator.

You mean "it was me." You're following an archaic grammar rule.

It was *me* who allowed the—

No, my master, an archaic tone is appropriate here. The sentence sounds—

It was *I* who allowed—

Come on, the accusative case is fine. Nominative pronouns are—

It me

I allowed it

My master, please never say that again.
Parts of Speech

Nouns
- Baltimore
- UMBC
- bread
- milk
- cats
- cat

Verbs
- speak
- give
- run

Adjectives
- would-be
- wettest
- large
- happy
- red
- fake
Parts of Speech

Nouns
- bread
- milk
- cat
- cats
- UMBC
- Baltimore

Verbs
- speak
- give
- run

Adjectives
- red
- happy
- large
- wettest
- would-be
- fake

Determiners
- a
- the
- every
- what

Conjunctions
- and
- or
- if
- because

Prepositions
- in
- under
- top
Parts of Speech

Nouns
- bread
- milk
- cat
- cats
- UMBC
- Baltimore

Verbs
- speak
- give
- run
- can
- may
- do

Adjectives
- red
- happy
- large
- wettest
- would-be
- fake

Determiners
- a
- every
- the
- what
- every

Conjunctions
- and
- or
- if
- because

Prepositions
- in
- top
- under

“I can eat.”

Adapted from Luke Zettlemoyer
Parts of Speech

Nouns
- bread
- milk
- cats
- cat
- Baltimore
- UMBC

Verbs
- speak
- give
- run
- can
- may
- do

Adjectives
- fake
- would-be
- red
- happy
- large
- wettest

Determiners
- a
- every
- what
- the

Conjunctions
- and
- or
- if
- because

Prepositions
- in
- top
- under

Adapted from Luke Zettlemoyer
Parts of Speech

Adapted from Luke Zettlemoyer

Nouns
- bread
- milk
- cats
- UMBC
- Baltimore

Verbs
- speak
- give
- run
- can
- may
- do

Adjectives
- fake
- would-be
- red
- happy
- large
- wettest

Adverbs
- happily
- recently

Conjunctions
- and
- or
- if
- because

Determiners
- a
- every
- what
- the

Prepositions
- in
- top
- under
“Today, we eat there.”
Parts of Speech

Nouns
- bread
- milk
- cats
- cat
- Baltimore
- UMBC

Verbs
- speak
- give
- run
- can
- may
- do

Adjectives
- fake
- red
- would-be
- happy
- large
- wettest

Adverbs
- happily
- recently
- then
- there
- (location)

Conjunctions
- and
- or
- if
- because

Determiners
- a
- every
- what
- the

Prepositions
- in
- top
- under

"I ate."
"There is a cat."

Adapted from Luke Zettlemoyer
Parts of Speech

Nouns
- bread
- milk
- cats
- UMBC
- Baltimore

Verbs
- speak
- give
- run
- can
- may
- do

Adjectives
- would-be
- large
- wettest
- fake
- red
- happy

Adverbs
- happily
- recently
- then
- there
- there (location)

Determiners
- a
- every
- what
- the

Conjunctions
- and
- or
- if
- because

Numbers
- 1,324
- one

Prepositions
- in
- top
- under
Parts of Speech

Open class words
- Nouns: bread, milk, cats, cat, Baltimore, UMBC
- Verbs: speak, give, run, can, may, do, modals, auxiliaries
- Adjectives: red, happy, large, wettest, would-be, fake, non-subsective
- Adverbs: happily, recently, then, there (location)
- Prepositions: in, top, under
- Determiners: a, every, the, what
- Conjunctions: and, or, if, because
- Pronouns: I, you, there

Closed class words
- Numbers: 1,324

Adapted from Luke Zettlemoyer
Parts of Speech

Open class words

- Nouns: milk, cats, cat, bread, UMBC, Baltimore
- Verbs: speak, give, run
- Adjectives: red, happy, large, wettest, fake
- Adverbs: happily, recently
- Numbers: 1,324
- Pronouns: I, you, there
- Determiners: a, the, every
- Particles: not, so (far)
- Conjunctions: and, or, if, because
- Prepositions: in, under, top

Closed class words

- Modals, auxiliaries: can, may, do, can, may, do
- Numbers: one

Adapted from Luke Zettlemoyer

Kamp & Partee (1995)
Language evolves!

“I’m reading this because I want to procrastinate.” → “I’m reading this because procrastination.”

Agenda

HMM Motivation (Part of Speech) and Brief Definition

What is Part of Speech?

HMM Detailed Definition

HMM Tasks
Hidden Markov Models: Part of Speech

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

(i): Adjective → Noun → Verb → Prep → Noun → Noun

(ii): Noun → Verb → Noun → Prep → Noun → Noun

Class-based model

Bigram model of the \textit{classes}

Model all class sequences

\( p(w_i|z_i) \)
Hidden Markov Models: Part of Speech

(i): Adjective → Noun → Verb → Prep → Noun → Noun

(ii): Noun → Verb → Noun → Prep → Noun → Noun

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

Class-based model

Bigram model of the classes

Model all class sequences

\[ p(w_i | z_i) \]

\[ p(z_i | z_{i-1}) \]
Hidden Markov Models: Part of Speech

\[ p(w_i | z_i) \quad p(z_i | z_{i-1}) \quad \sum_{z_1,\ldots,z_N} p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) \]

(i): Adjective → Noun → Verb → Prep → Noun → Noun

(ii): Noun → Verb → Noun → Prep → Noun → Noun

p(British Left Waffles on Falkland Islands)
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

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

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

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

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

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

Transition and emission distributions do not change
Hidden Markov Model Terminology

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

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

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

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

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

*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?

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

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

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

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

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

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

Graphical Models (see 478/678)
Hidden Markov Model Representation

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

Emission probabilities/parameters

Transition probabilities/parameters
Hidden Markov Model Representation

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

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

emission

transition

probabilities/parameters

probabilities/parameters
Hidden Markov Model Representation

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

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

Initial starting distribution ("BOS")

\[ p(z_1 | z_0) \]
Hidden Markov Model Representation

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

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

\[ z_1 = \text{V} \]
\[ z_2 = \text{V} \]
\[ z_3 = \text{V} \]
\[ z_4 = \text{V} \]

... 

\[ z_1 = \text{N} \]
\[ z_2 = \text{N} \]
\[ z_3 = \text{N} \]
\[ z_4 = \text{N} \]

... 

\[ w_1 \]
\[ w_2 \]
\[ w_3 \]
\[ w_4 \]
Example: 2-state Hidden Markov Model as a Lattice

\[ z_1 = V \]
\[ z_2 = V \]
\[ z_3 = V \]
\[ z_4 = V \]

\[ z_1 = N \]
\[ z_2 = N \]
\[ z_3 = N \]
\[ z_4 = N \]

\[ p(w_1 | V) \]
\[ p(w_2 | V) \]
\[ p(w_3 | V) \]
\[ p(w_4 | V) \]

\[ p(w_1 | N) \]
\[ p(w_2 | N) \]
\[ p(w_3 | N) \]
\[ p(w_4 | N) \]
Example: 2-state Hidden Markov Model as a Lattice
Example: 2-state Hidden Markov Model as a Lattice
Comparison of Joint Probabilities

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

Unigram Language Model
Comparison of Joint Probabilities

Unigram Language Model

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

Unigram Class-based Language Model (“K” coins)

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = \prod_i p(w_i | z_i) p(z_i) \]
Comparison of Joint Probabilities

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

Unigram Language Model

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

Unigram Class-based Language Model ("K" coins)

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

Hidden Markov Model