Feature Rich Models
(Log Linear Models)

Yejin Choi
University of Washington

[Many slides from Dan Klein, Luke Zettlemoyer]
Announcements

- HW #3 Due
  - Feb 16 Fri?
  - Feb 19 Mon?

- Feb 5 – guest lecture by Max Forbes!
  - VerbPhysics (using a “factor graph” model)
    - Related models: Conditional Random Fields, Markov Random Fields, log-linear models
    - Related algorithms: belief propagation, sum-product algorithm, forward backward
Goals of this Class

- How to construct a feature vector $f(x)$
- How to extend the feature vector to $f(x,y)$
- How to construct a probability model using any given $f(x,y)$
- How to learn the parameter vector $w$ for MaxEnt (log-linear) models
- Knowing the key differences between MaxEnt and Naïve Bayes
- How to extend MaxEnt to sequence tagging
**Structure in the output variable(s)?**

<table>
<thead>
<tr>
<th>What is the input representation?</th>
<th>No Structure</th>
<th>Structured Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generative models (classical probabilistic models)</td>
<td>Naïve Bayes</td>
<td>HMMs PCFGs IBM Models</td>
</tr>
<tr>
<td>Log-linear models (discriminatively trained feature-rich models)</td>
<td>Perceptron Maximum Entropy Logistic Regression</td>
<td>MEMM CRF</td>
</tr>
<tr>
<td>Neural network models (representation learning)</td>
<td>Feedforward NN CNN</td>
<td>RNN LSTM GRU …</td>
</tr>
</tbody>
</table>
Feature Rich Models

- Throw anything (features) you want into the stew (the model)

- Log-linear models
- Often lead to great performance.

Why want richer features?

- POS tagging: more information about the context?
  - Is previous word “the”?
  - Is previous word “the” and the next word “of”?
  - Is previous word capitalized and the next word is numeric?
  - Is there a word “program” within [-5,+5] window?
  - Is the current word part of a known idiom?
  - Conjunctions of any of above?

- Desiderata:
  - Lots and lots of features like above: > 200K
  - No independence assumption among features

- Classical probability models, however
  - Permit very small amount of features
  - Make strong independence assumption among features
HMMs: \( P(\text{tag sequence}|\text{sentence}) \)

- We want a model of sequences \( y \) and observations \( x \)

\[
p(x_1 \ldots x_n, y_1 \ldots y_{n+1}) = q(\text{STOP}|y_n) \prod_{i=1}^{n} q(y_i|y_{i-1}) e(x_i|y_i)
\]

where \( y_0 = \text{START} \) and we call \( q(y'|y) \) the transition distribution and \( e(x|y) \) the emission (or observation) distribution.

- Assumptions:
  - Tag/state sequence is generated by a markov model
  - Words are chosen independently, conditioned only on the tag/state
  - These are totally broken assumptions: why?
PCFGs: $P(\text{parse tree}|\text{sentence})$

$$t_2 = \text{The man saw the woman with the telescope}$$

$$p(t_2) = 1.8 \times 0.3 \times 1.0 \times 0.7 \times 0.2 \times 0.4 \times 1.0 \times 0.3 \times 1.0 \times 0.2 \times 0.4 \times 0.5 \times 0.3 \times 1.0 \times 0.1$$

- Probability of a tree $t$ with rules
  $$\alpha_1 \rightarrow \beta_1, \alpha_2 \rightarrow \beta_2, \ldots, \alpha_n \rightarrow \beta_n$$

  is
  $$p(t) = \prod_{i=1}^{n} q(\alpha_i \rightarrow \beta_i)$$

  where $q(\alpha \rightarrow \beta)$ is the probability for rule $\alpha \rightarrow \beta$. 

Rich features for long range dependencies

- What’s different between basic PCFG scores here?
- What (lexical) correlations need to be scored?
Generative process: (1) generate the very first word conditioning on the special symbol START, then, (2) pick the next word conditioning on the previous word, then repeat (2) until the special word STOP gets picked.

Graphical Model:

Subtleties:

- If we are introducing the special START symbol to the model, then we are making the assumption that the sentence always starts with the special start word START, thus when we talk about \( p(x_1 \ldots x_n) \) it is in fact \( p(x_1 \ldots x_n| x_0 = \text{START}) \)
- While we add the special STOP symbol to the vocabulary \( \mathcal{V}^* \), we do not add the special START symbol to the vocabulary. Why?
Internals of probabilistic models: nothing but adding log-prob

- **LM:** $... + \log p(w_7 \mid w_5, w_6) + \log p(w_8 \mid w_6, w_7) + ...$
- **PCFG:** $\log p(NP \ VP \mid S) + \log p(Papa \mid NP) + \log p(VP \ PP \mid VP) ...$
- **HMM tagging:** $... + \log p(t_7 \mid t_5, t_6) + \log p(w_7 \mid t_7) + ...$
- **Noisy channel:** $[\log p(source)] + [\log p(data \mid source)]$
- **Naïve Bayes:**
  $\log p(Class) + \log p(feature_1 \mid Class) + \log p(feature_2 \mid Class) ...$
arbitrary scores instead of log probs?

Change \( \log p(\text{this} | \text{that}) \) to \( \Phi(\text{this} ; \text{that}) \)

- **LM:** \( ... + \Phi(w7 ; w5, w6) + \Phi(w8 ; w6, w7) + ... \)
- **PCFG:** \( \Phi(\text{NP VP} ; \text{S}) + \Phi(\text{Papa} ; \text{NP}) + \Phi(\text{VP PP} ; \text{VP}) + ... \)
- **HMM tagging:** \( ... + \Phi(t7 ; t5, t6) + \Phi(w7 ; t7) + ... \)
- **Noisy channel:** \( [\Phi(\text{source})] + [\Phi(\text{data} ; \text{source})] \)
- **Naïve Bayes:** \( \Phi(\text{Class}) + \Phi(\text{feature1} ; \text{Class}) + \Phi(\text{feature2} ; \text{Class}) + ... \)
arbitrary scores instead of log probs?

Change \( \log p(\text{this} \mid \text{that}) \) to \( \Phi(\text{this} ; \text{that}) \)

- **LM:** \( \ldots + \Phi(w7 ; w5, w6) + \Phi(w8 ; w6, w7) + \ldots \)
- **PCFG:** \( \Phi(\text{NP VP} ; \text{S}) + \Phi(\text{Papa} ; \text{NP}) + \Phi(\text{VP PP} ; \text{VP}) \ldots \)
- **HMM tagging:** \( \ldots + \Phi(t7 ; t5, t6) + \Phi(w7 ; t7) + \ldots \)
- **Noisy channel:** \( [\Phi(\text{source})] + [\Phi(\text{data} ; \text{source})] \)
- **Naïve Bayes:** \( \Phi(\text{Class}) + \Phi(\text{feature1} ; \text{Class}) + \Phi(\text{feature2} ; \text{Class}) \ldots \)

logistic regression / max-ent
Running example: POS tagging

- Roadmap of (known / unknown) accuracies:
  - Strawman baseline:
    - Most freq tag: ~90% / ~50%
  - Generative models:
    - Trigram HMM: ~95% / ~55%
    - TnT (HMM++): 96.2% / 86.0% (with smart UNK’ing)
  - Feature-rich models?
    - Upper bound: ~98%
### Structure in the output variable(s)?

<table>
<thead>
<tr>
<th>Generative models (classical probabilistic models)</th>
<th>No Structure</th>
<th>Structured Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td></td>
<td>HMMs, PCFGs, IBM Models</td>
</tr>
<tr>
<td>Log-linear models (discriminatively trained feature-rich models)</td>
<td>Perceptron, Maximum Entropy, Logistic Regression</td>
<td>MEMM, CRF</td>
</tr>
<tr>
<td>Neural network models (representation learning)</td>
<td>Feedforward NN, CNN</td>
<td>RNN, LSTM, GRU ...</td>
</tr>
</tbody>
</table>
Throw in various features about the context:

- $f_1 := \text{Is previous word "the" and the next word "of"?}$
- $f_2 := \text{Is previous word capitalized and the next word is numeric?}$
- $f_3 := \text{Frequencies of "the" within $[-15, +15]$ window?}$
- $f_4 := \text{Is the current word part of a known idiom?}$

Given a sentence “the blah ... the truth of ... the blah”

Let’s say $x = \text{“truth”}$ above, then

$f(x) := (f_1, f_2, f_3, f_4)$

$f(\text{truth}) = (true, false, 3, false)$

$=>$

$f(x) = (1, 0, 3, 0)$
Rich features for rich contextual information

- Throw in various features about the context:
  - $f_1 := \text{Is previous word “the” and the next word “of”?}$
  - $f_2 := \ldots$

- You can also define features that look at the output ‘y’!
  - $f_{1\_N} := \text{Is previous word “the” and the next tag is “N”?}$
  - $f_{2\_N} := \ldots$
  - $f_{1\_V} := \text{Is previous word “the” and the next tag is “V”?}$
  - \ldots (replicate all features with respect to different values of y)

$f(x) := (f_1, f_2, f_3, f_4)$
$f(x, y) := (f_{1\_N}, f_{2\_N}, f_{3\_N}, f_{4\_N},$
$f_{1\_V}, f_{2\_V}, f_{3\_V}, f_{4\_V},$
$f_{1\_D}, f_{2\_D}, f_{3\_D}, f_{4\_D},$
\ldots)$
Rich features for rich contextual information

- You can also define features that look at the output ‘y’!
  - \( f_{1 \_N} := \) Is previous word “the” and the next tag is “N”?
  - \( f_{2 \_N} := \) …
  - \( f_{1 \_V} := \) Is previous word “the” and the next tag is “V”?
  - …. (replicate all features with respect to different values of y)

Given a sentence “the blah … the truth of … the blah ”
Let’s say \( x = \) “truth” above, and \( y = \) “N”, then

\[
f(\text{truth}) = (\text{true}, \text{false}, 3, \text{false})
\]

\[
f(x,y) := (f_{1 \_N}, f_{2 \_N}, f_{3 \_N}, f_{4 \_N}, \) \[f(\text{truth}, \text{N}) = ?
\]

\[
f_{1 \_V}, f_{2 \_V}, f_{3 \_V}, f_{4 \_V}, \)
\[
f_{1 \_D}, f_{2 \_D}, f_{3 \_D}, f_{4 \_D}, \)
\[
\text{…..)}
\]
Rich features for rich contextual information

- Throw in various features about the context:
  - $f_1 := \text{Is previous word "the" and the next word "of"?}$
  - $f_2 := \text{Is previous word capitalized and the next word is numeric?}$
  - $f_3 := \text{Frequencies of "the" within [-15,+15] window?}$
  - $f_4 := \text{Is the current word part of a known idiom?}$

- You can also define features that look at the output ‘y’!
  - $f_{1_N} := \text{Is previous word "the" and the next tag is "N"?}$
  - $f_{1_V} := \text{Is previous word "the" and the next tag is "V"?}$

- You can also take any conjunctions of above.

$$f(x, y) = [0, 0, 0, 1, 0, 0, 0, 0, 3, 0.2, 0, 0, 0, \ldots]$$

- Create a very long feature vector with dimensions often $>200K$
- Overlapping features are fine – no independence assumption among features
Goals of this Class

- How to construct a feature vector $f(x)$
- How to extend the feature vector to $f(x,y)$
- How to construct a probability model using any given $f(x,y)$
- How to learn the parameter vector $w$ for MaxEnt (log-linear) models
- Knowing the key differences between MaxEnt and Naïve Bayes
- How to extend MaxEnt to sequence tagging
Maximum Entropy (MaxEnt) Models

- Output: $y$
  - One POS tag for one word (at a time)
- Input: $x$ (any words in the context)
  - Represented as a feature vector $f(x, y)$
- Model parameters: $w$
- Make probability using **SoftMax** function:
- Also known as “**Log-linear**” Models (linear if you take log)

$$p(y|x) = \frac{\exp(w \cdot f(x, y))}{\sum_{y'} \exp(w \cdot f(x, y'))}$$

Make positive! Normalize!
Training MaxEnt Models

- Make probability using **SoftMax** function

\[
p(y|x) = \frac{\exp(w \cdot f(x, y))}{\sum_{y'} \exp(w \cdot f(x, y'))}
\]

- Training:
  - maximize log likelihood of training data \( \{(x^i, y^i)\}_{i=1}^n \)

\[
L(w) = \log \prod_i p(y^i|x^i) = \sum_i \log \frac{\exp(w \cdot f(x^i, y^i))}{\sum_{y'} \exp(w \cdot f(x^i, y'))}
\]

- which also incidentally maximizes the entropy (hence “maximum entropy”)
Training MaxEnt Models

- Make probability using **SoftMax** function

\[
p(y|x) = \frac{\exp(w \cdot f(x, y))}{\sum_{y'} \exp(w \cdot f(x, y'))}
\]

- Training:
  - maximize log likelihood

\[
L(w) = \log \prod_i p(y^i|x^i) = \sum_i \log \frac{\exp(w \cdot f(x^i, y^i))}{\sum_{y'} \exp(w \cdot f(x^i, y'))}
\]

\[
= \sum_i \left( w \cdot f(x^i, y^i) - \log \sum_{y'} \exp(w \cdot f(x^i, y')) \right)
\]
Training MaxEnt Models

\[ L(w) = \sum_i \left( w \cdot f(x^i, y^i) - \log \sum_{y'} \exp(w \cdot f(x^i, y')) \right) \]

Take partial derivative for each \( w_k \) in the weight vector \( w \):

\[ \frac{\partial L(w)}{\partial w_k} = \sum_i \left( f_k(x^i, y^i) - \sum_{y'} p(y'|x^i) f_k(x^i, y') \right) \]

Total count of feature \( k \) with respect to the correct predictions

Expected count of feature \( k \) with respect to the predicted output
Convex Optimization for Training

The likelihood function is convex. (can get global optimum)

Many optimization algorithms/software available.
  - Gradient ascent (descent), Conjugate Gradient, L-BFGS, etc

All we need are:
  1. evaluate the function at current ‘w’
  2. evaluate its derivative at current ‘w’
Goals of this Class

- How to construct a feature vector $f(x)$
- How to extend the feature vector to $f(x,y)$
- How to construct a probability model using any given $f(x,y)$
- How to learn the parameter vector $w$ for MaxEnt (log-linear) models
- Knowing the key differences between MaxEnt and Naïve Bayes
- How to extend MaxEnt to sequence tagging
Graphical Representation of MaxEnt

\[ p(y|x) = \frac{\exp(w \cdot f(x, y))}{\sum_{y'} \exp(w \cdot f(x, y'))} \]
Graphical Representation of Naïve Bayes

\[ p(x|y) = \prod_{j} p(x_j|y) \]
### Naïve Bayes Classifier

<table>
<thead>
<tr>
<th>“Generative” models</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(\text{input}</td>
</tr>
<tr>
<td>For instance, for text categorization, ( P(\text{words}</td>
</tr>
<tr>
<td>Unnecessary efforts on generating input</td>
</tr>
</tbody>
</table>

| Independent assumption among input variables: Given the category, each word is generated independently from other words (too strong assumption in reality!) |
| Cannot incorporate arbitrary/redundant/overlapping features |

### Maximum Entropy (MaxEnt) Classifier

<table>
<thead>
<tr>
<th>“Discriminative” models</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(\text{output}</td>
</tr>
<tr>
<td>For instance, for text categorization, ( P(\text{category}</td>
</tr>
<tr>
<td>Focus directly on predicting the output</td>
</tr>
</tbody>
</table>

| By conditioning on the entire input, we don’t need to worry about the independent assumption among input variables |
| Can incorporate arbitrary features: redundant and overlapping features |
Overview: POS tagging Accuracies

- Roadmap of (known / unknown) accuracies:
  - Most freq tag: ~90% / ~50%
  - Trigram HMM: ~95% / ~55%
  - TnT (HMM++): 96.2% / 86.0%
  - Maxent $P(s_i|x)$: 96.8% / 86.8%

- Q: what’s missing in MaxEnt compared to HMM?

- Upper bound: ~98%
### What is the input representation?

### Structure in the output variable(s)?

| Generative models (classical probabilistic models) | Naïve Bayes | HMMs, PCFGs, IBM Models |
| Log-linear models (discriminatively trained feature-rich models) | Perceptron, Maximum Entropy Logistic Regression | MEMM, CRF |
| Neural network models (representation learning) | Feedforward NN, CNN | RNN, LSTM, GRU, ... |
Goals of this Class

- How to construct a feature vector $f(x)$
- How to extend the feature vector to $f(x,y)$
- How to construct a probability model using any given $f(x,y)$
- How to learn the parameter vector $w$ for MaxEnt (log-linear) models
- Knowing the key differences between MaxEnt and Naïve Bayes
- How to extend MaxEnt to sequence tagging
MEMM Taggers

- **One step up:** also condition on previous tags

\[
p(s_1 \ldots s_m | x_1 \ldots x_m) = \prod_{i=1}^{m} p(s_i | s_1 \ldots s_{i-1}, x_1 \ldots x_m)
\]

\[
= \prod_{i=1}^{m} p(s_i | s_{i-1}, x_1 \ldots x_m)
\]

- Train up \( p(s_i | s_{i-1}, x_1 \ldots x_m) \) as a discrete log-linear (maxent) model, then use to score sequences

\[
p(s_i | s_{i-1}, x_1 \ldots x_m) = \frac{\exp (w \cdot \phi(x_1 \ldots x_m, i, s_{i-1}, s_i))}{\sum_{s'} \exp (w \cdot \phi(x_1 \ldots x_m, i, s_{i-1}, s'))}
\]

- This is referred to as an MEMM tagger [Ratnaparkhi 96]
<table>
<thead>
<tr>
<th>HMM</th>
<th>MEMM</th>
</tr>
</thead>
</table>
| “Generative” models  
⇒ joint probability $p(\text{words, tags})$  
⇒ “generate” input (in addition to tags)  
⇒ but we need to predict tags, not words!  
Probability of each slice = emission * transition =  
$p(\text{word}_i | \text{tag}_i) \times p(\text{tag}_i | \text{tag}_{i-1}) =$  
⇒ Cannot incorporate long distance features | “Discriminative” or “Conditional” models  
⇒ conditional probability $p(\text{tags} | \text{words})$  
⇒ “condition” on input  
⇒ Focusing only on predicting tags  
Probability of each slice =  
$p(\text{tag}_i | \text{tag}_{i-1}, \text{word}_i)$  
or  
$p(\text{tag}_i | \text{tag}_{i-1}, \text{all words})$  
⇒ Can incorporate long distance features |
The HMM State Lattice / Trellis (repeat slide)

\[ e(\text{Fed}|N) \]
\[ q(V|N) \]
\[ e(\text{raises}|V) \]
\[ q(V|V) \]
\[ e(\text{interest}|V) \]
\[ q(V|V) \]
\[ e(\text{rates}|J) \]
\[ q(V|J) \]
\[ e(\text{STOP}|V) \]

START       Fed           raises       interest       rates       STOP
The MEMM State Lattice / Trellis

\[ x = \text{START} \quad \text{Fed} \quad \text{raises} \quad \text{interest} \quad \text{rates} \quad \text{STOP} \]
Decoding: \[ p(s_1 \ldots s_m | x_1 \ldots x_m) = \prod_{i=1}^{m} p(s_i | s_{i-1}, x_1 \ldots x_m) \]

- Decoding maxent taggers:
  - Just like decoding HMMs
  - Viterbi, beam search, posterior decoding

- Viterbi algorithm (HMMs):
  - Define \( \pi(i,s_i) \) to be the max score of a sequence of length \( i \) ending in tag \( s_i \)
    \[ \pi(i, s_i) = \max_{s_{i-1}} e(x_i | s_i)q(s_i | s_{i-1})\pi(i - 1, s_{i-1}) \]

- Viterbi algorithm (Maxent):
  - Can use same algorithm for MEMMs, just need to redefine \( \pi(i,s_i) \)
    \[ \pi(i, s_i) = \max_{s_{i-1}} p(s_i | s_{i-1}, x_1 \ldots x_m)\pi(i - 1, s_{i-1}) \]
Overview: Accuracies

- **Roadmap of (known / unknown) accuracies:**
  - Most freq tag: ~90% / ~50%
  - Trigram HMM: ~95% / ~55%
  - TnT (HMM++): 96.2% / 86.0%
  - Maxent P(s_i|x): 96.8% / 86.8%
  - MEMM tagger: 96.9% / 86.9%

- Upper bound: ~98%
<table>
<thead>
<tr>
<th>What is the input representation?</th>
<th>Structure in the output variable(s)?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Structure</td>
</tr>
<tr>
<td>Generative models (classical probabilistic models)</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>Log-linear models (discriminatively trained feature-rich models)</td>
<td>Perceptron Maximum Entropy Logistic Regression</td>
</tr>
<tr>
<td>Neural network models (representation learning)</td>
<td>Feedforward NN CNN</td>
</tr>
</tbody>
</table>
MEMM v.s. CRF
(Conditional Random Fields)

**MEMM**
- **NNP**: Secretariat
- **VBZ**: is
- **VBN**: expected
- **TO**: to
- **VB**: race
- **NR**: tomorrow

**CRF**
- **NNP**: Secretariat
- **VBZ**: is
- **VBN**: expected
- **TO**: to
- **VB**: race
- **NR**: tomorrow
**MEMM**

- Directed graphical model
- "Discriminative" or "Conditional" models
  - conditional probability $p(\text{tags} | \text{words})$
- Probability is defined for each slice =
  - $P(\text{tag}_i | \text{tag}_{i-1}, \text{word}_i)$
  - or $p(\text{tag}_i | \text{tag}_{i-1}, \text{all words})$

**CRF**

- Undirected graphical model
- Instead of probability, potential (energy function) is defined for each slice =
  - $\phi(\text{tag}_i, \text{tag}_{i-1}) \ast \phi(\text{tag}_i, \text{word}_i)$
    - or $\phi(\text{tag}_i, \text{tag}_{i-1}, \text{all words}) \ast \phi(\text{tag}_i, \text{all words})$

- Can incorporate long distance features
Conditional Random Fields (CRFs)

- Maximum entropy (logistic regression)

\[ p(s|x; w) = \frac{\exp (w \cdot \Phi(x, s))}{\sum_{s'} \exp (w \cdot \Phi(x, s'))} \]

- Learning: maximize the (log) conditional likelihood of training data \( \{(x^i, s^i)\}_{i=1}^{n} \)

\[ \frac{\partial}{\partial w_j} L(w) = \sum_{i=1}^{n} \left( \Phi_j(x_i, s_i) - \sum_{s} p(s|x_i; w) \Phi_j(x_i, s) \right) \]

- Computational Challenges?
  - Most likely tag sequence, normalization constant, gradient

[Lafferty, McCallum, Pereira 01]
Decoding

- **CRFs**
  - Features must be local, for \( x=x_1\ldots x_m \), and \( s=s_1\ldots s_m \)

\[
p(s|x; w) = \frac{\exp (w \cdot \Phi(x, s))}{\sum_{s'} \exp (w \cdot \Phi(x, s'))}
\]

\[
\Phi(x, s) = \sum_{j=1}^{m} \phi(x, j, s_{j-1}, s_j)
\]

\[
\arg \max_{s} \frac{\exp (w \cdot \Phi(x, s))}{\sum_{s'} \exp (w \cdot \Phi(x, s'))} = \arg \max_{s} \exp (w \cdot \Phi(x, s))
\]

\[
= \arg \max_{s} w \cdot \Phi(x, s)
\]

- **Viterbi recursion**

\[
\pi(i, s_i) = \max_{s_{i-1}} w \cdot \phi(x, i, s_{i-1}, s_i) + \pi(i - 1, s_{i-1})
\]
CRFs: Computing Normalization

\[ p(s|x; w) = \frac{\exp (w \cdot \Phi(x, s))}{\sum_{s'} \exp (w \cdot \Phi(x, s'))} \]
\[ \Phi(x, s) = \sum_{j=1}^{m} \phi(x, j, s_{j-1}, s_j) \]
\[ \sum_{s'} \exp (w \cdot \Phi(x, s')) = \sum_{s'} \exp \left( \sum_{j} w \cdot \phi(x, j, s_{j-1}, s_j) \right) \]
\[ = \sum_{s'} \prod_{j} \exp (w \cdot \phi(x, j, s_{j-1}, s_j)) \]

Define \( \text{norm}(i,s_i) \) to sum of scores for sequences ending in position \( i \)

\[ \text{norm}(i, y_i) = \sum_{s_{i-1}} \exp (w \cdot \phi(x, i, s_{i-1}, s_i)) \text{norm}(i - 1, s_{i-1}) \]

- **Forward Algorithm! Remember HMM case:**

\[ \alpha(i, y_i) = \sum_{y_{i-1}} e(x_i|y_i)q(y_i|y_{i-1})\alpha(i - 1, y_{i-1}) \]

- Could also use backward?
CRFs: Computing Gradient*

\[ p(s \mid x; w) = \frac{\exp (w \cdot \Phi(x, s))}{\sum_{s'} \exp (w \cdot \Phi(x, s'))} \]

\[ \Phi(x, s) = \sum_{j=1}^{m} \phi(x, j, s_{j-1}, s_j) \]

\[ \frac{\partial}{\partial w_j} L(w) = \sum_{i=1}^{n} \left( \Phi_j(x_i, s_i) - \sum_{s} p(s \mid x_i; w) \Phi_j(x_i, s) \right) \]

\[ \sum_{s} p(s \mid x_i; w) \Phi_j(x_i, s) = \sum_{s} p(s \mid x_i; w) \sum_{j=1}^{m} \phi_k(x_i, j, s_{j-1}, s_j) \]

\[ = \sum_{j=1}^{m} \sum_{a,b} \sum_{s:s_{j-1}=a,s_b=b} p(s \mid x_i; w) \phi_k(x_i, j, s_{j-1}, s_j) \]

- Need forward and backward messages

See notes for full details!
Overview: Accuracies

- Roadmap of (known / unknown) accuracies:
  - Most freq tag: ~90% / ~50%
  - Trigram HMM: ~95% / ~55%
  - TnT (HMM++): 96.2% / 86.0%
  - Maxent P(s_i|x): 96.8% / 86.8%
  - MEMM tagger: 96.9% / 86.9%
  - CRF (untuned) 95.7% / 76.2%
  - Upper bound: ~98%
Cyclic Network

- Train two MEMMs, multiple together to score
- And be very careful
  - Tune regularization
  - Try lots of different features
  - See paper for full details

(a) Left-to-Right CMM

(b) Right-to-Left CMM

(c) Bidirectional Dependency Network
Overview: Accuracies

- Roadmap of (known / unknown) accuracies:
  - Most freq tag: ~90% / ~50%
  - Trigram HMM: ~95% / ~55%
  - TnT (HMM++): 96.2% / 86.0%
  - Maxent P(s_i|x): 96.8% / 86.8%
  - MEMM tagger: 96.9% / 86.9%
  - Perceptron: 96.7% / ??
  - CRF (untuned): 95.7% / 76.2%
  - Cyclic tagger: 97.2% / 89.0%
  - Upper bound: ~98%
Locally normalized models

- HMMs, MEMMs
- Local scores are probabilities
- However: one issue in local models
  - “Label bias” and other **explaining away effects**
  - MEMM taggers’ local scores can be near one without having both good “transitions” and “emissions”
  - This means that often evidence doesn’t flow properly
  - Why isn’t this a big deal for POS tagging?

Globally normalized models

- Local scores are arbitrary scores
- Conditional Random Fields (CRFs)
- Slower to train (structured inference at each iteration of learning)
- Neural Networks (global training w/o structured inference)
### Structure in the output variable(s)?

<table>
<thead>
<tr>
<th>What is the input representation?</th>
<th>No Structure</th>
<th>Structured Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generative models (classical probabilistic models)</td>
<td>Naïve Bayes</td>
<td>HMMs PCFGs IBM Models</td>
</tr>
<tr>
<td>Log-linear models (discriminatively trained feature-rich models)</td>
<td>Perceptron Maximum Entropy Logistic Regression</td>
<td>MEMM CRF</td>
</tr>
<tr>
<td>Neural network models (representation learning)</td>
<td>Feedforward NN CNN</td>
<td>RNN LSTM GRU …</td>
</tr>
</tbody>
</table>
Supplementary Material
Graphical Models

- Conditional probability for each node
  - e.g. \( p( Y_3 | Y_2, X_3 ) \) for \( Y_3 \)
  - e.g. \( p( X_3 ) \) for \( X_3 \)

- Conditional independence
  - e.g. \( p( Y_3 | Y_2, X_3 ) = p( Y_3 | Y_1, Y_2, X_1, X_2, X_3) \)

- Joint probability of the entire graph
  = product of conditional probability of each node
Undirected Graphical Model Basics

- Conditional independence
  - e.g. \( p( Y3 \mid \text{all other nodes} ) = p( Y3 \mid Y3' \text{' neighbor} ) \)
- No conditional probability for each node
- Instead, "potential function" for each clique
  - e.g. \( \phi ( X1, X2, Y1 ) \) or \( \phi ( Y1, Y2 ) \)
- Typically, log-linear potential functions
  \[ \phi ( Y1, Y2 ) = \exp \sum_k w_k f_k (Y1, Y2) \]
Undirected Graphical Model Basics

- Joint probability of the entire graph

\[ P(\vec{Y}) = \frac{1}{Z} \prod_{\text{clique } C} \varphi(\vec{Y}_C) \]

\[ Z = \sum_{\vec{Y}} \prod_{\text{clique } C} \varphi(\vec{Y}_C) \]