Leftovers: Unsupervised TBL

- Initialization: a list of allowable part of speech tags
- Transformations: Change the tag of a word from $\chi$ to $Y$ in context $C$, where $\gamma \in \chi$.
  
  Example: “From NN/VBP to VBP if previous tag is NNS”
- Scoring criterion:
  
  $$ R = \arg \max_{Z \times Z \neq Y} \frac{freq(X)}{freq(Z) \times incontext(Z, C)} $$

  $$ score = incontext(\gamma, C) - \frac{Y}{freq(R) \times incontext(R, C')} $$

Leftovers: POS distribution

The number of word types in Brown corpus by degree of ambiguity

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous (1 tag)</td>
<td>35,340</td>
</tr>
<tr>
<td>Ambiguous (2-7 tags)</td>
<td>4,100</td>
</tr>
<tr>
<td>2 tags</td>
<td>3,764</td>
</tr>
<tr>
<td>3 tags</td>
<td>264</td>
</tr>
<tr>
<td>4 tags</td>
<td>61</td>
</tr>
<tr>
<td>5 tags</td>
<td>12</td>
</tr>
<tr>
<td>6 tags</td>
<td>2</td>
</tr>
<tr>
<td>7 tags</td>
<td>1</td>
</tr>
</tbody>
</table>
The General Problem

- We have some **input domain** $\chi$
- We have some **label set** $\gamma$
- Goal: learn a **conditional probability** $P(y|x)$ for any $x \in \chi$ and $y \in \gamma$

POS tagging: Representation

*Our/PRP$ enemies/NNS are/VBP innovative/JJ and/CC resourceful/JJ ,/, and/CC so/RB are/VB we/PRP ?/?.*

- History is a 4-tuples $(t_1, t_2, w[1:n], i)$
- $t_1, t_2$ are the previous two tags
- $w[1:n]$ are the $n$ words in the input sentence
- $i$ is the index of the word being tagged

$\chi$ is the set of all possible histories

POS tagging

*Our/PRP$ enemies/NNS are/VBP innovative/JJ and/CC resourceful/JJ ,/, and/CC so/RB are/VB we/PRP ?/?.*

- Input domain: $\chi$ is the set of possible histories
- Label set: $\gamma$ is the set of all possible tags
- Goal: learn a **conditional probability** $P(tag|history)$
Feature Vector Representation

- A feature is a function $f : \chi \times \gamma \rightarrow 0$

$$f(h, t) = \begin{cases} 1 & \text{if current word } w_i \text{ is "are" and } t = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

- We have $m$ features $f_k$ for $k = 1 \ldots m$

Maximum Entropy and Log-linear Models

POS Representation

- Word/tag features for all word/tag pairs:

$$f_{55}(h, t) = \begin{cases} 1 & \text{if current word } w_i \text{ is "our" and } t = \text{PRP} \\ 0 & \text{otherwise} \end{cases}$$

- Spelling features for all prefixes/suffixes of certain length:

$$f_{70}(h, t) = \begin{cases} 1 & \text{if current word } w_i \text{ ends in "ing" and } t = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

- Contextual features:

$$f_{112}(h, t) = \begin{cases} 1 & \text{if previous word } w_i \text{ is "the" and } t = \text{Vt} \\ 0 & \text{otherwise} \end{cases}$$

Maximum Entropy and Log-linear Models

Maximum Entropy: Motivating Example

Estimate probability distribution $p(a, b)$, given the constraint: $p(x, 0) + p(y, 0) = 0.6$, where $a \in \{x, y\}$ and $b \in 0, 1$.

<table>
<thead>
<tr>
<th>$p(a, b)$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>y</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>total</td>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Goal: learn a conditional probability $P(tag|history)$
Another Way To Satisfy Constraints

<table>
<thead>
<tr>
<th>p(a, b)</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>y</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>total</td>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

One Way To Satisfy Constraints

<table>
<thead>
<tr>
<th>p(a, b)</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>y</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>total</td>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Maximum Entropy and Log-linear Models

Representing Evidence

Constraint: *observed expectation* of each feature has to be the same as *the model’s expectation* of the feature:

\[ E_p f_j = E_{p'} f_j (j = 1 \ldots m), \]

\[ E_p f_j = \sum_{x \in \{X \neq \gamma\}} p(x) f_j(x) \text{ (model’s expectation)} \]
\[ E_{p'} f_j = \sum_{x \in \{X \neq \gamma\}} p'(x) f_j(x) \text{ (observed expectation)} \]

Maximum Entropy Modeling

Given a set of training examples, we wish to find a distribution which:

- satisfies the input constraints
- maximizes the uncertainty

\[
\ldots \text{in making inference on the basis of partial information we must use the probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have. Jaynes, 1957}
\]
**Outline**

- We will first show that
  \[ p^*(x) = \pi \prod_{j=1}^{k} \alpha_j^{f_j}(x), \quad 0 < \alpha_j < \infty, \]
  where \( \pi \) is a normalization constant and the \( \alpha \)'s are the model parameters.
- Then, we will consider an estimation procedure for finding the \( \alpha \)'s.

**Principle of Maximum Entropy**

\[ P = \{ p | E_p f_j = E_{p'} f_j, j = \{1 \ldots m\} \} \]
\[ p^* = \arg \max_{p \in P} H(p) \]

**Relative Entropy (Kullback-Liebler Distance)**

- **Definition:** The relative entropy \( D \) between two probability distributions \( p \) and \( q \) is given by:
  \[ D(p, q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)} \]
- **Lemma 1:** For any two probability distributions \( p \) and \( q \), \( D(p, q) \geq 0 \), and \( D(p, q) = 0 \) if and only if \( p = q \)

**Notations**

- \( \chi \) is the set of possible histories
- \( \gamma \) is the set of all possible tags
- \( S \) finite training sample of events
- \( p'(x) \) observed probability of \( x \) in \( S \)
- \( p(x) \) the model's probability of \( x \)
- \( f_j \) function of type \( \chi \times \gamma \rightarrow \{0, 1\} \)
- \( E_p f_j = \sum_{x \in \chi \times \gamma} p(x) f_j(x) \)
- \( E_{p'} f_j = \sum_{x \in \chi \times \gamma} p'(x) f_j(x) \)
- \( P = \{ p | E_p f_j = E_{p'} f_j, j = \{1 \ldots m\} \} \)
- \( Q = \{ p | p(x) = \pi \prod_{j=1}^{k} \alpha_j^{f_j}(x) \} \)
- \( H(p) = -\sum_{x \in \chi \times \gamma} p(x) \log p(x) \)
- \( L(p) = \sum_{x \in \chi \times \gamma} p'(x) \log p(x) \)
Let \( p \in P \), \( q \in Q \) and \( p^* \in P \cap Q \):

\[
D(p, p^*) + D(p^*, q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log p^*(x) + \sum_x p^*(x) \log p^*(x) - \sum_x p^*(x) \log q(x)
\]

The Maximum Likelihood Solution

**Theorem 2** If \( p^* \in P \cap Q \), then \( p^* = \arg\max_{q \in Q} L(q) \). Furthermore, \( p^* \) is unique.

**Proof.** Let \( \hat{p}(x) \) be the observed distribution of \( x \) in the sample \( S \), \( \forall x \in \varepsilon \). Clearly \( \hat{p} \in P \).

Suppose \( q \in Q \) and \( p^* \in P \cap Q \).

- Show that \( L(q) \leq L(p^*) \):
  - By Lemma 2,
    \[
    D(\hat{p}, q) = D(\hat{p}, p^*) + D(p^*, q)
    \]
  - and by Lemma 1,
    \[
    D(\hat{p}, q) \geq D(\hat{p}, p^*)
    \]
  - \( -H(\hat{p}) - L(q) \geq -H(\hat{p}) - L(p^*) \)
  - \( L(q) \leq L(p^*) \)

- Show \( p^* \) is unique:
  \[
  L(q) = L(p^*) \implies D(\hat{p}, q) = D(\hat{p}, p^*) \implies D(p^*, q) = 0 \implies p^* = q
  \]

**Pythagorean Property**

**Lemma 2** (Pythagorean Property): If \( p \in P \) and \( q \in Q \), and \( p^* \in P \cap Q \), then

\[
D(p, q) = D(p, p^*) + D(p^*, q)
\]

**Proof.** For any \( r, x \in P \), and \( t \in Q \)

\[
\sum_x r(x) \log t(x) = \sum_x r(x) [\log \pi + \sum_j f_j(x) \log \alpha_j]
\]

\[
= \sum_x r(x) [\log \pi + \sum_j \log \alpha_j \sum_x r(x) f_j(x)]
\]

\[
= \sum_x s(x) [\log \pi + \sum_j \log \alpha_j \sum_x s(x) f_j(x)]
\]

\[
= \sum_x s(x) [\log \pi + \sum_j f_j(x) \log \alpha_j] = \sum_x s(x) \log t(x)
\]

**The Maximum Entropy Solution**

**Theorem 1** If \( p^* \in P \cap Q \), then \( p^* = \arg\max_{p \in P} H(p) \). Furthermore, \( p^* \) is unique.

**Proof.** Suppose \( p \in P \) and \( p^* \in P \cap Q \). Let \( u \in Q \) be the uniform distribution so that \( \forall x \in \varepsilon u(x) = \frac{1}{|\varepsilon|} \).

- Show that \( H(p) \leq H(p^*) \):
  - By Lemma 2,
    \[
    D(p, u) = D(p, p^*) + D(p^*, u)
    \]
  - and by Lemma 1,
    \[
    D(p, u) \geq D(p^*, u)
    \]
  - \( -H(p) - \log \frac{1}{|\varepsilon|} \geq -H(p^*) - \log \frac{1}{|\varepsilon|} \)
  - \( H(p) \leq H(p^*) \)

- Show \( p^* \) is unique:
  \[
  H(p) = H(p^*) \implies D(p, u) = D(p^*, u) \implies D(p, p^*) = 0 \implies p = p^*
  \]
Duality Theorem

There is a unique distribution \( p^* \)

1. \( p^* \in P \cap Q \)
2. \( p^* = \arg\max_{p \in P} H(p) \) (Max-ent solution)
3. \( p^* = \arg\max_{q \in Q} L(q) \) (Max-likelihood solution)

**Implications:**

- The maximum entropy solution can be written in log-linear form
- Finding the maximum-likelihood solution also gives the maximum entropy solution
**Summary**

- Modeling conditional probabilities with log-linear models
- Maximum-entropy properties of log-linear models
- Optimization via iterative scaling

Some implementations:

**ME classifiers**

- Can handle lots of features
- Sparsity is an issue
  - apply smoothing and feature selection
- Feature interaction
  - ME classifiers do not assume feature independence
  - However, they do not explicitly model feature interaction