6.891: Lecture 6 (September 24, 2003)
Log-Linear Models

Overview

- Log-linear models
- The maximum-entropy property
- Smoothing, feature selection etc. in log-linear models

Tagging Problems

- Mapping strings to Tagged Sequences

a b e e a f h j $\Rightarrow$ a/C b/D e/C e/C a/D f/C h/D j/C

Part-of-Speech Tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

N = Noun
V = Verb
P = Preposition
Adv = Adverb
Adj = Adjective
. . .

Information Extraction

Named Entity Recognition

INPUT: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT: Profits soared at [Company Boeing Co.], easily topping forecasts on [Location Wall Street], as their CEO [Person Alan Mulally] announced first quarter results.

Named Entity Extraction as Tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

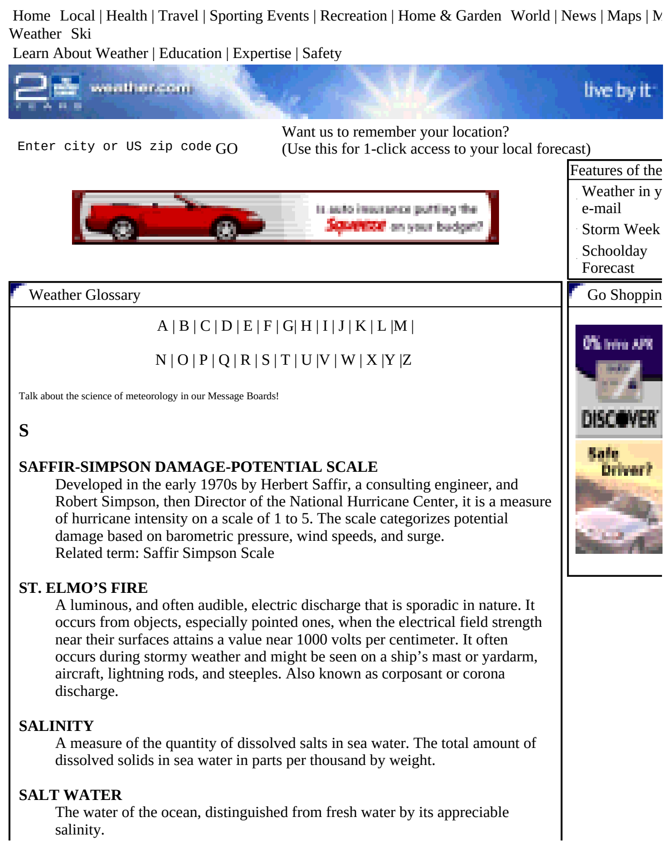
OUTPUT:

Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ,/NA as/NA their/NA CEO/NA Alan/SP Mulally/CP announced/NA first/NA quarter/NA results/NA ./NA

NA = No entity
SC = Start Company
CC = Continue Company
SL = Start Location
CL = Continue Location
. . .

Extracting Glossary Entries from the Web

Input:



Home Local | Health | Travel | Sporting Events | Recreation | Home & Garden World | News | Maps | M
Weather Ski
Learn About Weather | Education | Expertise | Safety

weather.com live by it

Enter city or US zip code GO

Want us to remember your location?
(Use this for 1-click access to your local forecast)

Is auto insurance putting the Squeeze on your budget?

Features of the
Weather in y
e-mail
Storm Week
Schoolday
Forecast

Weather Glossary

A | B | C | D | E | F | G| H | I | J | K | L |M |

N | O | P | Q | R | S | T | U |V | W | X |Y |Z

Talk about the science of meteorology in our Message Boards!

S

SAFFIR-SIMPSON DAMAGE-POTENTIAL SCALE
Developed in the early 1970s by Herbert Saffir, a consulting engineer, and Robert Simpson, then Director of the National Hurricane Center, it is a measure of hurricane intensity on a scale of 1 to 5. The scale categorizes potential damage based on barometric pressure, wind speeds, and surge.
Related term: Saffir Simpson Scale

ST. ELMO'S FIRE
A luminous, and often audible, electric discharge that is sporadic in nature. It occurs from objects, especially pointed ones, when the electrical field strength near their surfaces attains a value near 1000 volts per centimeter. It often occurs during stormy weather and might be seen on a ship's mast or yardarm, aircraft, lightning rods, and steeples. Also known as corposant or corona discharge.

SALINITY
A measure of the quantity of dissolved salts in sea water. The total amount of dissolved solids in sea water in parts per thousand by weight.

SALT WATER
The water of the ocean, distinguished from fresh water by its appreciable salinity.

Go Shoppin

0% Intro APR
DISCOVER
Safe Driver?

Output: **St. Elmo's Fire:** A luminous, and often audible, electric discharge that is sporadic in nature. It occurs from objects, especially pointed ones, when the electrical field strength near their surfaces attains a value near 100 volts per centimeter...

The General Problem

- We have some **input domain** $\mathcal{X}$
- Have a finite **label set** $\mathcal{Y}$
- Aim is to provide a **conditional probability** $P(y \mid x)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

An Example

Hispaniola/NNP quickly/RB became/VB an/DT important/JJ base/?? from which Spain expanded its empire into the rest of the Western Hemisphere .

- There are many possible tags in the position ??
$\mathcal{Y} = \{$NN, NNS, Vt, Vi, IN, DT, $\ldots\}$

- The input domain $\mathcal{X}$ is the set of all possible **histories** (or contexts)

- Need to learn a function from (history, tag) pairs to a probability $P(tag|history)$

Representation: Histories

- A **history** is a 4-tuple $\langle t_{-1}, t_{-2}, w_{[1:n]}, i \rangle$
- 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
- $\mathcal{X}$ is the set of all possible histories

Hispaniola/NNP quickly/RB became/VB an/DT important/JJ base/?? from which Spain expanded its empire into the rest of the Western Hemisphere .

- $t_{-1}, t_{-2} =$ DT, JJ
- $w_{[1:n]} = \langle Hispaniola, quickly, became, \ldots, Hemisphere, . \rangle$
- $i = 6$

Feature Vector Representations

- We have some input domain $\mathcal{X}$, and a finite label set $\mathcal{Y}$. Aim is to provide a conditional probability $P(y \mid x)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- A **feature** is a function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$
(Often **binary features** or **indicator functions** $f : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}$).

- Say we have m features ϕ_k for $k = 1 \ldots m$
$\Rightarrow$ A **feature vector** $\phi(x, y) \in \mathbb{R}^m$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

An Example (continued)

- $\mathcal{X}$ is the set of all possible histories of form $\langle t_{-1}, t_{-2}, w_{[1:n]}, i \rangle$
- $\mathcal{Y} = \{$NN, NNS, Vt, Vi, IN, DT, $\ldots\}$
- We have m features $\phi_k : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ for $k = 1 \ldots m$

For example:

$$\phi_1(h, t) = \begin{cases} 1 & \text{if current word } w_i \text{ is } \texttt{base} \text{ and } t = \texttt{Vt} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_2(h, t) = \begin{cases} 1 & \text{if current word } w_i \text{ ends in } \texttt{ing} \text{ and } t = \texttt{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$\ldots$$

$\phi_1(\langle \text{JJ, DT, } \langle \text{ Hispaniola, } \ldots \rangle, 6 \rangle, \text{Vt}) = 1$
$\phi_2(\langle \text{JJ, DT, } \langle \text{ Hispaniola, } \ldots \rangle, 6 \rangle, \text{Vt}) = 0$
$\ldots$

The Full Set of Features in [Ratnaparkhi 96]

- Word/tag features for all word/tag pairs, e.g.,

$$\phi_{100}(h,t) = \begin{cases} 1 & \text{if current word } w_i \text{ is } \texttt{base} \text{ and } t = \texttt{Vt} \\ 0 & \text{otherwise} \end{cases}$$

- Spelling features for all prefixes/suffixes of length ≤ 4, e.g.,

$$\phi_{101}(h,t) = \begin{cases} 1 & \text{if current word } w_i \text{ ends in } \texttt{ing} \text{ and } t = \texttt{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{102}(h,t) = \begin{cases} 1 & \text{if current word } w_i \text{ starts with } \texttt{pre} \text{ and } t = \texttt{NN} \\ 0 & \text{otherwise} \end{cases}$$

The Full Set of Features in [Ratnaparkhi 96]

- Contextual Features, e.g.,

$$\phi_{103}(h,t) = \begin{cases} 1 & \text{if } \langle t_{-2}, t_{-1}, t\rangle = \langle \text{DT, JJ, Vt}\rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{104}(h,t) = \begin{cases} 1 & \text{if } \langle t_{-1}, t\rangle = \langle \text{JJ, Vt}\rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{105}(h,t) = \begin{cases} 1 & \text{if } \langle t\rangle = \langle \text{Vt}\rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{106}(h,t) = \begin{cases} 1 & \text{if previous word } w_{i-1} = \textit{the} \text{ and } t = \text{Vt} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{107}(h,t) = \begin{cases} 1 & \text{if next word } w_{i+1} = \textit{the} \text{ and } t = \text{Vt} \\ 0 & \text{otherwise} \end{cases}$$

The Final Result

- We can come up with practically any questions (*features*) regarding history/tag pairs.

- For a given history $x \in \mathcal{X}$, each label in $\mathcal{Y}$ is mapped to a different feature vector

$$
\begin{aligned}
\phi(\langle \text{JJ, DT, } \langle \text{ Hispaniola, } \ldots \rangle, 6 \rangle, \text{Vt}) &= 1001011001001100110 \\
\phi(\langle \text{JJ, DT, } \langle \text{ Hispaniola, } \ldots \rangle, 6 \rangle, \text{JJ}) &= 0110010101011110010 \\
\phi(\langle \text{JJ, DT, } \langle \text{ Hispaniola, } \ldots \rangle, 6 \rangle, \text{NN}) &= 0001111101001100100 \\
\phi(\langle \text{JJ, DT, } \langle \text{ Hispaniola, } \ldots \rangle, 6 \rangle, \text{IN}) &= 0001011011000000010 \\
&\ldots
\end{aligned}
$$

Log-Linear Models

- We have some input domain $\mathcal{X}$, and a finite label set $\mathcal{Y}$. Aim is to provide a conditional probability $P(y \mid x)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- A feature is a function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$
(Often binary features or indicator functions $f : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}$).

- Say we have m features ϕ_k for $k = 1 \ldots m$
$\Rightarrow$ A feature vector $\phi(x, y) \in \mathbb{R}^m$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- We also have a **parameter vector W** $\in \mathbb{R}^m$

- We define

$$\boxed{P(y \mid x, \mathbf{W}) = \frac{e^{\mathbf{W} \cdot \phi(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x,y')}}}$$

More About Log-Linear Models

- Why the name?

$$\log P(y \mid x, \mathbf{W}) = \underbrace{\mathbf{W} \cdot \phi(x, y)}_{\text{Linear term}} - \underbrace{\log \sum_{y' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x, y')}}_{\text{Normalization term}}$$

- Maximum-likelihood estimates given training sample (x_i, y_i) for $i = 1 \ldots n$, each $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$:

$$\mathbf{W}_{ML} = \text{argmax}_{\mathbf{W} \in \mathbb{R}^m} L(\mathbf{W})$$

where

$$\begin{aligned} L(\mathbf{W}) &= \sum_{i=1}^{n} \log P(y_i \mid x_i) \\ &= \sum_{i=1}^{n} \mathbf{W} \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, y')} \end{aligned}$$

Calculating the Maximum-Likelihood Estimates

- Need to maximize:

$$L(\mathbf{W}) = \sum_{i=1}^{n} \mathbf{W} \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, y')}$$

- Calculating gradients:

$$\begin{aligned}
\left.\frac{dL}{d\mathbf{W}}\right|_{\mathbf{W}} &= \sum_{i=1}^{n} \phi(x_i, y_i) - \sum_{i=1}^{n} \frac{\sum_{y' \in \mathcal{Y}} \phi(x_i, y') e^{\mathbf{W} \cdot \phi(x_i, y')}}{\sum_{z' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, z')}} \\
&= \sum_{i=1}^{n} \phi(x_i, y_i) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} \phi(x_i, y') \frac{e^{\mathbf{W} \cdot \phi(x_i, y')}}{\sum_{z' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, z')}} \\
&= \underbrace{\sum_{i=1}^{n} \phi(x_i, y_i)}_{\text{Empirical counts}} \underbrace{- \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} \phi(x_i, y') P(y' \mid x_i, \mathbf{W})}_{\text{Expected counts}}
\end{aligned}$$

Gradient Ascent Methods

- Need to maximize $L(\mathbf{W})$ where

$$\left.\frac{dL}{d\mathbf{W}}\right|_{\mathbf{W}} = \sum_{i=1}^{n} \phi(x_i, y_i) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} \phi(x_i, y') P(y' \mid x_i, \mathbf{W})$$

Initialization: $\mathbf{W} = 0$

Iterate until convergence:

- Calculate $\Delta = \left.\frac{dL}{d\mathbf{W}}\right|_{\mathbf{W}}$
- Calculate $\beta_* = \text{argmax}_\beta L(\mathbf{W} + \beta\Delta)$ (Line Search)
- Set $\mathbf{W} \leftarrow \mathbf{W} + \beta_*\Delta$

Conjugate Gradient Methods

- (Vanilla) gradient ascent can be very slow

- Conjugate gradient methods require calculation of gradient at each iteration, but do a line search in **a direction which is a function of the current gradient, and the previous step taken**.

- Conjugate gradient packages are widely available
  In general: they require a function

$$\texttt{calc\_gradient}(\mathbf{W}) \rightarrow \left(L(\mathbf{W}), \left. \frac{dL}{d\mathbf{W}} \right|_{\mathbf{W}} \right)$$

and that's about it!

Iterative Scaling

Initialization:

$\mathbf{W} = 0$

Calculate $\mathbf{H} = \sum_i \phi(x_i, y_i)$ (Empirical counts)

Calculate $C = \max_{i=1\ldots n, y \in \mathcal{Y}} \left(\sum_{k=1}^{m} \phi_k(x_i, y)\right)$

Iterate until convergence:

Calculate $\mathbf{E}(\mathbf{W}) = \sum_i \sum_{y' \in \mathcal{Y}} \phi(x_i, y') P(y' \mid x_i, \mathbf{W})$
(Expected counts)

For $k = 1 \ldots m$, set $\mathbf{W}_k \leftarrow \mathbf{W}_k + \frac{1}{C} \log \frac{\mathbf{H}_k}{\mathbf{E}_k(\mathbf{W})}$

Converges to maximum-likelihood solution provided that $\phi_k(x_i, y_i) \geq 0$ for all i, k.

Derivation of Iterative Scaling

Consider a vector of updates $\delta \in \mathbb{R}^m$, so that $W_{k+1} = W_k + \delta$. The gain in log-likelihood is then $L(\mathbf{W} + \delta) - L(\mathbf{W})$.

$$\begin{aligned}
& L(\mathbf{W}+\delta) - L(\mathbf{W}) \\
= \quad & \sum_{i=1}^{n} (\mathbf{W}+\delta) \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} e^{(\mathbf{W}+\delta)\cdot\phi(x_i, y')} \\
& - \left(\sum_{i=1}^{n} \mathbf{W} \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} e^{\mathbf{W}\cdot\phi(x_i, y')} \right) && (1) \\
= \quad & \sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \frac{\sum_{y' \in \mathcal{Y}} e^{(\mathbf{W}+\delta)\cdot\phi(x_i, y')}}{\sum_{z \in \mathcal{Y}} e^{\mathbf{W}\cdot\phi(x_i, z)}} && (2) \\
= \quad & \sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) e^{\delta\cdot\phi(x_i, y')} && (3)
\end{aligned}$$

$$\geq \sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) + 1 - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) e^{\delta \cdot \phi(x_i, y')} \tag{4}$$

(From $\quad -\log(x) \geq 1 - x$)

$$= \sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) + 1$$

$$- \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) \exp\left\{(\delta \cdot \phi(x_i, y') + 0.(C - C_i(y')))\right\} \tag{5}$$

(Where $C_i(y') = \sum_k \phi_k(x_i, y')$, and $C = \max_{i,y'} C_i(y')$)

$$\geq \sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) + 1$$

$$- \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) \left(\sum_k \frac{\phi(x_i, y')}{C} e^{C\delta_k} + \frac{C - C_i(y')}{C} \right) \tag{6}$$

(From $e^{\sum_x q(x) f(x)} \leq \sum_x q(x) e^{f(x)}$ for any $q(x) \geq 0$, and $\sum_x q(x) = 1$)

$$= A(\mathbf{W}, \delta) \tag{7}$$

- We now have an auxilliary function $A(\mathbf{W}, \delta)$ such that

$$L(\mathbf{W}, \delta) - L(\mathbf{W}) \geq A(\mathbf{W}, \delta)$$

- Now maximize $A(\mathbf{W}, \delta)$ with respect to each δ_k:

$$\begin{aligned}\frac{dA}{d\delta_k} &= \sum_{i=1}^{n} \phi_k(x_i, y_i) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) \phi_k(x_i, y') e^{C\delta_k} \\ &= \mathbf{H}_k - e^{C\delta_k} \mathbf{E}_k(\mathbf{W})\end{aligned}$$

Setting derivatives equal to 0 gives iterative scaling:

$$\delta_k = \frac{1}{C} \log \frac{\mathbf{H}_k}{\mathbf{E}_k(\mathbf{W})}$$

Improved Iterative Scaling (Berger et. al)

$$\sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) + 1 - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) e^{\delta \cdot \phi(x_i, y')} \tag{8}$$

$$\geq \sum_{i=1}^{n} \delta \cdot \phi(x_i, y_i) + 1$$

$$- \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) \left(\sum_{k} \frac{\phi(x_i, y')}{f(x_i, y')} e^{f(x_i, y')\delta_k} \right) \tag{9}$$

(Where $f(x_i, y') = \sum_k \phi_(x_i, y')$, (10)

and from $e^{\sum_x q(x) f(x)} \leq \sum_x q(x) e^{f(x)}$ for any $q(x) \geq 0$, and $\sum_x q(x) = 1$)

$$= A(\mathbf{W}, \delta) \tag{11}$$

Maximizing $A(\mathbf{W}, \delta)$ w.r.t. δ involves finding δ_k's which solve:

$$\sum_{i=1}^{n} \phi_k(x_i, y_i) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} p(y' \mid x_i, \mathbf{W}) \phi_k(x_i, y') e^{f(x_i, y')\delta_k} = 0$$

Overview

- Log-linear models
- The maximum-entropy property
- Smoothing, feature selection etc. in log-linear models

Maximum-Entropy Properties of Log-Linear Models

- We define the set of distributions which satisfy linear constraints implied by the data:

$$\mathcal{P} = \{p \;:\; \underbrace{\sum_i \phi(x_i, y_i)}_{\text{Empirical counts}} = \underbrace{\sum_i \sum_{y \in \mathcal{Y}} p(y \mid x_i) \phi(x_i, y)}_{\text{Expected counts}}\}$$

  here, p is an $n \times |\mathcal{Y}|$ vector defining $P(y \mid x_i)$ for all i, y.

- Note that at least one distribution satisfies these constraints, i.e.,

$$p(y \mid x_i) = \begin{cases} 1 & \text{if } y = y_i \\ 0 & \text{otherwise} \end{cases}$$

Maximum-Entropy Properties of Log-Linear Models

- The **entropy** of any distribution is:

$$H(p) = -\left(\frac{1}{n}\sum_{i}\sum_{y\in\mathcal{Y}} p(y \mid x_i)\log p(y \mid x_i)\right)$$

- Entropy is a measure of “smoothness” of a distribution

- In this case, entropy is maximized by uniform distribution,

$$p(y \mid x_i) = \frac{1}{|\mathcal{Y}|} \quad \text{for all } y, x_i$$

The Maximum-Entropy Solution

- The maximum entropy model is

$$p_* = \mathrm{argmax}_{p \in \mathcal{P}} H(p)$$

- Intuition: find a distribution which

  1. satisfies the constraints
  2. is as smooth as possible

Maximum-Entropy Properties of Log-Linear Models

- We define the set of distributions which can be specified in log-linear form

$$\mathcal{Q} = \{p \ : \ p(y \mid x_i) = \frac{e^{\mathbf{W}\cdot\phi(x_i,y)}}{\sum_{y'\in\mathcal{Y}} e^{\mathbf{W}\cdot\phi(x_i,y')}}, \mathbf{W} \in \mathbb{R}^m\}$$

  here, each p is an $n \times |\mathcal{Y}|$ vector defining $p(y \mid x_i)$ for all i, y.

- Define the negative log-likelihood of the data

$$L(p) = -\sum_i \log p(y_i \mid x_i)$$

- Maximum likelihood solution:

$$q_* = \arg\min_{q\in\bar{\mathcal{Q}}} L(q)$$

  where $\bar{\mathcal{Q}}$ is the *closure* of $\mathcal{Q}$

Duality Theorem

- There is a unique distribution q_* satisfying

  1. $q_* \in$ intersection of P and $\bar{\mathcal{Q}}$
  2. $q_* = \text{argmax}_{p \in \mathcal{P}} H(p)$ (Max-ent solution)
  3. $q_* = \arg\min_{q \in \bar{\mathcal{Q}}} L(q)$ (Max-likelihood solution)

- This implies:

  1. The maximum entropy solution can be written in log-linear form
  2. Finding the maximum-likelihood solution also gives the maximum entropy solution

Developing Intuition Using Lagrange Multipliers

- Max-Ent Problem: Find $\max_{p \in \mathcal{P}} H(p)$

- Equivalent (unconstrained) problem

$$\max_{p \in \Delta} \inf_{\mathbf{W} \in \mathbb{R}^m} L(p, \mathbf{W})$$

  where Δ is the space of all probability distributions, and

$$L(p, \mathbf{W}) = \left(H(p) - \sum_{k=1}^{m} \mathbf{W}_k \left(\sum_i \phi_k(x_i, y_i) - \sum_i \sum_{y \in \mathcal{Y}} \phi_k(x_i, y) p(y \mid x_i) \right) \right)$$

- Why the equivalence?:

$$\inf_{\mathbf{W} \in \mathbb{R}^m} L(p, \mathbf{W}) = \begin{cases} H(p) & \text{if all constraints satisfied, i.e., } p \in \mathcal{P} \\ -\infty & \text{otherwise} \end{cases}$$

Developing Intuition Using Lagrange Multipliers

- We can now switch the min and max:

$$\max_{p \in \mathcal{P}} H(p) = \max_{p \in \Delta} \inf_{\mathbf{W} \in \mathbb{R}^m} L(p, \mathbf{W}) = \inf_{\mathbf{W} \in \mathbb{R}^m} \max_{p \in \Delta} L(p, \mathbf{W}) = \inf_{\mathbf{W} \in \mathbb{R}^m} L(\mathbf{W})$$

- where $L(\mathbf{W}) = \max_{p \in \Delta} L(p, \mathbf{W})$

- By differentiating $L(p, \mathbf{W})$ w.r.t. p, and setting the derivative to zero (making sure to include lagrange multipliers that ensure for all i, $\sum_y p(y \mid x_i) = 1$), and solving

$$p^* = \text{argmax}_{p \in \Delta} L(p, \mathbf{W})$$

gives

$$p^*(y \mid x_i, \mathbf{W}) = \frac{e^{\sum_k \mathbf{W}_k \phi_k(x_i, y)}}{\sum_{y' \in \mathcal{Y}} e^{\sum_k \mathbf{W}_k \phi_k(x_i, y')}}$$

- Also:

$$\begin{aligned} L(\mathbf{W}) = \max_{p \in \Delta} L(p, \mathbf{W}) &= L(p^*(y \mid x_i, \mathbf{W}), \mathbf{W}) \\ &= -\sum_i \log p^*(y \mid x_i, \mathbf{W}) \end{aligned}$$

i.e., the negative log-likelihood under parameters $\mathbf{W}$!

To Summarize

- We've shown that

$$\max_{p \in \mathcal{P}} H(p) = \inf_{\mathbf{W} \in \mathbb{R}^m} L(\mathbf{W})$$

where $L(\mathbf{W})$ is negative log-likelihood

- This argument is pretty informal, as we have to be careful about switching the $\max$ and $\inf$, and we need to relate $\inf_{\mathbf{W} \in \mathbb{R}^m} L(\mathbf{W})$ to finding $q_* = \arg\min_{q \in \bar{\mathcal{Q}}} L(q)$. See [Della Pietra, Della Pietra, and Lafferty 1997] for a proof of the duality theorem.

Is the Maximum-Entropy Property Useful?

- Intuition: find a distribution which
  1. satisfies the constraints
  2. is as smooth as possible

- One problem: the constraints are define by *empirical counts* from the data.

- Another problem: no formal relationship between maximum-entropy property and generalization(?) (at least none is given in the NLP literature)

Overview

- Log-linear models
- The maximum-entropy property
- Smoothing, feature selection etc. in log-linear models

Smoothing in Maximum Entropy Models

- Say we have a feature:

$$\phi_{100}(h,t) \;=\; \begin{cases} 1 & \text{if current word } w_i \text{ is } \texttt{base} \text{ and } t = \texttt{Vt} \\ 0 & \text{otherwise} \end{cases}$$

- In training data, `base` is seen 3 times, with `Vt` every time
- Maximum likelihood solution satisfies

$$\sum_i \phi_{100}(x_i, y_i) = \sum_i \sum_y p(y \mid x_i, \mathbf{W}) \phi_{100}(x_i, y)$$

$\Rightarrow p(\text{Vt} \mid x_i, \mathbf{W}) = 1$ for any history x_i where w_i = `base`
$\Rightarrow \mathbf{W}_{100} \rightarrow \infty$ at maximum-likelihood solution (most likely)
$\Rightarrow p(\text{Vt} \mid x, \mathbf{W}) = 1$ for any test data history x where w = `base`

A Simple Approach: Count Cut-Offs

- [Ratnaparkhi 1998] (PhD thesis): include all features that occur 5 times or more in training data. i.e.,

$$\sum_i \phi_k(x_i, y_i) \geq 5$$

for all features ϕ_k.

Gaussian Priors

- Modified loss function

$$L(\mathbf{W}) = \sum_{i=1}^{n} \mathbf{W} \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, y')} - \sum_{k=1}^{m} \frac{\mathbf{W}_k^2}{2\sigma^2}$$

- Calculating gradients:

$$\left.\frac{dL}{d\mathbf{W}}\right|_{\mathbf{W}} = \underbrace{\sum_{i=1}^{n} \phi(x_i, y_i)}_{\text{Empirical counts}} - \underbrace{\sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} \phi(x_i, y') P(y' \mid x_i, \mathbf{W})}_{\text{Expected counts}} - \frac{1}{\sigma^2} \mathbf{W}$$

- Can run conjugate gradient methods as before
- Adds a penalty for large weights

The Bayesian Justification for Gaussian Priors

- In *Bayesian* methods, combine the log-likelihood $P(data \mid \mathbf{W})$ with a prior over parameters, $P(\mathbf{W})$

$$P(\mathbf{W} \mid data) = \frac{P(data \mid \mathbf{W})P(\mathbf{W})}{\int_{\mathbf{W}} P(data \mid \mathbf{W})P(\mathbf{W})d\mathbf{W}}$$

- The **MAP** (Maximum A-Posteriori) estimates are

$$\begin{aligned} \mathbf{W}_{MAP} &= \text{argmax}_{\mathbf{W}} P(\mathbf{W} \mid data) \\ &= \text{argmax}_{\mathbf{W}} \left(\underbrace{\log P(data \mid \mathbf{W})}_{\text{Log-Likelihood}} + \underbrace{\log P(\mathbf{W})}_{\text{Prior}} \right) \end{aligned}$$

- Gaussian prior: $P(\mathbf{W}) \propto e^{-\sum_k \mathbf{W}_k^2/2\sigma^2}$
$\Rightarrow \log P(\mathbf{W}) = -\sum_k \mathbf{W}_k^2/2\sigma^2 + C$

Experiments with Gaussian Priors

- [Chen and Rosenfeld, 1998]: apply maximum entropy models to language modeling: Estimate $P(w_i \mid w_{i-2}, w_{i-1})$

- Unigram, bigram, trigram features, e.g.,

$$
\begin{aligned}
\phi_1(w_{i-2}, w_{i-1}, w_i) &= \begin{cases} 1 & \text{if trigram is } \texttt{(the,dog,laughs)} \\ 0 & \text{otherwise} \end{cases} \\
\phi_2(w_{i-2}, w_{i-1}, w_i) &= \begin{cases} 1 & \text{if bigram is } \texttt{(dog,laughs)} \\ 0 & \text{otherwise} \end{cases} \\
\phi_3(w_{i-2}, w_{i-1}, w_i) &= \begin{cases} 1 & \text{if unigram is } \texttt{(laughs)} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

$$
P(w_i \mid w_{i-2}, w_{i-1}) = \frac{e^{\sum_k \phi_k(w_{i-2}, w_{i-1}, w_i) \cdot \mathbf{W}}}{\sum_w e^{\sum_k \phi_k(w_{i-2}, w_{i-1}, w) \cdot \mathbf{W}}}
$$

Experiments with Gaussian Priors

- In regular (unsmoothed) maxent, if all n-gram features are included, then it's equivalent to maximum-likelihood estimates!

$$P(w_i \mid w_{i-2}, w_{i-1}) = \frac{Count(w_{i-2}, w_{i-1}, w_i)}{Count(w_{i-2}, w_{i-1})}$$

- [Chen and Rosenfeld, 1998]: with gaussian priors, get very good results. Performs as well as or better than standardly used "discounting methods" such as Kneser-Ney smoothing (see lecture 2).

- Note: their method uses development set to optimize σ parameters

- Downside: computing $\sum_w e^{\sum_k \phi_k(w_{i-2}, w_{i-1}, w) \cdot \mathbf{W}}$ is SLOW.

Feature Selection Methods

- Goal: find *a small number of features* which make good progress in optimizing log-likelihood

- A greedy method:

Step 1 Throughout the algorithm, maintain a set of active features. Initialize this set to be empty.

Step 2 Choose a feature from outside of the set of active features which has largest estimated impact in terms of increasing the log-likelihood and add this to the active feature set.

Step 3 Minimize $L(\mathbf{W})$ with respect to the set of active features. Return to **Step 2**.

Figures from [Ratnaparkhi 1998] (PhD thesis)

- The task: PP attachment ambiguity
- **ME Default:** Count cut-off of 5
- **ME Tuned:** Count cut-offs vary for 4-tuples, 3-tuples, 2-tuples, unigram features
- **ME IFS:** feature selection method

Experiment	Accuracy	Training Time	# of Features
ME Default	82.0%	10 min	4028
ME Tuned	83.7%	10 min	83875
ME IFS	80.5%	30 hours	387
DT Default	72.2%	1 min	
DT Tuned	80.4%	10 min	
DT Binary	-	1 week +	
Baseline	70.4%		

Table 8.2: Maximum Entropy (ME) and Decision Tree (DT) Experiments on PP attachment

Figures from [Ratnaparkhi 1998] (PhD thesis)

- A second task: text classification, identifying articles about acquisitions

Experiment	Accuracy	Training Time	# of Features
ME Default	95.5%	15 min	2350
ME IFS	95.8%	15 hours	356
DT Default	91.6%%	18 hours	
DT Tuned	92.1%	10 hours	

Table 8.4: Text Categorization Performance on the `acq` category

Summary

- Introduced log-linear models as general approach for modeling conditional probabilities $P(y \mid x)$.
- Optimization methods:
  - Iterative scaling
  - Gradient ascent
  - **Conjugate gradient ascent**
- Maximum-entropy properties of log-linear models
- Smoothing methods using Gaussian prior, and feature selection methods

References

[Altun, Tsochantaridis, and Hofmann, 2003] Altun, Y., I. Tsochantaridis, and T. Hofmann. 2003. Hidden Markov Support Vector Machines. In *Proceedings of ICML 2003*.

[Bartlett 1998] P. L. Bartlett. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Transactions on Information Theory, 44(2): 525-536, 1998.

[Bod 98] Bod, R. (1998). *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications/Cambridge University Press.

[Booth and Thompson 73] Booth, T., and Thompson, R. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22(5), pages 442–450.

[Borthwick et. al 98] Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. (1998). Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. *Proc. of the Sixth Workshop on Very Large Corpora*.

[Collins and Duffy 2001] Collins, M. and Duffy, N. (2001). Convolution Kernels for Natural Language. In *Proceedings of NIPS 14*.

[Collins and Duffy 2002] Collins, M. and Duffy, N. (2002). New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *Proceedings of ACL 2002*.

[Collins 2002a] Collins, M. (2002a). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with the Perceptron Algorithm. In *Proceedings of EMNLP 2002*.

[Collins 2002b] Collins, M. (2002b). Parameter Estimation for Statistical Parsing Models: Theory and Practice of Distribution-Free Methods. To appear as a book chapter.

[Crammer and Singer 2001a] Crammer, K., and Singer, Y. 2001a. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. In *Journal of Machine Learning Research*, 2(Dec):265-292.

[Crammer and Singer 2001b] Koby Crammer and Yoram Singer. 2001b. Ultraconservative Online Algorithms for Multiclass Problems In *Proceedings of COLT 2001*.

[Freund and Schapire 99] Freund, Y. and Schapire, R. (1999). Large Margin Classification using the Perceptron Algorithm. In *Machine Learning*, 37(3):277–296.

[Helmbold and Warmuth 95] Helmbold, D., and Warmuth, M. On Weak Learning. *Journal of Computer and System Sciences*, 50(3):551-573, June 1995.

[Hopcroft and Ullman 1979] Hopcroft, J. E., and Ullman, J. D. 1979. *Introduction to automata theory, languages, and computation*. Reading, Mass.: Addison–Wesley.

[Johnson et. al 1999] Johnson, M., Geman, S., Canon, S., Chi, S., & Riezler, S. (1999). Estimators for stochastic 'unification-based" grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann.

[Lafferty et al. 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML-01, pages 282-289, 2001.

[Littlestone and Warmuth, 1986] Littlestone, N., and Warmuth, M. 1986. Relating data compression and learnability. *Technical report, University of California, Santa Cruz*.

[MSM93] Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of english: The Penn treebank. *Computational Linguistics*, *19*, 313-330.

[McCallum et al. 2000] McCallum, A., Freitag, D., and Pereira, F. (2000) Maximum entropy markov models for information extraction and segmentation. In *Proceedings of ICML 2000*.

[Miller et. al 2000] Miller, S., Fox, H., Ramshaw, L., and Weischedel, R. 2000. A Novel Use of Statistical Parsing to Extract Information from Text. In *Proceedings of ANLP 2000*.

[Ramshaw and Marcus 95] Ramshaw, L., and Marcus, M. P. (1995). Text Chunking Using Transformation-Based Learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, Association for Computational Linguistics, 1995.

[Ratnaparkhi 96] A maximum entropy part-of-speech tagger. In *Proceedings of the empirical methods in natural language processing conference*.

[Schapire et al., 1998] Schapire R., Freund Y., Bartlett P. and Lee W. S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651-1686.

[Zhang, 2002] Zhang, T. 2002. Covering Number Bounds of Certain Regularized Linear Function Classes. In *Journal of Machine Learning Research*, 2(Mar):527-550, 2002.