Maximum Entropy

CL1: Jordan Boyd-Graber

University of Maryland

October 21, 2013

Adapted from material by Robert Malouf, Philipp Koehn, and Matthew Leingang

Roadmap

- Why we need more powerful probabilistic modeling formalism
- Example: POS tagging
- Introducing key concepts from information theory
- Maximum Entropy Models
  - Formulation
  - Estimation

Outline

Modeling Distributions

- Modeling Distributions
- Estimating from data
- Thus far, only counting
  - MLE
  - Priors
  - Backoff

Modeling Distributions

- Modeling Distributions
- Estimating from data
- Thus far, only counting
  - MLE
  - Priors
  - Backoff
- What about features?

Supervised Learning

- Problem Setup
  - Given: some annotated data
  - Goal: Build a model
  - Task: Apply it to unseen data
- Issues
  - More data help
  - How to represent the data

Supervised Learning

- Problem Setup
  - Given: some annotated data
  - Goal: Build a model
  - Task: Apply it to unseen data
- Issues
  - More data help
  - How to represent the data
- Part of speech tagging

Supervised Learning

- Problem Setup
  - Given: some annotated data (words with POS tags)
  - Goal: Build a model (using some feature representation)
  - Task: Apply it to unseen data (POS tags for untagged sentences)
- Issues
  - More data help
  - How to represent the data
- Part of speech tagging

Contrast: Hidden Markov Models

- HMMs are useful and simple; three parameters
  - Initial distribution
  - Transition
  - Conditional emission
- Training is easy from tagged data
- Find best sequence using Vitterbi

Contrast: Hidden Markov Models

- HMMs are useful and simple; three parameters
  - Initial distribution
  - Transition
  - Conditional emission
- Training is easy from tagged data
- Find best sequence using Vitterbi
- But it ignores important clues that could help

Motivating Example: Features for POS Tagging

$$\begin{array}{ccccc} w_{n-2} & w_{n-1} & w_n & w_{n+1} & w_{n+2} \\ t_{n-2} & t_{n-1} & t_n & t_{n+1} & t_{n+2} \end{array}$$

- But we can do better

Motivating Example: Features for POS Tagging

$$\begin{array}{ccccc} w_{n-2} & w_{n-1} & w_n & w_{n+1} & w_{n+2} \\ t_{n-2} & t_{n-1} & t_n & t_{n+1} & t_{n+2} \end{array}$$

- But we can do better
- If **one** of the previous tags is `md`, then `vb` is likelier than `vbp` (basic verb form instead of singular)

Motivating Example: Features for POS Tagging

$$\begin{array}{ccccc} w_{n-2} & w_{n-1} & w_n & w_{n+1} & w_{n+2} \\ t_{n-2} & t_{n-1} & t_n & t_{n+1} & t_{n+2} \end{array}$$

- But we can do better
- If **one** of the previous tags is `md`, then `vb` is likelier than `vbp` (basic verb form instead of singular)
- If **next** tag is `jj`, `rbr` is likelier than `jjr` (adverb instead of adjective)

Motivating Example: Features for POS Tagging

$$\begin{array}{ccccc} w_{n-2} & w_{n-1} & w_n & w_{n+1} & w_{n+2} \\ t_{n-2} & t_{n-1} & t_n & t_{n+1} & t_{n+2} \end{array}$$

- But we can do better
- If **one** of the previous tags is `md`, then `vb` is likelier than `vbp` (basic verb form instead of singular)
- If **next** tag is `jj`, `rbr` is likelier than `jjr` (adverb instead of adjective)
- If one of the previous words is "not", the `vb` is likelier than `vbp`

Motivating Example: Features for POS Tagging

$$\begin{array}{ccccc} w_{n-2} & w_{n-1} & \color{red}{w_n} & w_{n+1} & w_{n+2} \\ t_{n-2} & t_{n-1} & t_n & t_{n+1} & t_{n+2} \end{array}$$

- But we can do better
- If **one** of the previous tags is `md`, then `vb` is likelier than `vbp` (basic verb form instead of singular)
- If **next** tag is `jj`, `rbr` is likelier than `jjr` (adverb instead of adjective)
- If one of the previous words is "not", the `vb` is likelier than `vbp`
- If a word ends in "-tion" it is likely a `nn`, but "-ly" implies adverb

Encoding Features

- Much more powerful and expressive than counting **single** observations
  - $\vec{f}(x)$ a vector with the feature count for observation x
  - $f_i(x)$: count of feature i in observation x

Encoding Features

- Much more powerful and expressive than counting **single** observations
  - $\vec{f}(x)$ a vector with the feature count for observation x
  - $f_i(x)$: count of feature i in observation x
- Typical example

$$f(w_{n-2}, w_{n-1}, w_n, w_{n+1}, w_{n+1}, t_n) = \begin{cases} 1, & \text{if } w_{n-1} = \text{``angry'' and } t_n = \texttt{NNP} \\ 0, & \text{otherwise} \end{cases}$$

Where Maximum Entropy Models Fit

- Suppose we have some data-driven information about these features
- What distribution should we use to model these features?
- "Maximum Entropy" models provide a solution

Where Maximum Entropy Models Fit

- Suppose we have some data-driven information about these features
- What distribution should we use to model these features?
- "Maximum Entropy" models provide a solution ... but first we need some definitions

Outline

Expectation

An *expectation* of a random variable is a weighted average:

$$\mathbb{E}\left[f(X)\right] = \sum_{x=1}^{\infty} f(x)\, p(x) \qquad \text{(discrete)}$$

$$= \int_{-\infty}^{\infty} f(x)\, p(x)\, dx \qquad \text{(continuous)}$$

Alternate formulation for positive random variables:

$$\mathbb{E}\left[X\right] = \sum_{x=1}^{\infty} P(X > x) \qquad \text{(discrete)}$$

$$= \int_{0}^{\infty} P(X > x)\, dx \qquad \text{(continuous)}$$

Expectation

Expectations of constants or known values:

- $\mathbb{E}[a] = a$
- $\mathbb{E}[Y \mid Y = y] = y$

Expectation of die / dice

What is the expectation of the roll of die?

Expectation of die / dice

What is the expectation of the roll of die?

One die

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} =$$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

What is the expectation of the sum of two dice?

Expectation of die / dice

What is the expectation of the roll of die?

One die

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

What is the expectation of the sum of two dice?

Two die

$$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} =$$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

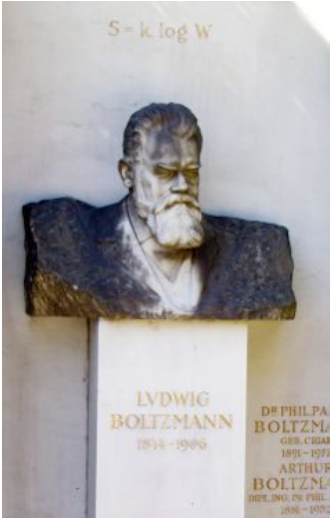
What is the expectation of the sum of two dice?

Two die

$$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$$

Entropy

- Measure of disorder in a system
- In the real world, entroy in a system tends to increase
- Can also be applied to probabilities:
  - Is one (or a few) outcomes certain (low entropy)
  - Are things equiprobable (high entropy)



Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$\begin{aligned} H(X) &= -\mathbb{E}\left[\lg(p(X))\right] \\ &= -\sum_x p(x)\lg(p(x)) && \text{(discrete)} \\ &= -\int_{-\infty}^{\infty} p(x)\lg(p(x))\,dx && \text{(continuous)} \end{aligned}$$

Does not account for the values of the random variable, only the spread of the distribution.

- $H(X) \geq 0$
- uniform distribution = highest entropy, point mass = lowest
- suppose $P(X=1)=p$, $P(X=0)=1-p$ and $P(Y=100)=p$, $P(Y=0)=1-p$: X and Y have the same entropy

Entropy of a die / dice

What is the entropy of a roll of a die?

Entropy of a die / dice

What is the entropy of a roll of a die?

One die

$$-\left(\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)\right)=2.58$$

Entropy of a die / dice

What is the entropy of a roll of a die?

One die

$$-\left(\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)+\frac{1}{6}\lg\left(\frac{1}{6}\right)\right)=2.58$$

What is the entropy of the sum of two die? Tricky question: will it be higher or lower than the first one?

Entropy of a die / dice

What is the entropy of a roll of a die?

One die

$$-\left(\tfrac{1}{6}\lg\left(\tfrac{1}{6}\right)+\tfrac{1}{6}\lg\left(\tfrac{1}{6}\right)+\tfrac{1}{6}\lg\left(\tfrac{1}{6}\right)+\tfrac{1}{6}\lg\left(\tfrac{1}{6}\right)+\tfrac{1}{6}\lg\left(\tfrac{1}{6}\right)+\tfrac{1}{6}\lg\left(\tfrac{1}{6}\right)\right)=2.58$$

What is the entropy of the sum of two die? Tricky question: will it be higher or lower than the first one?

Two die

$$\begin{aligned}-\Bigg(\frac{1}{36}\lg\left(\frac{1}{36}\right) &+ \frac{2}{36}\lg\left(\frac{2}{36}\right)+\frac{3}{36}\lg\left(\frac{3}{36}\right)+\frac{4}{36}\lg\left(\frac{4}{36}\right)+\frac{5}{36}\lg\left(\frac{5}{36}\right)\\ &+\frac{6}{36}\lg\left(\frac{6}{36}\right)+\frac{5}{36}\lg\left(\frac{5}{36}\right)+\frac{4}{36}\lg\left(\frac{4}{36}\right)+\frac{3}{36}\lg\left(\frac{3}{36}\right)\\ &+\frac{2}{36}\lg\left(\frac{2}{36}\right)+\frac{1}{36}\lg\left(\frac{1}{36}\right)\Bigg)=3.27\end{aligned}$$

Principles for Modeling Distributions

Maximum Entropy Principle (Jaynes)

All else being equal, we should prefer distributions that maximize the Entropy

Principles for Modeling Distributions

Maximum Entropy Principle (Jaynes)

All else being equal, we should prefer distributions that maximize the Entropy

- What additional constraints do we want to place on the distribution?
- How, mathematically, do we optimize the entropy?

Outline

The obvious one ...

- We're attempting to model a probability distribution p
- By definition, our probability distribution must sum to one

$$\sum_x p(x) = 1 \tag{1}$$

Feature constraints

- We observe features across many outcomes
- We're modeling a distribution p over observations x. What is the correct model of features under this distribution?
- The whole point of this is that we **don't** want to count outcomes (we've discussed those methods)

Feature constraints

- We observe features across many outcomes
- We're modeling a distribution p over observations x. What is the correct model of features under this distribution?
- The whole point of this is that we **don't** want to count outcomes (we've discussed those methods)
- Ideally, the expected count of the features should be consistent with observations

Estimated Counts

$$\mathbb{E}_p \left[f_i(x) \right] = \sum_x p(x) f_i(x) \tag{2}$$

Empirical Counts

$$\hat{\mathbb{E}}_{\hat{p}} \left[f_i(x) \right] = \hat{p}(x) f_i(x) \tag{3}$$

- Empirical distribution is just what we've observed in data

Optimizing Constrained Functions

Theorem: Lagrange Multiplier Method

Given functions $f(x_1, \ldots x_n)$ and $g(x_1, \ldots x_n)$, the critical points of f restricted to the set $g = 0$ are solutions to equations:

$$\begin{aligned} \frac{\partial f}{\partial x_i}(x_1, \ldots x_n) =& \lambda \frac{\partial g}{\partial x_i}(x_1, \ldots x_n) \quad \forall i \\ g(x_1, \ldots x_n) =& 0 \end{aligned}$$

This is $n + 1$ equations in the $n + 1$ variables $x_1, \ldots x_n, \lambda$.

Lagrange Example

Maximize $f(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

Lagrange Example

Maximize $f(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial f}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \qquad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial f}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \qquad \frac{\partial g}{\partial y} = 10$$

Lagrange Example

Maximize $f(x,y)=\sqrt{xy}$ subject to the constraint $20x+10y=200$.

- Compute derivatives

$$\frac{\partial f}{\partial x}=\frac{1}{2}\sqrt{\frac{y}{x}} \qquad \frac{\partial g}{\partial x}=20$$
$$\frac{\partial f}{\partial y}=\frac{1}{2}\sqrt{\frac{x}{y}} \qquad \frac{\partial g}{\partial y}=10$$

- Create new systems of equations

Lagrange Example

Maximize $f(x,y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial f}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial f}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

- Create new systems of equations

$$\frac{1}{2}\sqrt{\frac{y}{x}} = 20\lambda$$
$$\frac{1}{2}\sqrt{\frac{x}{y}} = 10\lambda$$
$$20x + 10y = 200$$

Lagrange Example

- Dividing the first equation by the second gives us

$$\frac{y}{x} = 2 \tag{4}$$

- which means $y = 2x$, plugging this into the constraint equation gives:

$$20x + 20(2x) = 200$$
$$x = 5 \Rightarrow y = 10$$

Outline

Objective Function

- We want a distribution p that maximizes

$$H(p) \equiv -\sum_x p(x) \log p(x) \tag{5}$$

- Under the constraints that

$$\sum_x p(x) = 1 \tag{6}$$

- and, for every feature f_i

$$\mathbb{E}_p[f_i] = \hat{\mathbb{E}}_{\hat{p}}[f_i]. \tag{7}$$

Augmented Objective Function

$$L(p, \lambda, \gamma) = \\ - \sum_x p(x) \log p(x) \\ - \sum_i \lambda_i \left(\sum_x p(x) f_i(x) - \hat{\mathbb{E}}\left[f_i\right] \right) \\ - \gamma \left(\sum_x p(x) - 1 \right)$$

Plan for solution:

- Take derivative
- Set it equal to zero
- Solve for the $p(x)$ that optimizes equation
- This will give the functional form of our solution

Form of Solution

- Derivation in class
- (Feel free to work out for yourself)

$$p(x) = \frac{\exp\left\{\lambda^\top \vec{f}(x)\right\}}{\sum_{x'} \exp\left\{\lambda^\top \vec{f}(x)\right\}} \tag{8}$$

- Thus, distribution is parameterized by $\vec{\lambda}$ (one for each feature)

Finding Parameters

- Form is simple
- However, finding parameteris is difficult
- Solutions take iterative form
  1. Start with $\vec{\lambda}^{(0)} = \vec{0}$
  2. For $k = 1 \dots$
     1. Determine update $\vec{\delta}^{(k)}$
     2. $\vec{\lambda}^{(k)} \rightarrow \vec{\lambda}^{(k-1)} + \vec{\delta}^{(k)}$

Method for finding updates

- Our objective is a function of $\vec{\lambda}$

$$L(\lambda) = \sum_x \frac{\exp\left\{\lambda^\top f(x)\right\}}{\sum_{x'} \exp\left\{\lambda^\top f(x')\right\}} \tag{9}$$

(in practice, we typically use the log probability)

- Strategy: Move $\vec{\lambda}$ by walking up the gradient $G(\lambda^{(k)})$
- Gradient

$$G_i(\lambda) = \frac{\partial L(\lambda)}{\partial \lambda_i} = -\left[\left(\sum_x p_\lambda(x) f_i(x)\right) - \hat{\mathbb{E}}\left[f_i\right]\right] \tag{10}$$

Method for finding updates

- Set the update of the form

$$\delta^{(k)} = \alpha^{(k)} G(\lambda^{(k)}) \tag{11}$$

- Use the new parameter

$$\vec{\lambda}^{(k)} \rightarrow \vec{\lambda}^{(k-1)} + \vec{\delta}^{(k)} \tag{12}$$

- What value of α?

Method for finding updates

- Set the update of the form

$$\delta^{(k)} = \alpha^{(k)} G(\lambda^{(k)}) \tag{11}$$

- Use the new parameter

$$\vec{\lambda}^{(k)} \rightarrow \vec{\lambda}^{(k-1)} + \vec{\delta}^{(k)} \tag{12}$$

- What value of α?
  - Try lots of different values, pick the one that optimizes $L(\lambda)$ (grid search)

Other parameter estimation techniques

- Iterative scaling
- Conjugate gradient methods
- Real difference is speed and scalability

Regularization / Priors

- We often want to prefer small parameters over large ones, all else being equal

$$L(\lambda) = \sum_x \frac{\exp\left\{\lambda^\top f(x)\right\}}{\sum_{x'} \exp\left\{\lambda^\top f(x')\right\}} - \sum_i \frac{\lambda^2}{\sigma^2} \tag{13}$$

- This is equivalent to having a Gaussian prior on the weights λ
- Also possible to use **informed** priors when you have an idea of what the weights should be (e.g. for domain adaptation)

All sorts of distributions

- We talked about a simple distribution $p(x)$
- But could just as easily be joint distribution $p(y, x)$

$$p(y, x) = \frac{\exp\left\{\lambda^\top f(y, x)\right\}}{\sum_{y', x'} \exp\left\{\lambda^\top f(y', x')\right\}} \tag{14}$$

- Or a conditional distribution $p(y|x)$

$$p(y|x) = \frac{\exp\left\{\lambda^\top f(y, x)\right\}}{\sum_{y'} \exp\left\{\lambda^\top f(y', x)\right\}} \tag{15}$$

Uses of MaxEnt Distributions

- POS Tagging (state of the art)
- Supervised classification: spam vs. not spam
- Parsing (head or not)
- Many other NLP applications

In class . . .

- HW 3 Results
- Quiz
- Deriving MaxEnt formula
- Defining feature functions