Language and Statistics II

Lecture 5: Log-Linear Models
(The Details)
Noah Smith
Today’s Plan

• (Anonymous) pop quiz
• Maximum Entropy modeling
• Relationship to log-linear models
• How to do it!
• Feature selection
• Regularization
• Conditional estimation
Data

How to assign probability to each type?
Maximum Likelihood (Multinomial)

Overfitting?

11 df
Maximum Likelihood Estimation

• Given a model family, pick the parameters to maximize

\[ p(\text{data} \mid \text{model}) \]

• Examples:
  – Gaussian: \( \hat{\mu} = \frac{1}{n} \sum x_i, \hat{\sigma} = \sqrt{\frac{1}{n} \sum (x_i - \hat{\mu})^2} \)
  – Bernoulli: \( \hat{p} = \frac{n_{\text{success}}}{n} \)
  – Multinomial: \( \forall i, \hat{p}_i = \frac{n_i}{n} \)
  – n-gram model?
  – HMM?

\{ \text{closed form solution} \}
Using the Chain Rule

Pr(Color, Shape, Size) = Pr(Color) \cdot Pr(Shape | Color) \cdot Pr(Size | Color, Shape)

These two are the same!

These two are the same!

11 df
Add an Independence Assumption?

Pr(Color, Shape, Size) = Pr(Color) \cdot Pr(Shape) \cdot Pr(Size | Color, Shape)
Pr(Color, Shape, Size) = Pr(Size) \cdot Pr(Shape | Size) \cdot Pr(Color | Size)
Strong Independence?

\[
\Pr(\text{Color}, \text{Shape}, \text{Size}) = \Pr(\text{Size}) \cdot \Pr(\text{Shape}) \cdot \Pr(\text{Color})
\]
This Is Hard!

- Different **factorizations** affect
  - Model size (e.g., number of parameters or df)
  - Complexity of inference
  - “Interpretability”
  - Goodness of fit to the data
  - Generalization
  - Smoothing methods

- How would it change if we used **log-linear** models?

- Arguable: some major “innovations” in NLP involved really good choices about independence assumptions, directionality, and smoothing!
A Log-Linear Shape Model

\[ p(\text{shape}) = \exp \sum_{i} f_i(\text{shape}) \cdot \theta_i / Z(\vec{\theta}) \]

How do we pick the features?
How do we set the weights?

Desideratum: after we pick features, picking the weights should be the computer’s job!
Some Intuitions

• Simpler models are better
  – (E.g., fewer degrees of freedom)
  – Why?

• Want to fit the data

• Don’t want to assume that an unobserved event has probability 0
Occam’s Razor

One should not increase, beyond what is necessary, the number of entities required to explain anything.
### Uniform model

<table>
<thead>
<tr>
<th></th>
<th>△</th>
<th>○</th>
<th>□</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>0.083</td>
<td>0.083</td>
<td>0.083</td>
</tr>
<tr>
<td>small</td>
<td>0.083</td>
<td>0.083</td>
<td>0.083</td>
</tr>
<tr>
<td>large</td>
<td>0.083</td>
<td>0.083</td>
<td>0.083</td>
</tr>
<tr>
<td>large</td>
<td>0.083</td>
<td>0.083</td>
<td>0.083</td>
</tr>
</tbody>
</table>
Constraint: $\Pr(\text{small}) = 0.625$

\[
\begin{array}{c|c|c|c}
& \triangle & \bigcirc & \square \\
\hline
\text{small} & 0.104 & 0.104 & 0.104 \\
\hline
\text{small} & 0.104 & 0.104 & 0.104 \\
\hline
\text{large} & 0.063 & 0.063 & 0.063 \\
\hline
\text{large} & 0.063 & 0.063 & 0.063 \\
\end{array}
\]

Where did the constraint come from?
\[
\Pr(\triangle, \text{ small}) = 0.048
\]

<table>
<thead>
<tr>
<th></th>
<th>(\triangle)</th>
<th>(\bigcirc)</th>
<th>(\square)</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>0.024</td>
<td>0.144</td>
<td>0.144</td>
</tr>
<tr>
<td>small</td>
<td>0.024</td>
<td>0.144</td>
<td>0.144</td>
</tr>
<tr>
<td>large</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
</tr>
<tr>
<td>large</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
</tr>
</tbody>
</table>
$\Pr(\text{large, } \square) = 0.125$

<table>
<thead>
<tr>
<th></th>
<th>$\triangle$</th>
<th>$\bigcirc$</th>
<th>$\square$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>small</strong></td>
<td>0.024</td>
<td>0.144</td>
<td>0.144</td>
</tr>
<tr>
<td><strong>small</strong></td>
<td>0.024</td>
<td>0.144</td>
<td>0.144</td>
</tr>
<tr>
<td><strong>large</strong></td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
</tr>
<tr>
<td><strong>large</strong></td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
</tr>
</tbody>
</table>
Maximum Entropy

\[
\max_p H(p) \equiv \max_p \sum_x -p(x) \log p(x)
\]

subject to

\[
\sum_x p(x) = 1, \quad \forall x, p(x) \geq 0
\]

\[
\forall j \in \{1, 2, \ldots, m\}, \quad E_p \left[ f_j(X) \right] = \alpha_j
\]

\[
\sum_x p(x) f_j(x) = \alpha_j
\]
Questions Worth Asking

• Does a solution always exist?
  – What to do if it doesn’t?
• How to find the solution?
Entropy Review

\[ H(p) = \sum_{x} -p(x)\log p(x) \]

- Measurement on a distribution
- Value in \([0, \log|\mathcal{X}|]\)
- High entropy \(\rightarrow\) uniform
- Low entropy \(\rightarrow\) determinism
- Concave in \(p\)
Maximum Entropy

\[
\max_p H(p) \equiv \max_p \sum_x -p(x) \log p(x)
\]

subject to

\[
\sum_x p(x) = 1, \quad \forall x, p(x) \geq 0
\]

\[
\forall j \in \{1, 2, \ldots, m\}, \quad \mathbb{E}_p \left[ f_j(X) \right] = \alpha_j
\]

\[
\sum_x p(x)f_j(x) = \alpha_j
\]
Marginal Constraints

\[ \sum_x p(x)f_j(x) = \alpha_j \]

\[ \sum_x p(x)f_j(x) = \frac{1}{D} \sum_{i=1}^{D} f_j(\tilde{x}_i) \]

Example:

\[ \sum_x p(x) \begin{cases} 1 \text{ if } x \text{ is square} \\ 0 \text{ otherwise} \end{cases} = \frac{1}{D} \sum_{i=1}^{D} \begin{cases} 1 \text{ if } \tilde{x}_i \text{ is square} \\ 0 \text{ otherwise} \end{cases} = \frac{\text{count(square)}}{D} \]

Let \( \mathcal{P} \) represent the set of distributions \( p \) that meet the constraints.
Claim 1

The unique solution to the maximum entropy problem

$$\arg\max_{p \in \mathcal{P}} H(p)$$

is a log-linear model on the same features as $\mathcal{P}$. 
Claim 2

The unique solution to the maximum entropy problem

\[ \text{arg max}_{p \in \mathcal{P}} H(p) \]

is the log-linear model on the same features as \( \mathcal{P} \) that also solves

\[ \text{arg max}_{p \in \text{Loglinear}} p(\tilde{x}) \]
Max constrained $|X|$ variables ($p$) concave in $p$

unconstrained $m$ variables ($\theta$) concave in $\theta$
Mathematical Magic

For details: see handout on course page.

1. Use Lagrangean multipliers (one per constraint).
2. Take the gradient, set equal to zero.
3. Algebra …
4. Voilà! Maximum likelihood problem!
What if we took out $f_2$?
Additional Point

• If the constraints are empirical, then they are satisfiable (solution exists).

• So there is a unique solution to:
  \[
  \text{Max Ent} = \text{Log-linear MLE}
  \]
Slightly More General View

• Instead of “maximize entropy,” can describe this as “minimize divergence” to a base distribution $q$ (which happens so far to be uniform, but needn’t have been).

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

• Everything goes through pretty much the same.
Training the Weights

• Old answer: “iterative scaling”
  – Specialized method for this problem
  – Later versions: Generalized IS (Darroch and Ratliff, 1972) and Improved IS (Della Pietra, Della Pietra, and Lafferty, 1995)

• More recent answer:
  – It’s unconstrained, convex optimization!
  – See Malouf (2002) for comparison.
Improved Iterative Scaling (Della Pietra et al., 1997)

• Initialize each $\theta_j$ arbitrarily.
• Let: $f_\#(x) = \sum_j f_j(x)$
• Repeat until convergence:
  – Solve for each $\delta_j$: \[
  \sum_x \tilde{p}(x) f_j(x) = \sum_x \frac{\exp f(x) \cdot \bar{\theta}}{Z(\bar{\theta})} f_j(x) e^{\delta_j f_\#(x)}
  \]
  – Update: $\theta_j \leftarrow \theta_j + \delta_j$

Berger’s IIS tutorial gives a derivation.
Gradient Ascent

• Initialize each $\theta_j$ arbitrarily.

• Repeat until convergence:
  – Line search for step size:
    $$\hat{\alpha} \leftarrow \arg\max_{\alpha} f(\bar{\theta} + \alpha \nabla f(\bar{\theta}))$$

  – Gradient step:
    $$\bar{\theta} \leftarrow \bar{\theta} + \hat{\alpha} \nabla f(\bar{\theta})$$
Quasi-Newton Methods

- Use the same information as gradient ascent: function value and gradient.
- Build up an approximate Hessian matrix (second derivatives) over time.
- Converge much faster.
- There are existing implementations: you provide a function that computes $f$ and $\nabla f$.
- (Could use true Hessian, but $n \times n$ second derivatives to compute!)
- Common examples: conjugate gradient, L-BFGS.
What are the Function and Gradient?

\[ L(\theta) = \frac{1}{D} \sum_j \theta_j \sum_{i=1}^D f_j(\tilde{x}_i) - \log \sum_x \exp \sum_j f_j(x) \cdot \theta_j \]

\[ \frac{\partial L}{\partial \theta_j} = \frac{1}{D} \sum_{i=1}^D f_j(\tilde{x}_i) - \mathbb{E}_{p_\theta(x)}[f_j(X)] \]

Should remind you of Max Ent constraints!