Learning Generative Models

October 22, 2012

Generative Models

- A generative model assigns probability jointly to structures and data
- Examples
  - Hidden Markov Models (structure = state sequence, data = observation sequence)
  - PCFGs (structure = tree, data = word observations)
  - Naïve Bayes (“structure” = class, data = word observations)
- Non-examples
  - Conditional random fields
  - Perceptron

Learning Generative Models

$$\mathcal{T} = (\langle \mathbf{x}_1, \mathbf{y}_1 \rangle, \langle \mathbf{x}_2, \mathbf{y}_2 \rangle, \ldots, \langle \mathbf{x}_n, \mathbf{y}_n \rangle)$$

$$p(\mathcal{T}) = \prod_{\langle \mathbf{x}, \mathbf{y} \rangle \in \mathcal{T}} p(\mathbf{x}, \mathbf{y})$$

View 1: MLE

- Find parameters of the model that maximize the likelihood of the training data

$$\begin{aligned}
\boldsymbol{w}^* &= \arg\max_{\boldsymbol{w}} p(\mathcal{T}) \\
&= \arg\max_{\boldsymbol{w}} \prod_{(\mathbf{x},\mathbf{y})\in\mathcal{T}} p(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y};\boldsymbol{w}) \\
&= \arg\max_{\boldsymbol{w}} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}} \log p(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y};\boldsymbol{w})
\end{aligned}$$

View 1: ERM

- The predictor *h* is a probability distribution, and we use the log loss

$$cost(\mathbf{x}, \mathbf{y}, h) = -\log p(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$$

$$p^* = \arg\min_{p \in \mathcal{P}} \underbrace{\frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}} -\log p(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})}_{\text{empirical risk!}}$$

Multinomials and MLE

- Multinomials (or, more properly, Categorical distributions) with N outcomes generalize the notion of a die

- The parameters of a categorical distribution are a N-dimensional vector $\boldsymbol{\theta}$

$$\sum_{i=1}^{N} \theta_i = 1 \qquad\qquad \theta_i \geq 0, \quad \forall i = [1, N]$$

MLE of Multinomials

$$\begin{aligned}
p(T) &= \prod_{x \in T} p(x; \boldsymbol{\theta}) \\
&= \prod_{x \in \mathcal{X}} p(x; \boldsymbol{\theta})^{f(x \in T)} \\
&= \prod_{x \in \mathcal{X}} \theta_x^{f(x \in T)}
\end{aligned}$$

MLE of Multinomials

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_{x \in \mathcal{X}} -f(x \in T) \log \theta_x$$

$$\text{s.t. } \boldsymbol{\theta} > \mathbf{0} \ \wedge \ \sum_{x'} \theta_{x'} = 1$$

*How do we solve this constrained optimization problem?

MLE of Multinomials

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_{x \in \mathcal{X}} -f(x \in T) \log \theta_x$$

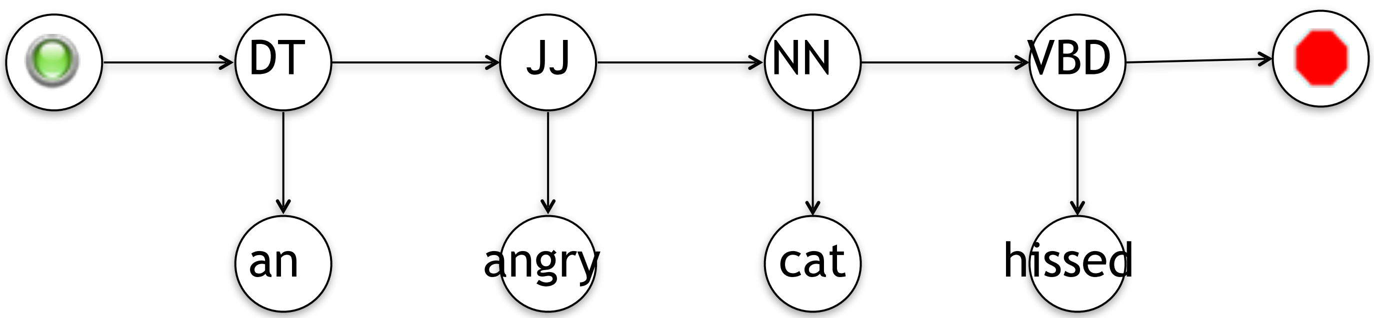
$$\text{s.t. } \boldsymbol{\theta} > \mathbf{0} \;\wedge\; \sum_{x'} \theta_{x'} = 1$$

*How do we solve this constrained optimization problem?

$$\implies \theta_x^* = \frac{f(x \in T)}{|T|}$$

Back to HMMs

- We just have a collection of observations from multinomials!
  - Remember: we are assuming the fully supervised case



MLE for HMMs

- Maximizing values have the following form:

$$p(x \mid y) = \frac{N(x, y)}{N(\cdot, y)}$$

Penalized Maximum Likelihood

- Generally
  - We want good performance on held-out data
  - Zero probabilities are “sampling zeros”
- Solutions
  - “Smoothing”
  - “MAP Estimation”

MAP Estimation of Models

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathcal{T})$$

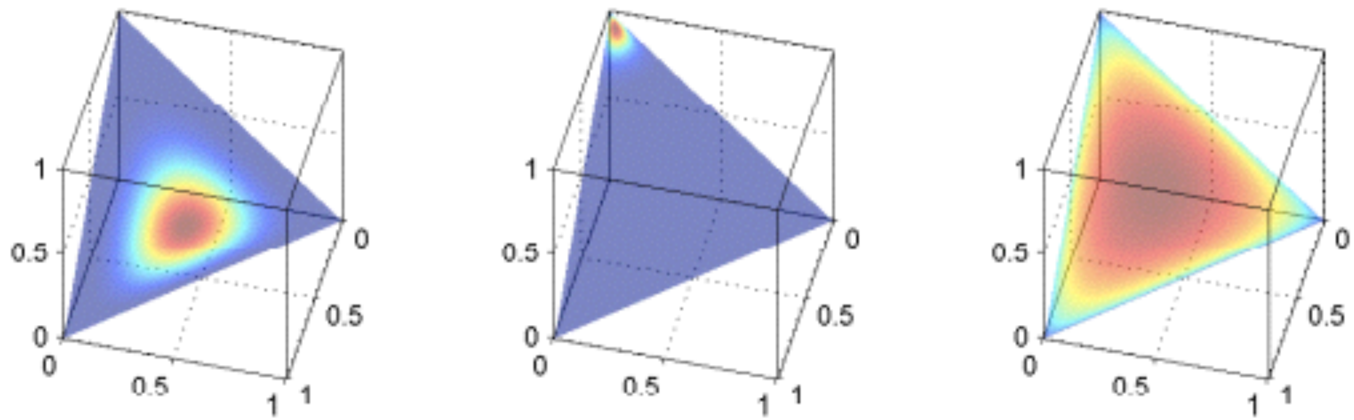
$$\arg\max_{\boldsymbol{\theta}} \frac{p(\mathcal{T} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int d\boldsymbol{\theta}' \;\, p(\mathcal{T} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}$$

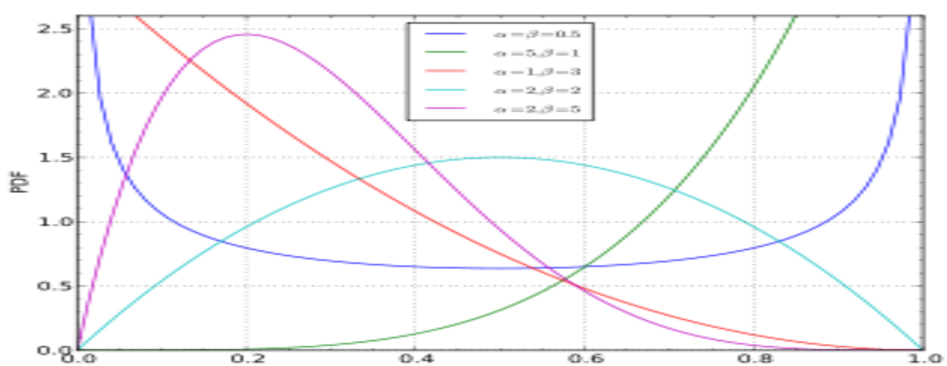
$$\arg\max_{\boldsymbol{\theta}} p(\mathcal{T} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$$

$p(\boldsymbol{\theta})$ encodes prior beliefs about what a good model
will look like. These may be: uniformity of the
distribution ("entropic priors"), sparsity, etc.

Dirichlet & Beta Distributions

- Distributions over multinomial/Bernouilli parameters





Dirichlet/Beta Distributions

- Two parameters, a mean parameter vector $\boldsymbol{\mu}$ and a "concentration" $\alpha > 0$

$$\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\alpha\boldsymbol{\mu})$$

$$p_{\alpha,\boldsymbol{\mu}}(\boldsymbol{\theta}) = \frac{\Gamma(\alpha)}{\prod_{x\in\mathcal{X}}\Gamma(\alpha\mu_x)}\prod_{x\in\mathcal{X}}\theta_x^{\alpha\mu_x-1}$$

MAP Estimation

- Estimation with Dirichlet distributions has the following attractive form when

$$\alpha\mu_x > 1 \quad \forall x \in \mathcal{X}$$

This produces a series of extra "pseudo counts" that are added to the observations

$$\langle \alpha\mu_1 - 1, \alpha\mu_2 - 1, \ldots, \alpha\mu_d - 1 \rangle$$

From this, you show that add-1 smoothing is an instance of MAP inference with a Dir.

MAP Estimation

- This then reduces to:

$$\hat{\theta}_x = \frac{N(x) + \alpha_x - 1}{N(\cdot) + \sum_{x' \in \mathcal{X}} (\alpha_{x'} - 1)}$$

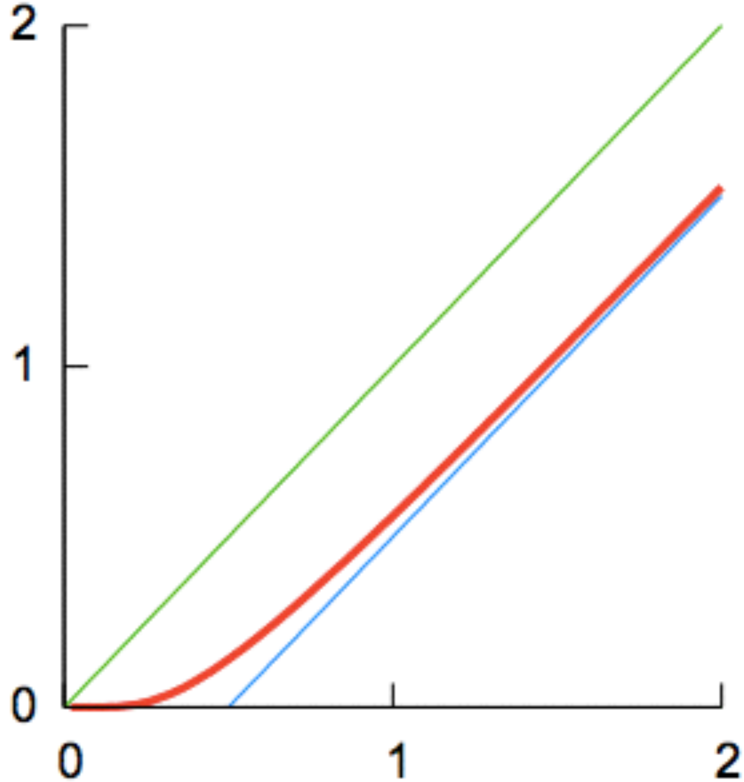
- When does the MAP solution = the MLE solution?

MAP Estimation When $\alpha\mu_x < 1$ '

- When pseudo counts are less than zero, you end up with a sparser (less uniform) solution than the data would warrant
- However, the mode does not have a closed form solution.
  - It may be estimated using Monte Carlo techniques
  - It may be estimated using variational techniques

$$\hat{\theta}_x = \frac{\exp \Psi(N(x) + \alpha_x)}{\exp \Psi(N(\cdot) + \sum_{x' \in \mathcal{X}} (\alpha_{x'} - 1))}$$

Variational Approximation



$$y = \exp \Psi(x)$$

$$y = x - \frac{1}{2}$$

Locally Normalized Log-Linear Models

- Hidden Markov Models

$$p(\text{state } r \mid \text{state } q) = \frac{\boldsymbol{w}^\top \boldsymbol{f}(q, r)}{Z(q)}$$

- PCFGs

$$p(\mathrm{S} \to \mathrm{NP\ VP}) = \frac{\boldsymbol{w}^\top \boldsymbol{f}(\mathrm{S}, \mathrm{NP}, \mathrm{VP})}{Z(\mathrm{S})}$$

Derivation of MLE

- Work out on the board

Globally Normalized Log-Linear Models

- These are not widely used, but it is possible to define a globally normalized generative log-linear model
- These are also called Markov Random Fields or undirected models

$$p(\mathbf{x}, \mathbf{y}) = \frac{\exp \boldsymbol{w}^\top \boldsymbol{F}(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{x}', \mathbf{y}'} \exp \boldsymbol{w}^\top \boldsymbol{F}(\mathbf{x}, \mathbf{y})}$$