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

\[ \mathcal{T} = (\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle) \]

\[ p(\mathcal{T}) = \prod_{\langle x, y \rangle \in \mathcal{T}} p(x, y) \]
View 1: MLE

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

\[ w^* = \arg \max_w p(T) \]

\[ = \arg \max_w \prod_{(x,y) \in T} p(X = x, Y = y; w) \]

\[ = \arg \max_w \sum_{(x,y) \in T} \log p(X = x, Y = y; w) \]
View 1: ERM

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

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

$$p^* = \arg\min_{p \in P} \frac{1}{|\mathcal{T}|} \sum_{(x, y) \in \mathcal{T}} -\log p(X = x, Y = y)$$

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 $\theta$:

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

\[ p(T) = \prod_{x \in T} p(x; \theta) \]

\[ = \prod_{x \in \mathcal{X}} p(x; \theta)^{f(x \in T)} \]

\[ = \prod_{x \in \mathcal{X}} \theta_x^{f(x \in T)} \]
MLE of Multinomials

\[ \theta_{\text{MLE}} = \arg \max_{\theta} \sum_{x \in \mathcal{X}} -f(x \in T) \log \theta_x \]

s.t. \( \theta > 0 \land \sum_{x'} \theta_{x'} = 1 \)

*How do we solve this constrained optimization problem?
MLE of Multinomials

\[ \theta_{\text{MLE}} = \arg \max_{\theta} \sum_{x \in \mathcal{X}} -f(x \in T) \log \theta_x \]

s.t. \( \theta > 0 \land \sum_{x'} \theta_{x'} = 1 \)

*How do we solve this constrained optimization problem?

\[ \Rightarrow \theta^*_x = \frac{f(x \in T)}{|T|} \]
Back to HMMs

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

an angry cat hissed
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

\[ \theta = \arg \max_{\theta} p(\theta \mid \mathcal{T}) \]

\[ \arg \max_{\theta} \frac{p(\mathcal{T} \mid \theta)p(\theta)}{\int d\theta' \ p(\mathcal{T} \mid \theta)p(\theta)} \]

\[ \arg \max_{\theta} p(\mathcal{T} \mid \theta)p(\theta) \]

\( p(\theta) \) 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 $\mu$ and a “concentration” $\alpha > 0$

$$\theta \sim \text{Dirichlet}(\alpha \mu)$$

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

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

\[ y = \exp \Psi(x) \]

\[ y = x - \frac{1}{2} \]
Locally Normalized Log-Linear Models

- Hidden Markov Models

\[ p(\text{state } r \mid \text{state } q) = \frac{\mathbf{w}^\top \mathbf{f}(q, r)}{Z(q)} \]

- PCFGs

\[ p(\text{S } \rightarrow \text{ NP VP}) = \frac{\mathbf{w}^\top \mathbf{f}(\text{S, NP, VP})}{Z(\text{S})} \]
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

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