Lecture 21:

Features and Maximum Entropy Models

Sam Roweis

March 27, 2002

Features

- Consider a subset of random variables $\mathbf{x}_C$ in a graphical model, e.g. three consecutive characters $c_1c_2c_3$ in a string of English text.
- How would we build a model pf $p(c_1c_2c_3)$?
- The full joint clique potential might be huge, e.g. 26^3.
- However, we often know that some particular joint settings of the variables in a clique are quite likely or quite unlikely.
  e.g. `ing, ?ed, qu?, zzz,...`
- For example, we might have $f_{\texttt{ing}}(c_1c_2c_3)$ which is 1 if the string is `'ing'` and 0 otherwise.
- If we use a relatively small number of features, we can combine them into clique potentials using only a reasonable number of parameters but preserving the dependencies that we want.

Constrained Clique Potentials

- So far we have discussed the most general form of a graphical model in which maximal cliques are parametrized by general potential functions $\psi_C(\mathbf{x}_C)$.
- But for large cliques these general potentials are exponentially costly for inference and have exponential numbers of parameters that we must learn from limited data.
- One solution: change the graphical model to make cliques smaller. But this changes the dependencies, and may make more independence assumptions than we would like.
- Another solution: use *feature-based models*.

Combining Features

- How can we combine these features into a probability model?
- Using a generalized exponential family distribution:

$$p(c_1c_2c_3) \propto \exp\{ \quad \theta_{\texttt{ing}} f_{\texttt{ing}}(c_1c_2c_3) + \theta_{\texttt{?ed}} f_{\texttt{?ed}}(c_1c_2c_3) + \theta_{\texttt{qu?}} f_{\texttt{qu?}}(c_1c_2c_3) + \theta_{\texttt{zzz}} f_{\texttt{zzz}}(c_1c_2c_3) + \ldots\}$$

- Each feature has a weight which tells us how important it is and whether it increases or decreases the probability of the clique.
- In general, the features may be overlapping, unconstrained indicators of any function of the clique variables:

$$\psi_c(\mathbf{x}_c) \equiv \prod_{i \in I_C} \exp\{\theta_i f_i(\mathbf{x}_{Ci})\} = \exp\left\{ \sum_{i \in I_C} \theta_i f_i(\mathbf{x}_{Ci}) \right\}$$

Feature Based Model

- We can multiply these clique potentials as usual:

$$\begin{aligned} p(\mathbf{x}|\theta) &= \frac{1}{Z(\theta)} \prod_C \psi_C(\mathbf{x}_C) \\ &= \frac{1}{Z(\theta)} \prod_C \exp\left\{ \sum_{i \in I_C} \theta_i f_i(\mathbf{x}_{Ci}) \right\} \\ &= \frac{1}{Z(\theta)} \exp\left\{ \sum_C \sum_{i \in I_C} \theta_i f_i(\mathbf{x}_{Ci}) \right\} \end{aligned}$$

- However, in general we can forget about associating features with cliques and just use a simplified form:

$$p(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}_{Ci}) \right\}$$

- This is just our friend the exponential model, with the features as sufficient statistics!

Maximum Entropy

- We can approach the modeling problem from an entirely different point of view. *Begin* with some fixed feature expectations:

$$\sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) = \alpha_i$$

- Assuming expectations are consistent, there exist many distributions which satisfy them. Which one should we select? The most uncertain or flexible one: i.e. the one with maximum entropy.
- This yields a new optimization problem:

$$\begin{aligned} \max\ \mathcal{H}[p(\mathbf{x})] &= -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \\ \text{subject to } \sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) &= \alpha_i \\ \sum_{\mathbf{x}} p(\mathbf{x}) &= 1 \end{aligned}$$

Review: Maximum Likelihood

$$\begin{aligned} \ell(\theta; \mathcal{D}) &= \sum_{\mathbf{x}} n(\mathbf{x}) \log p(\mathbf{x}|\theta) \\ &= \sum_{\mathbf{x}} n(\mathbf{x}) \left(\sum_i \theta_i f_i(\mathbf{x}) - \log Z(\theta) \right) \\ &= \sum_{\mathbf{x}} n(\mathbf{x}) \sum_i \theta_i f_i(\mathbf{x}) - N \log Z(\theta) \\ \frac{\partial \ell}{\partial \theta_i} &= \sum_{\mathbf{x}} n(\mathbf{x}) f_i(\mathbf{x}) - N \frac{\partial}{\partial \theta_i} \log Z(\theta) \\ &= \sum_{\mathbf{x}} n(\mathbf{x}) f_i(\mathbf{x}) - N \sum_{\mathbf{x}} p(\mathbf{x}|\theta) f_i(\mathbf{x}) \\ \Rightarrow \quad \sum_{\mathbf{x}} p(\mathbf{x}|\theta) f_i(\mathbf{x}) &= \sum_{\mathbf{x}} \frac{n(\mathbf{x})}{N} f_i(\mathbf{x}) = \sum_{\mathbf{x}} \bar{p}(\mathbf{x}) f_i(\mathbf{x}) \end{aligned}$$

Derivative of log partition function is the expectation of the feature.
At ML estimate, model expectations match empirical feature counts.

Solution to the MaxEnt Problem

- To solve the maxent problem, we use Lagrange multipliers:

$$\begin{aligned} L &= -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_i \theta_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) - \alpha_i \right) - \mu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right) \\ \frac{\partial L}{\partial p(\mathbf{x})} &= 1 + \log p(\mathbf{x}) - \sum_i \theta_i f_i(\mathbf{x}) - \mu \\ p^*(\mathbf{x}) &= e^{\mu - 1} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \\ Z(\theta) &= e^{1-\mu} = \sum_{\mathbf{x}} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \\ p(\mathbf{x}|\theta) &= \frac{1}{Z(\theta)} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \end{aligned}$$

- So feature constraints + maxent implies exponential family.
- Problem is convex, so solution is unique.

Constraints from Data

- Where do the constraints α_i come from?
- Just as before, measure the empirical counts on the training data:

$$\alpha_i = \sum_{\mathbf{x}} \frac{n(\mathbf{x})}{N} f_i(\mathbf{x}) = \sum_{\mathbf{x}} \bar{p}(\mathbf{x}) f_i(\mathbf{x})$$

- This also ensures consistency automatically.
- Known as the "method of moments".
- We have seen a case of *convex duality*:
  In one case, we assume exponential family and show that ML implies feature expectations match observed counts.
  In the other case, we assume model expectations must match empirical feature counts and show that maxent implies exponential family distribution.

Geometric Interpretation

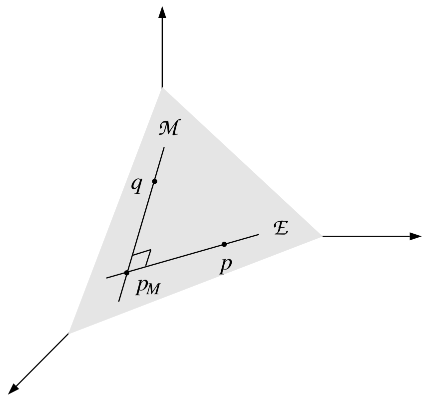
Define two submanifolds on the probability simplex $p(\mathbf{x})$.

The first is $\mathcal{E}$, the set of all exponential family distributions based on a particular set of features $f_i(\mathbf{x})$.

The second is $\mathcal{M}$, the set of all distributions that satisfy the feature expectation constraints.

They intersect at a single distribution p_M, the maxent, maximum likelihood distribution.

A generalized Pythagorean theorem holds for cross entropies.



$$\mathcal{E} = \left\{ p(\mathbf{x}) \ : \ \frac{1}{Z(\theta)} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \right\}$$

$$\mathcal{M} = \left\{ p(\mathbf{x}) \ : \ \sum_{\mathbf{x}} p(\mathbf{x}|\theta) f_i(\mathbf{x}) = \sum_{\mathbf{x}} \frac{n(\mathbf{x})}{N} f_i(\mathbf{x}) = \sum_{\mathbf{x}} \bar{p}(\mathbf{x}) f_i(\mathbf{x}) \right\}$$

$\mathsf{KL}[q\|p] = \mathsf{KL}[q\|p_M] + \mathsf{KL}[\,p_M\|p\,]$