

CS281A/Stat241A recitation 7: iterative proportional fitting

October 15, 2007

Percy Liang

Multinomials

- This section addresses some of the questions that were asked about multinomial distributions after recitation. Multinomial distributions are very important, as they are the building blocks for all distributions over discrete data.
- Definitions
 - The standard definition of a multinomial distribution is a distribution over the number of occurrences of each of K possible outcomes, $\{1, \dots, K\}$. A multinomial draw X is a K -dimensional vector, where X_k is the number of occurrences of outcome k in a total of $N = \sum_{k=1}^K X_k$ trials. N is usually assumed to be a constant.
 - A very important special case is when the number of trials is 1. Suppose the outcome of that one trial is k . In this case, X which is 1 in component k and 0 in all other components. In other words, $X_j = \delta(j, k)$. A directed graphical model is simply a set of multinomial draws, one for each node, each of which involves exactly one trial.
 - When the number of trials is 1, a convenient way to think about a multinomial random variable X is that it takes on values in $\{1, \dots, K\}$. So we could write $p(X = k)$ for the probability of a multinomial draw (dice roll) taking on value $k \in \{1, \dots, K\}$. This is the natural notation for talking about the values of nodes in graphical models, but for parameter estimation, the vectorial representation is more convenient.
- Representation
 - There are three ways to parameterize a multinomial distribution which have various advantages and disadvantages.
 - Constrained:

$$p(X = k) = \theta_k \quad [\text{Parameters: } \theta = (\theta_1, \dots, \theta_K) \text{ such that } \theta_k \geq 0, \sum_{k=1}^K \theta_k = 1]$$

This corresponds to a curved exponential family. Advantage: the most straightforward. Disadvantages: if we want to find the maximum likelihood estimate, we need Lagrange multipliers to enforce the sum-to-one constraint. The general optimality condition for the MLE for an exponential family doesn't hold: $\mathbb{E}_\eta T(X) \neq \nabla A(\eta)$.

- Unconstrained minimal:

$$p(X = k) = \begin{cases} \theta_k & 1 \leq k < K \\ 1 - \sum_{k'=1}^{K-1} \theta_{k'} & k = K \end{cases} \quad [\text{Parameters: } \theta = (\theta_1, \dots, \theta_{K-1}) \text{ such that } \theta_k \geq 0]$$

Minimal means that there is a one-to-one mapping between parameters θ and distributions over X . Advantage: the maximum likelihood estimate is unique. Disadvantage: the algebra is messier and less intuitive due to the asymmetry of outcome K .

- Unconstrained overcomplete:

$$p(X = k) = \frac{\theta_k}{\sum_{k'=1}^K \theta_{k'}} \quad [\text{Parameters: } \theta = (\theta_1, \dots, \theta_K) \text{ such that } \theta_k \geq 0]$$

Written as an exponential family (switching to the vectorial representation):

$$p(X) = \exp\left\{\sum_{k=1}^K X_k \log \theta_k - \log \sum_{k=1}^K \theta_k\right\}$$

Overcomplete means that the parameterization is not minimal, that there exists two settings of the parameters that yield the same distribution over X . Advantage: algebra is simple; get that

$$p(X = k) = \mathbb{E}_\eta T_k(X) = \frac{\partial}{\partial \eta_k} A(\eta) = \frac{\theta_k}{\sum_{k'=1}^K \theta_{k'}}.$$

Disadvantage: MLE is not unique since we can scale all the θ_k s by a constant without affecting the distribution.

- Useful algebra trick

- Suppose we observe n i.i.d. multinomial draws X_1, \dots, X_n . The likelihood is

$$p(X_1, \dots, X_n) = \theta_{X_1} \cdots \theta_{X_n}.$$

- But for purposes of parameter estimation, it is more convenient to organize the factors by the outcome k . So the likelihood can also be written as

$$p(X_1, \dots, X_n) = \prod_{k=1}^K \theta_k^{\sum_{i=1}^n \delta(X_i, k)}.$$

It is easy to check that the two are equivalent. This can be generalized and becomes increasingly useful when we have many multinomial draws.

Introduction

- We eventually want to do maximum likelihood estimation for undirected graphical models using the iterative proportional fitting algorithm. But let us present estimation in the context of exponential families, which is a powerful framework that gives us many results without having to work out the specific cases.
- Recall definition of exponential family:

$$p(x; \eta) = \exp\{\eta^T T(x) - A(\eta)\}$$

If we have n i.i.d. data points, then

$$p(x_1, \dots, x_n; \eta) = \exp \left\{ n \left(\eta^T \underbrace{\left(\frac{1}{n} \sum_{i=1}^n T(x_i) \right)}_{T(x_1, \dots, x_n)} - A(\eta) \right) \right\}$$

Note that the sufficient statistics of all the data points is just an average of the individual sufficient statistics.

- Our goal: maximum likelihood estimation

$$\eta^* = \max_{\eta} p(x_1, \dots, x_n; \eta)$$

Taking the derivative and setting it to zero, we see that η^* satisfies a moment-matching property, that the expected sufficient statistics $\mathbb{E}_{\eta^*} T(X)$ matches the empirical sufficient statistics $\hat{\mathbb{E}} T(X)$:

$$\mu^* \stackrel{\text{def}}{=} \mathbb{E}_{\eta^*} T(X) = \hat{\mathbb{E}} T(X) = \frac{1}{n} \sum_{i=1}^n T(X_i)$$

Technically, we also need to verify that the Hessian is positive negative definite, which follows from the fact that $\nabla^2 A(\eta) = \text{cov}_{\eta} T(X)$ and covariance matrices are always positive semi-definite. In general, it is non-trivial to find η^* that satisfies this criterion, but in certain cases, it is possible. A general approach to learning is gradient ascent:

$$\eta^{(t+1)} \leftarrow \eta^{(t)} + \rho (T(x_1, \dots, x_n) - \mathbb{E}_{\eta} T(X))$$

Unfortunately, even just $\mathbb{E}_{\eta} T(X)$ is hard to compute for general graphical models.

- There are two opposite problems, inference and learning.
 - Inference: map from parameters η to expected sufficient statistics $\mathbb{E}_{\eta} T(X)$

- Learning (parameter estimation): map empirical sufficient statistics $\hat{\mathbb{E}}T(X)$ to parameters η^*
- Examples: Gaussian, multinomial, directed graphical models
- Undirected graphical models (today)
 - * Junction tree, belief propagation: potentials \rightarrow marginals
 - * Iterative proportional fitting: marginals \rightarrow potentials
- Examples of exponential families $p(x \mid \eta)$:
 - Simple exponential family (Gaussian, multinomial, etc.)
 - * Moment matching: $\mu(\eta^*) = \frac{1}{n} \sum_{i=1}^n x_i$ where μ is the mean of the Gaussian
 - Directed graphical model

$$p(\tilde{x}; \eta) = \prod_v p(x_v \mid x_{\pi(v)}; \eta) = \prod_v \frac{\theta(x_v, x_{\pi(v)}; \eta)}{\sum_{x'_v} \theta(x'_v, x_{\pi(v)}; \eta)}$$

We can write it in exponential family form:

$$p(\tilde{x}; \eta) = \exp \left\{ \sum_{v, x_v, x_{\pi(v)}} \delta(x_v, \tilde{x}_v) \delta(x_{\pi(v)}, \tilde{x}_{\pi(v)}) \log \theta(x_v \mid x_{\pi(v)}; \eta) - A(\eta) \right\},$$

where

$$A(\eta) = \sum_v \sum_{x_{\pi(v)}} \log \sum_{x_v} \theta(x_v, x_{\pi(v)}; \eta).$$

- * Sufficient statistics:

$$T(x_1, \dots, x_n) = \{\tilde{p}(x_v, x_{\pi(v)})\},$$

where

$$\tilde{p}(x_v, x_{\pi(v)}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \delta(x_v, \tilde{x}_{iv}) \delta(x_{\pi(v)}, \tilde{x}_{i\pi(v)})$$

is the fraction of times the particular configuration shows up in the data.

- * Moment matching property says at the MLE η^* , we must have:

$$\frac{\theta(x_v, x_{\pi(v)}; \eta^*)}{\sum_{x'_v} \theta(x'_v, x_{\pi(v)}; \eta^*)} = p(x_v, x_{\pi(v)}; \eta^*) = \tilde{p}(x_v, x_{\pi(v)}),$$

namely, that the marginals must agree. This implies that the conditionals must agree.

$$p(x_v \mid x_{\pi(v)}; \eta^*) = \tilde{p}(x_v \mid x_{\pi(v)})$$

We're in luck because the conditional probabilities are exactly the parameters of an directed graphical model.

- Undirected graphical model

$$p(\tilde{x}; \eta) = \frac{1}{Z(\eta)} \prod_C \psi_C(\tilde{x}_C; \eta) = \exp \left\{ \sum_{C, x_C} \delta(x_C, \tilde{x}_C) \log \psi_C(x_C; \eta) - \log Z(\eta) \right\}$$

- * Sufficient statistics:

$$T(x_1, \dots, x_n) = \{\tilde{p}(x_C)\}$$

where

$$\tilde{p}(x_C) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \delta(x_C, \tilde{x}_{iC}).$$

- * The moment matching property tells us that at the MLE η , we must have

$$p(x_C; \eta^*) = \tilde{p}(x_C)$$

- * Problem: unlike the directed case, there is not a simple mapping between the potentials $\psi_C(x_C; \eta^*)$ and the marginals $p(x_C; \eta^*)$. We know what the marginals at the optimum should be, but to convert them into potentials, we need the the IPF algorithm.

- Side remarks (didn't talk about in recitation)

- Tied parameters: when several potentials $\psi(x_C)$ or conditional probabilities $p(x_v | p_{\pi(v)})$ share the same underlying parameters (as in problem 4 on homework 3)
 - * Doesn't complicated learning in directed graphical models
 - * Does complicate learning in undirected graphical models; for this, IPF doesn't work, so we need an alternative like iterative scaling (chapter 20)

So let's work with non-tied graphical models (we have separate parameters for each clique).

- Decomposable undirected graphical models (triangulated graphs)
 - * Given a set of marginals, we can set the potentials to match those marginals (a generalized version of problem 5 on homework 1).

- General undirected graphical models

- Finally, we present the IPF algorithm. At each iteration, we pick a clique C and update its potentials:

$$\psi_C^{(t+1)}(x_C) \leftarrow \psi_C^{(t)}(x_C) \frac{\tilde{p}(x_C)}{p^{(t)}(x_C)}$$

We need to compute $p^{(t)}(x_C)$ using junction tree, which requires inference (could be expensive for densely connected graphical models).

– Probabilistic interpretation

- * Claim: $p(x_{V \setminus C} \mid x_C)$ doesn't change ($p^{(t)}(x_{V \setminus C}) = p^{(t+1)}(x_{V \setminus C})$)
 - Only $\psi_C(x_C)$ is different between iterations t and $t + 1$.
 - Since we condition on x_C , any factor that only depends on x_C is treated as a constant and is sucked into the normalization constant.
- * The following derivation will give us two important properties:

$$\frac{p^{(t+1)}(x_C)}{p^{(t)}(x_C)} = \frac{p^{(t+1)}(x)}{p^{(t)}(x)} = \frac{\frac{1}{Z^{(t+1)}} \prod_D \psi_D^{(t+1)}(x_D)}{\frac{1}{Z^{(t)}} \prod_D \psi_D^{(t)}(x_D)} = \frac{Z^{(t)}}{Z^{(t+1)}} \frac{\psi_C^{(t+1)}(x_C)}{\psi_C^{(t)}(x_C)}$$

Substituting the IPF update and canceling some terms yields

$$p^{(t+1)}(x_C) = \frac{Z^{(t)}}{Z^{(t+1)}} \tilde{p}(x_C).$$

- * Property 1: normalization constant doesn't change ($Z^{(t)} = Z^{(t+1)}$)
 - To see this, sum both sides over x_C and note that both $p^{(t+1)}(x_C)$ and $\tilde{p}(x_C)$ are distributions.
 - This is an important property because it makes the IPF algorithm efficient.
- * Property 2: we have $p^{(t+1)}(x_C) = \tilde{p}(x_C)$
 - Recall that this is one of the optimality conditions for the MLE η^* , so we are making progress.
 - However, it does not suffice just to make one pass through the cliques, because changing $\psi_C(x_C)$ affects all the marginals (another the reason why there is not a simple relationship between potentials and marginals), which causes other optimality conditions to be violated again.

– Optimization interpretation: coordinate-wise ascent

- * From general exponential families, we know that

$$\frac{\partial p(x_1, \dots, x_n)}{\partial \eta_C} = \underbrace{\tilde{p}(x_C)}_{T_C(x_1, \dots, x_n)} - \underbrace{p(x_C)}_{\mathbb{E}_\eta T_C(X)},$$

where we use η_C and T_C to denote the components that correspond to clique C .

- * Property 2 tells us that the gradient is 0. Therefore, the IPF update finds the $\psi_C(x_C)$ that maximizes the log-likelihood (we would also need to verify that the log-likelihood is concave).
- * We could also do gradient ascent, but that has two disadvantages:
 - Need to choose step size ρ
 - Z changes changes, unlike IPF