CS281A/Stat241A: Statistical Learning Theory

Parameter Estimation

Lecturer: Michael I. Jordan *Scribes: Ravi Kolluri and Mukund Seshadri*

1 Parameter Estimation in the Completely Observed Case

In the first half of this lecture, we will prove that Iterative Proportional Fitting (IPF) is coordinate ascent in the log likelihood.

Consider an undirected graphical model parameterized by a set of clique potentials $\psi_C(X_C)$ for $C \in \mathcal{C}$ where $\mathcal{C}$ is a set of cliques. [1] The joint probability is then defined as:

$$p(x|\theta) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where $\theta = \{\psi_C(x_C), C \in \mathcal{C}\}$ is the collection of parameters, and where Z is the normalization factor.

Iterative Proportional Fitting (IPF) in this model is characterized by the following update equation (the superscripts indicate the step of the iteration):

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

The maximum likelihood estimate (MLE) is characterized by the following two conditions (as applied to the IPF model):

$$p^{(t+1)}(x_C) = \tilde{p}(x_C),$$

for fixed C, and

$$Z^{(t+1)} = Z^{(t)}$$

A coordinate ascent algorithm is characterized by maximization of a function $f(X_1, X_2, .., X_m)$ on the coordinates $(X_1, X_2, .., X_m)$, by iterating in the following manner through some order of the index i (the superscript indicates the step of the iteration):

[1] It is not necessary that this set is exhaustive or that the cliques are maximal.

$$X^{(t+1)} = (X_1^{(t)}, X_2^{(t)}, .., X_i^{(*)}, .., X_m^{(t)})$$

where

$$X_i^{(*)} = argmax_{X_i} f(X_1^{(t)}, X_2^{(t)}, .., X_i, .., X_m^{(t)})$$

To derive IPF as a coordinate ascent algorithm, we consider the derivative of the log likelihood w.r.t. the coordinate (which in our case is $\psi_C(x_C)$). In each iteration, we fix C and, for varying x_C, we solve for maximizing values of the the parameters governing $\psi_C(x_C)$, while holding the remaining potentials fixed. Since the log likelihood is convex, the coordinate ascent method will yield the global maximum.

We now utilize a decomposition obtained from the KL divergence view.

Lemma 1. *If $A \cup B = V$ represents a partition of all components of our coordinate, and writing $p(x_A, x_B) = p(x_B|x_A)p(x_A)$ and $q(x_A, x_B) = q(x_B|x_A)q(x_A)$, then:*

$$D(p(x_A, x_B)||q(x_A, x_B)) = D(p(x_A)||q(x_A)) + \sum_{x_A} p(x_A) D(p(x_B|x_A)||q(x_B|x_A))$$

Recall that the problem of maximum likelihood estimation is equivalent to the minimization of the KL divergence $D(\tilde{p}(x)||p(x|\theta))$, which in turn is defined as:

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

Therefore we will perform coordinate descent in $D(\tilde{p}(x)||p(x|\theta))$ by choosing a clique C and setting $\psi_C(x_C)$ so as to minimize the KL divergence. Using Lemma 1, we get:

$$D(\tilde{p}(x)||p(x|\theta)) = D(\tilde{p}(x_C)||p(x_C|\theta)) + \sum_{x_C} \tilde{p}(x_C) D(\tilde{p}(x_{\nu \setminus C}|x_C)||p(x_{\nu \setminus C}|x_C, \theta))$$

Note that changes to $\psi_C(x_C)$ have no effect on $p(x_{\nu \setminus C}|x_C, \theta)$. We assume the truth of this claim for now, for the sake of continuity, and verify it shortly. Also note that $\tilde{p}(x_{\nu \setminus C}|x_C)$ is independent of $\psi_C(x_C)$. Thus minimization of the LHS reduces to minimization of the first term of the RHS: $D(\tilde{p}(x_C)||p(x_C|\theta))$. This is accomplished when $p(x_C|\theta) = \tilde{p}(x_C)$. But this is exactly what is achieved by the IPF update. Thus we have proved that IPF is coordinate descent in the KL divergence, which in turn is coordinate ascent in the log likelihood.

All that remains is to prove the following claim:

Claim 1. *$p(x_{\nu \setminus C}|x_C, \theta)$ is independent of $\psi_C(x_C)$ for a given $C \in \mathcal{C}$, where $\mathcal{C}$ is the set of cliques.*

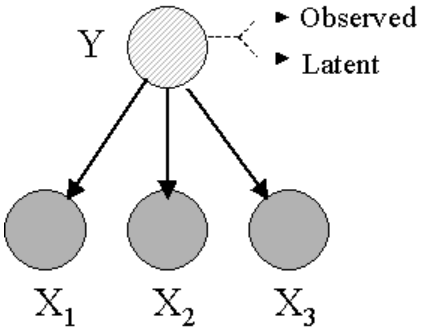

Figure 1: Parameter Estimation with Latent Variable Y

Proof.

$$\begin{aligned} p(x_{\nu\setminus C}|x_C,\theta) &= \frac{p(x_{\nu\setminus C},x_C,\theta)}{p(x_C)} \\ &= \frac{\frac{1}{Z}\prod_{F\in\mathcal{C}}\psi_F(x_F)}{\sum_{x_{\nu\setminus C}}\frac{1}{Z}\prod_{D\in\mathcal{C}}\psi_D(x_D)} \\ &= \frac{\prod_{F\in\mathcal{C}}\psi_F(x_F)}{\psi_C(x_C)\sum_{x_{\nu\setminus C}}\prod_{D\in\mathcal{C},D\neq C}\psi_D(x_D)} \\ &= \frac{\prod_{F\in\mathcal{C},F\neq C}\psi_F(x_F)}{\sum_{x_{\nu\setminus C}}\prod_{D\in\mathcal{C},D\neq C}\psi_D(x_D)} \end{aligned}$$

This expression is independent of $\psi_C(x_C)$. Thus our claim is proved. □

1.1 Latent Variables

When some of the variables are unobserved, we have latent variables, and the parameter estimation problem does not decouple as conveniently as in the completely observed case.

Consider the graphical model shown in Figure 1. The nodes X_i are conditionally independent given Y. When Y is observed, the log likelihood decouples conveniently:

$$\log p(x,y|\theta) = \log p(y|\theta_y) + \sum_{i=1..3}\log p(x_i|y,\theta_i)$$

However, when Y is latent, then the log likelihood does not exhibit the same decoupling:

$$\log p(x,y|\theta) = \log \sum_y [p(y|\theta_y)\prod_{i=1..3} p(x_i|y,\theta_i)]$$

As an example of a problem with latent variables consider finite mixture models. The graphical model corresponding to the finite mixture models is shown in Figure 2. Let Z represent a multinomial random variable with components Z^i. We have:

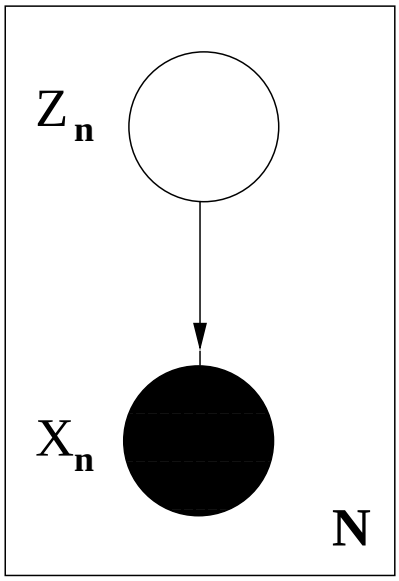

Figure 2: The mixture model.

$$\begin{aligned} p(x|\theta) &= \sum_i p(Z^i = 1|\pi_i) p(x|Z^i = 1, \theta_i) \\ &= \sum_{i=1}^{k} \pi_i p(x|\theta_i) \end{aligned}$$

where the parameters are $\theta = (\pi_i, \pi_2 \ldots \pi_K, \theta_1, \theta_2, \ldots \theta_K)$ and the π_i are constrained to sum to one. The π_i are called the mixing proportions and the densities $p(x|Z^i = 1, \theta_i)$ are the mixture components. We will now look at the special case of Gaussian mixture models, where the mixture components are Gaussians: $P(x|Z^i = 1, \theta_i) = N(x, \mu_i, \Sigma_i)$. The problem is to estimate the parameters(means,covariance matrix) from an IID set of observations $calD = \{x_n : n = 1, 2, \ldots, N\}$. Note that the data is unlabeled.

To solve these latent variable problems, we will (in the next lecture) learn the Expectation Maximization(EM) algorithm. As a warm up let us look at the K-means algorithm. Our goal here is to group the data into K clusters, for a given integer K. In practice, estimating K is considered a model selection problem.

The algorithm maintains two kinds of variables, μ_i, which are the estimates of the means of our clusters, and vectors z_n, which represent assignments. The i^{th} component of z_n is set to one when x_n has been assigned to cluster i.

The K-means algorithm first begins by guessing some initial values for the means μ_i. The algorithm consists of two phases: In the first phase the indicator variables are set by assigning each data x_n to the closest mean μ_i.

$$z_n^i = \begin{cases} 1 & if\ i = arg\ min_j ||x_n - \mu_j||^2 \\ 0 & otherwise \end{cases}$$

In the second phase, the values of the mean μ_i is computed by taking the average of all x_n that have been assigned to the i^{th} cluster.

$$\mu_i = \frac{\sum_n z_n^i x_n}{\sum_n z_n^i}$$

These two phases are repeated until there is no change in the assignments which also means that the values of μ_i have converged. It can be shown that K-means is coordinate descent for a distortion measure given by:

$$J = \sum_{n=1}^{K} \sum_{i=i}^{K} z_n^i ||x_n - \mu_i||^2$$