LECTURE 13: MIDTERM REVIEW

STAT 545: Intro. to Computational Statistics

Vinayak Rao
Purdue University

October 18, 2016

Point estimation for exponential family models

Exponential family distribution:

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))$$

$$\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_D(x)] : \quad \text{(feature) vector of sufficient statistics}$$

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D] : \quad \text{vector of natural parameters}$$

Point estimation for exponential family models

Exponential family distribution:

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))$$

$$\boldsymbol{\phi}(x) = [\phi_1(x), \dots, \phi_D(x)] : \quad \text{(feature) vector of sufficient statistics}$$

$$\boldsymbol{\theta} = [\theta_1, \dots, \theta_D] : \quad \text{vector of natural parameters}$$

Maximum likelihood estimation is Moment matching.

Given data $X = \{x_1, \dots, x_N\}$, set $\boldsymbol{\theta}$ so that:

$$\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\phi}(x_i) = \mathbb{E}_\theta[\phi(x)] := \boldsymbol{\mu} \quad \text{(Moment parameters)}$$

Point estimation for exponential family models

Exponential family distribution:

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))$$

$$\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_D(x)] : \qquad \text{(feature) vector of sufficient statistics}$$

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D] : \qquad \text{vector of natural parameters}$$

Maximum likelihood estimation is Moment matching.

Given data $X = \{x_1, \ldots, x_N\}$, set $\boldsymbol{\theta}$ so that:

$$\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\phi}(x_i) = \mathbb{E}_\theta[\phi(x)] := \boldsymbol{\mu} \quad \text{(Moment parameters)}$$

Clean, analytic solution.

Often mapping from moment to natural parameters is easy.

Joint probability:

$$p(x, y|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x, y) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

Joint probability:

$$p(x, y|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x, y) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

Given data $X = \{x_1, \ldots, x_N\}$, we want $\boldsymbol{\theta}_{MLE}$:

$$p(x|\boldsymbol{\theta}) = \int_{\mathcal{Y}} \frac{1}{Z(\theta)} h(x, y) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y)) \mathrm{d}y$$

Joint probability:

$$p(x, y|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x, y) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

Given data $X = \{x_1, \ldots, x_N\}$, we want $\boldsymbol{\theta}_{MLE}$:

$$p(x|\boldsymbol{\theta}) = \int_{\mathcal{Y}} \frac{1}{Z(\theta)} h(x, y) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y)) \mathrm{d}y$$

This marginal probability is NOT exp. family. Need iterative algorithms.

If we knew Y, match moments to

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i, y_i)$$

If we knew Y, match moments to

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i, y_i)$$

EM algorithm:

- Initialize with arbitrary $\boldsymbol{\theta}_0$.
- Repeat for $i = 1$ till convergence:
  - Calculate $q(Y) = P(Y|X, \boldsymbol{\theta}_i)$

Point est. for exp. fam. models w. missing data

If we knew Y, match moments to

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i, y_i)$$

EM algorithm:

- Initialize with arbitrary $\boldsymbol{\theta}_0$.
- Repeat for $i = 1$ till convergence:
  - Calculate $q(Y) = P(Y|X, \boldsymbol{\theta}_i)$
  - Calculate $\boldsymbol{\theta}_{i+1}$ by matching moments to

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_q[\phi(x_i, y_i)]$$

Point est. for exp. fam. models w. missing data

If we knew Y, match moments to

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i, y_i)$$

EM algorithm:

- Initialize with arbitrary $\boldsymbol{\theta}_0$.
- Repeat for $i = 1$ till convergence:
  - Calculate $q(Y) = P(Y|X, \boldsymbol{\theta}_i)$
  - Calculate $\boldsymbol{\theta}_{i+1}$ by matching moments to

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_q[\phi(x_i, y_i)]$$

If matching moments for the first equation is easy, so is for the second.

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x)) = h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x) - \log Z(\theta))$$

We have prior info. about $\boldsymbol{\theta}$ in the form of a prob. distribution.

We are interested not just in a point estimate, but the entire posterior distribution (mean, mode, variance etc).

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)}h(x)\exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x)) = h(x)\exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x) - \log Z(\theta))$$

We have prior info. about $\boldsymbol{\theta}$ in the form of a prob. distribution. We are interested not just in a point estimate, but the entire posterior distribution (mean, mode, variance etc).

If the prior is conjugate

$$p(\boldsymbol{\theta}|\mathsf{a}, b) \propto \exp(\boldsymbol{\theta}^\top \mathsf{a} - \log Z(\theta)b)$$

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x)) = h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x) - \log Z(\theta))$$

We have prior info. about $\boldsymbol{\theta}$ in the form of a prob. distribution. We are interested not just in a point estimate, but the entire posterior distribution (mean, mode, variance etc).

If the prior is conjugate

$$p(\boldsymbol{\theta}|\mathbf{a}, b) \propto \exp(\boldsymbol{\theta}^\top \mathbf{a} - \log Z(\theta) b)$$

$$p(\boldsymbol{\theta}|X, \mathbf{a}, b) \propto \exp(\boldsymbol{\theta}^\top (\mathbf{a} + \sum_{i=1}^{N} \boldsymbol{\phi}(x_i)) - \log Z(\theta)(b + N))$$

$$= P(\boldsymbol{\theta}|\mathbf{a} + \sum_{i=1}^{N} \boldsymbol{\phi}(x_i), b + N) \quad \text{(Same family as prior)}$$