LECTURE 11: REVIEW

STAT 545: Intro. to Computational Statistics

Vinayak Rao

Purdue University

September 30, 2019

Point estimation for exponential family models

Exponential family distribution:

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

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

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

Point estimation for exponential family models

Exponential family distribution:

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

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

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

Maximum likelihood estimation is Moment matching.

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

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

Point estimation for exponential family models

Exponential family distribution:

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

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

$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]$: 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:

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

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

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

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

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 $q(Y) = P(Y|X, \boldsymbol{\theta}_i)$

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

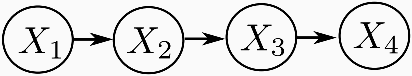
$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i, y_i)$$

EM algorithm:

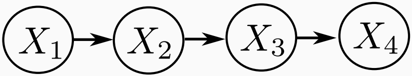
- Initialize with arbitrary $\boldsymbol{\theta}_0$.
- Repeat for $i = 1$ till convergence:
  - Calculate $q(Y) = P(Y|X, \boldsymbol{\theta}_i)$
  - Calculate $\boldsymbol{\theta}_{i+1}$ by matching moments to

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

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



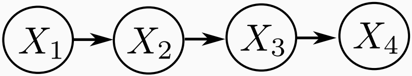
Consider an N-state Markov chain: $X_1 \sim \pi$, $p(X_{t+1}|X_t) = A$.



Consider an N-state Markov chain: $X_1 \sim \pi$, $p(X_{t+1}|X_t) = A$.

More precisely:

$$p(X_1) = \prod_{i=1}^{N} \pi_i^{\delta(X_1=i)} = \exp(\sum_{i=1}^{N} \delta(X_1 = i) \log \pi_i)$$



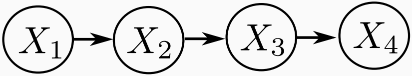
Consider an N-state Markov chain: $X_1 \sim \pi$, $p(X_{t+1}|X_t) = A$.

More precisely:

$$p(X_1) = \prod_{i=1}^{N} \pi_i^{\delta(X_1=i)} = \exp(\sum_{i=1}^{N} \delta(X_1 = i) \log \pi_i)$$

$$p(X_{t+1}|X_t) = \prod_{i=1}^{N}\prod_{j=1}^{N} A_{ij}^{\delta(X_t=i)\delta(X_{t+1}=j)}$$
$$= \exp(\sum_{i=1}^{N}\sum_{j=1}^{N} \delta(X_t = i, X_{t+1} = j) \log A_{ij})$$

HMMs and exp-fam distributions

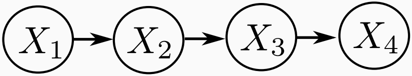


Consider an N-state Markov chain: $X_1 \sim \pi$, $p(X_{t+1}|X_t) = A$.

$$P(X_2, \ldots, X_T|X_1) = \prod_{t=1}^{T} p(X_{t+1}|X_t)$$

$$= \exp\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T} \delta(X_t = i, X_{t+1} = j) \log A_{ij}\right)$$

HMMs and exp-fam distributions



Consider an N-state Markov chain: $X_1 \sim \pi$, $p(X_{t+1}|X_t) = A$.

$$
\begin{aligned}
P(X_2, \ldots, X_T|X_1) &= \prod_{t=1}^{T} p(X_{t+1}|X_t) \\
&= \exp\Big(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T} \delta(X_t = i, X_{t+1} = j) \log A_{ij}\Big) \\
&= \exp\Big(\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i\to j}(X) \log A_{ij}\Big)
\end{aligned}
$$

HMMs and exp-fam distributions

$$p(Y_t|X_t) = \prod_{i=1}^{N} (\text{Poiss}(Y_t|\lambda_i))^{\delta(X_t=i)} = \prod_{i=1}^{N} \left(\frac{\lambda_i^{Y_t} \exp(-\lambda_i)}{Y_t!} \right)^{\delta(X_t=i)}$$

HMMs and exp-fam distributions

$$p(Y_t|X_t) = \prod_{i=1}^{N} (\mathrm{Poiss}(Y_t|\lambda_i))^{\delta(X_t=i)} = \prod_{i=1}^{N} \left(\frac{\lambda_i^{Y_t} \exp(-\lambda_i)}{Y_t!} \right)^{\delta(X_t=i)}$$

$$\propto \exp(\sum_{i=1}^{N} (\delta(X_t = i) Y_t) \log \lambda_i - \delta(X_t = i)\lambda_i)$$

HMMs and exp-fam distributions

$$p(Y_t|X_t) = \prod_{i=1}^{N} (\text{Poiss}(Y_t|\lambda_i))^{\delta(X_t=i)} = \prod_{i=1}^{N} \left(\frac{\lambda_i^{Y_t} \exp(-\lambda_i)}{Y_t!} \right)^{\delta(X_t=i)}$$

$$\propto \exp(\sum_{i=1}^{N} (\delta(X_t = i) Y_t) \log \lambda_i - \delta(X_t = i) \lambda_i)$$

$$p(Y|X) \propto \exp(\sum_{i=1}^{N} \sum_{t=1}^{T} (\delta(X_t = i) Y_t) \log \lambda_i - \delta(X_t = i) \lambda_i)$$

$$p(Y_t|X_t) = \prod_{i=1}^{N} (\text{Poiss}(Y_t|\lambda_i))^{\delta(X_t=i)} = \prod_{i=1}^{N} \left(\frac{\lambda_i^{Y_t} \exp(-\lambda_i)}{Y_t!} \right)^{\delta(X_t=i)}$$

$$\propto \exp(\sum_{i=1}^{N} (\delta(X_t = i)Y_t) \log \lambda_i - \delta(X_t = i)\lambda_i)$$

$$p(Y|X) \propto \exp(\sum_{i=1}^{N} \sum_{t=1}^{T} (\delta(X_t = i)Y_t) \log \lambda_i - \delta(X_t = i)\lambda_i)$$

$$\propto \exp(\sum_{i=1}^{N} M_i \log \lambda_i - C_i \lambda_i)$$

($M_i = C_i V_i$, where M_i is total value of Y when in state i, C_i = #-times in state i, V_i = avg value of Y when in state i)

HMMs and the EM algorithm

$$C_i = \sum_{t=1}^{T} \delta(X_t = i), \quad C_{i \to j} = \sum_{t=1}^{T} \delta(X_t = i, X_{t+1} = j)$$

$$B_i = \delta(X_1 = i), \qquad M_i = \sum_{t=1}^{T} (\delta(X_t = i) Y_t),$$

HMMs and the EM algorithm

$$C_i = \sum_{t=1}^{T} \delta(X_t = i), \quad C_{i \to j} = \sum_{t=1}^{T} \delta(X_t = i, X_{t+1} = j)$$

$$B_i = \delta(X_1 = i), \qquad M_i = \sum_{t=1}^{T} (\delta(X_t = i) Y_t),$$

For an HMM, we don't observe X. Use EM instead:

- Given current parameters $\boldsymbol{\theta}$, calculate $q(X) = p(X|Y, \boldsymbol{\theta})$

HMMs and the EM algorithm

$$C_i = \sum_{t=1}^{T} \delta(X_t = i), \quad C_{i\to j} = \sum_{t=1}^{T} \delta(X_t = i, X_{t+1} = j)$$

$$B_i = \delta(X_1 = i), \qquad M_i = \sum_{t=1}^{T} (\delta(X_t = i) Y_t),$$

For an HMM, we don't observe X. Use EM instead:

- Given current parameters $\boldsymbol{\theta}$, calculate $q(X) = p(X|Y, \boldsymbol{\theta})$
- Calculate expected sufficient statistics under q:

$$\mathbb{E}_q[C_i] = \sum_{t=1}^{T} p(X_t = i|Y, \boldsymbol{\theta}), \quad \mathbb{E}_q[C_{i\to j}] = \sum_{t=1}^{T} p(X_t = i, X_{t+1} = j|Y, \boldsymbol{\theta})$$

$$\mathbb{E}_q[B_i] = p(X_1 = i|Y, \boldsymbol{\theta}), \qquad \mathbb{E}_q[M_i] = \sum_{t=1}^{T} (p(X_t = i|Y, \boldsymbol{\theta}) Y_t)$$

$$C_i = \sum_{t=1}^{T} \delta(X_t = i), \quad C_{i \to j} = \sum_{t=1}^{T} \delta(X_t = i, X_{t+1} = j)$$

$$B_i = \delta(X_1 = i), \qquad M_i = \sum_{t=1}^{T} (\delta(X_t = i) Y_t),$$

For an HMM, we don't observe X. Use EM instead:

- Given current parameters $\boldsymbol{\theta}$, calculate $q(X) = p(X|Y, \boldsymbol{\theta})$
- Calculate expected sufficient statistics under q:

$$\mathbb{E}_q[C_i] = \sum_{t=1}^{T} p(X_t = i|Y, \boldsymbol{\theta}), \quad \mathbb{E}_q[C_{i \to j}] = \sum_{t=1}^{T} p(X_t = i, X_{t+1} = j|Y, \boldsymbol{\theta})$$

$$\mathbb{E}_q[B_i] = p(X_1 = i|Y, \boldsymbol{\theta}), \qquad \mathbb{E}_q[M_i] = \sum_{t=1}^{T} (p(X_t = i|Y, \boldsymbol{\theta}) Y_t)$$

- Match moments using expected suff. stats.

Run Baum-Welch each E-step.

Probability of a D-dim binary vector X_i under μ_k:

$$p(X_i|\mu_k) = \prod_{d=1}^{D} \exp(\delta(x_{id} = 1) \log \frac{\mu_{kd}}{1 - \mu_{kd}}) \quad (\log \frac{\mu_{kd}}{1 - \mu_{kd}} := \eta_{kd})$$

EM algorithm for mixture of Bernoulli vectors

Probability of a D-dim binary vector X_i under μ_k:

$$p(X_i|\mu_k) = \prod_{d=1}^{D} \exp(\delta(x_{id} = 1) \log \frac{\mu_{kd}}{1-\mu_{kd}}) \quad (\log \frac{\mu_{kd}}{1-\mu_{kd}} := \eta_{kd})$$

$$p(X_i, c_i|\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \prod_{d=1}^{D} (\pi_k p(X_i|\mu_k))^{\delta(c_i=k)}$$

EM algorithm for mixture of Bernoulli vectors

Probability of a D-dim binary vector X_i under μ_k:

$$p(X_i|\mu_k) = \prod_{d=1}^{D} \exp(\delta(x_{id} = 1) \log \frac{\mu_{kd}}{1 - \mu_{kd}}) \quad (\log \frac{\mu_{kd}}{1 - \mu_{kd}} := \eta_{kd})$$

$$p(X_i, c_i|\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \prod_{d=1}^{D} (\pi_k p(X_i|\mu_k))^{\delta(c_i=k)}$$

$$\log p(X_i, c_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \delta(c_i = k) \log(\pi_k) + \sum_{k=1}^{K} \sum_{d=1}^{D} \delta(c_i = k)\delta(x_{id} = 1)\eta_{kd} + C$$

EM algorithm for mixture of Bernoulli vectors

Probability of a D-dim binary vector X_i under μ_k:

$$p(X_i|\mu_k) = \prod_{d=1}^{D} \exp(\delta(x_{id} = 1) \log \frac{\mu_{kd}}{1-\mu_{kd}}) \quad (\log \frac{\mu_{kd}}{1-\mu_{kd}} := \eta_{kd})$$

$$p(X_i, c_i|\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \prod_{d=1}^{D} (\pi_k p(X_i|\mu_k))^{\delta(c_i=k)}$$

$$\log p(X_i, c_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \delta(c_i = k) \log(\pi_k) + \sum_{k=1}^{K} \sum_{d=1}^{D} \delta(c_i = k)\delta(x_{id} = 1)\eta_{kd} + C$$

Given N observations, MLE is moment matching:

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \delta(c_i = k), \quad \mu_{ik} = \frac{\sum_{i=1}^{N} \delta(c_i = k)\delta(x_{id} = 1)}{\sum_{i=1}^{N} \delta(c_i = k)}$$

EM algorithm for mixture of Bernoulli vectors

We don't know c_i's, but can calculate posterior $p(C|X, \boldsymbol{\theta})$. EM algorithm

EM algorithm for mixture of Bernoulli vectors

We don't know c_i's, but can calculate posterior $p(C|X,\boldsymbol{\theta})$. EM algorithm

- Set $q(C) = p(C|X,\boldsymbol{\theta})$ and define

$$\mathcal{F}_X(q,\boldsymbol{\theta}) = \mathbb{E}_q[\log p(X,C|\boldsymbol{\theta}]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} q(c_i = k)\log(\pi_k) + \sum_{k=1}^{K}\sum_{d=1}^{D} q(c_i = k)\delta(x_{id} = 1)\eta_{kd} + C(\boldsymbol{\theta})$$

Note: $C(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, and must be included to get a nondecreasing lower-bound.

EM algorithm for mixture of Bernoulli vectors

We don't know c_i's, but can calculate posterior $p(C|X, \boldsymbol{\theta})$. EM algorithm

- Set $q(C) = p(C|X, \boldsymbol{\theta})$ and define

$$\mathcal{F}_X(q, \boldsymbol{\theta}) = \mathbb{E}_q[\log p(X, C|\boldsymbol{\theta}]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} q(c_i = k) \log(\pi_k) + \sum_{k=1}^{K} \sum_{d=1}^{D} q(c_i = k)\delta(x_{id} = 1)\eta_{kd} + C(\boldsymbol{\theta})$$

  Note: $C(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, and must be included to get a nondecreasing lower-bound.

- M-step:

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} q(c_i = k), \quad \mu_{ik} = \frac{\sum_{i=1}^{N} q(c_i = k)\delta(x_{id} = 1)}{\sum_{i=1}^{N} q(c_i = k)}$$