Lecture 9: Latent Variable Models and EM Algorithm

Jacob Steinhardt

February 17, 2020

Last Time

Bayesian Inference

- Setup
- Conjugate priors
- Computing posteriors
- Inference
  - Full posterior
  - MAP, LMSE

This time: more complex models, fast algorithm (EM)

Recall: Heights and Gender

[Jupyter demo]

Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$

Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$
- $x_i \mid z_i \sim N(\mu_{z_i}, \sigma^2)$, i.e. $p(x_i \mid z_i) \propto \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)$

Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$
- $x_i \mid z_i \sim N(\mu_{z_i}, \sigma^2)$, i.e. $p(x_i \mid z_i) \propto \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)$
- $p(z_i) = \pi^{z_i}(1-\pi)^{1-z_i}$ (Bernoulli with probability π)

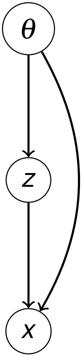
Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$
- $x_i \mid z_i \sim N(\mu_{z_i}, \sigma^2)$, i.e. $p(x_i \mid z_i) \propto \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)$
- $p(z_i) = \pi^{z_i}(1-\pi)^{1-z_i}$ (Bernoulli with probability π)

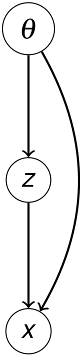
"Hyperparameters": $\mu_0, \mu_1, \sigma^2, \pi$

[draw graphical model]

Latent Variable Model: General Form



Latent Variable Model: General Form

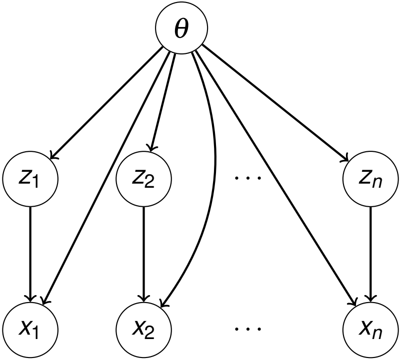


hyperparameters $(\mu_1, \mu_2, \sigma, \pi)$

latent structure (gender)

observed output (height)

Special Case: Hierarchical Model



"Bayesian hierarchical model"

Another Example: HMMs

Hidden Markov model

- Fish population, changing over time
  - $z_1, \ldots, z_T$; z_t: population at time t

Another Example: HMMs

Hidden Markov model

- Fish population, changing over time
  - $z_1, \ldots, z_T$; z_t: population at time t
- At each time t, randomly sample various locations in pond and count number of fish
  - Measurements: $x_1, \ldots, x_T$

Another Example: HMMs

Hidden Markov model

- Fish population, changing over time
  - $z_1, \dots, z_T$; z_t: population at time t
- At each time t, randomly sample various locations in pond and count number of fish
  - Measurements: $x_1, \dots, x_T$
- Model:

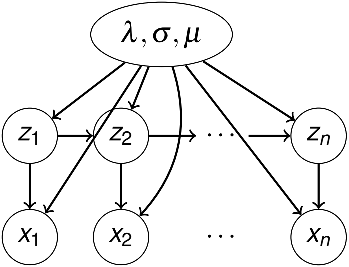
$$x_t \sim \text{Poisson}(\lambda z_t), \quad z_{t+1} \sim N(z_t, \sigma^2), \quad z_0 \sim N(\mu, \sigma^2)$$

Another Example: HMMs

Hidden Markov model

- Fish population, changing over time
  - $z_1, \dots, z_T$; z_t: population at time t
- At each time t, randomly sample various locations in pond and count number of fish
  - Measurements: $x_1, \dots, x_T$
- Model:

$$x_t \sim \text{Poisson}(\lambda z_t), \quad z_{t+1} \sim N(z_t, \sigma^2), \quad z_0 \sim N(\mu, \sigma^2)$$



Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state
- Sample true fraction of Clinton supporters for each state, look at how often Clinton wins

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state
- Sample true fraction of Clinton supporters for each state, look at how often Clinton wins

Something like this predicted 90% Clinton in 2016, but Trump won.

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state
- Sample true fraction of Clinton supporters for each state, look at how often Clinton wins

Something like this predicted 90% Clinton in 2016, but Trump won.

What is wrong with this analysis? [At least 2 things...]

Election Forecasting Model

[on board]

Election Forecasting Model

[on board]

Next: EM algorithm

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$
    - "half-Bayesian"

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$
    - “half-Bayesian”

How many terms in sum?

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$
  - "half-Bayesian"

How many terms in sum?

- 100 people, genders $z_1, \ldots, z_{100}$
- $2^{100} \approx 10^{30}$ possibilities

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$
  - “half-Bayesian”

How many terms in sum?

- 100 people, genders $z_1, \ldots, z_{100}$
- $2^{100} \approx 10^{30}$ possibilities

Need a better strategy!

Warm-up: Gaussian example

[on board: θ from z and z from θ]

Alternating Maximization

General observation (not just Gaussians):

- If z known, $\mathrm{argmax}_\theta \log p(x, z \mid \theta)$ often easy
- If θ known, computing $p(z \mid x, \theta)$ often easy

Alternating Maximization

General observation (not just Gaussians):

- If z known, $\text{argmax}_\theta \log p(x, z \mid \theta)$ often easy
- If θ known, computing $p(z \mid x, \theta)$ often easy

Idea: alternate between updating θ and updating z, repeat until convergence

EM Algorithm

- Alternates between updating two variables, θ and q
- $q(z)$: matches $p(z \mid \theta, x)$
- θ: optimizes $\underbrace{\mathbb{E}_{z\sim q(z)}[\log p(z, x \mid \theta)]}_{\text{average over } z \text{ drawn from } q}$

EM Algorithm

- Alternates between updating two variables, θ and q
- $q(z)$: matches $p(z \mid \theta, x)$
- θ: optimizes $\underbrace{\mathbb{E}_{z \sim q(z)}[\log p(z, x \mid \theta)]}_{\text{average over } z \text{ drawn from } q}$

Formally: initialize $\theta^{(1)}$ arbitrarily. Then for $t = 1, \ldots, T$:

$$q^{(t)}(z) \leftarrow p(z \mid x, \theta^{(t)}) \qquad \text{(E step)}$$
$$\theta^{(t+1)} \leftarrow \operatorname{argmax}_\theta \mathbb{E}_{z \sim q(z)}[\log p(z, x \mid \theta^{(t)})] \qquad \text{(M step)}$$

EM Algorithm

- Alternates between updating two variables, θ and q
- $q(z)$: matches $p(z \mid \theta, x)$
- θ: optimizes $\underbrace{\mathbb{E}_{z \sim q(z)}[\log p(z, x \mid \theta)]}_{\text{average over } z \text{ drawn from } q}$

Formally: initialize $\theta^{(1)}$ arbitrarily. Then for $t = 1, \ldots, T$:

$$q^{(t)}(z) \leftarrow p(z \mid x, \theta^{(t)}) \qquad \text{(E step)}$$

$$\theta^{(t+1)} \leftarrow \operatorname{argmax}_\theta \mathbb{E}_{z \sim q(z)}[\log p(z, x \mid \theta^{(t)})] \qquad \text{(M step)}$$

Can interpret as maximizing lower bound on $\log p(x \mid \theta)$.

EM Algorithm: Gaussian example

[on board]

Recap

- Many problems have unobserved structure / dependencies (hierarchical models, hidden Markov models, ...)
- Failing to model these can lead to wrong/overconfident predictions (election forecasting)
- Latent variables $\implies$ exponential sum $\implies$ need good algorithms!
- EM algorithm: works when we can handle z and θ individually.