Lecture 10: Approximate Inference via Sampling

Jacob Steinhardt

February 19, 2020

Last Time

- Latent variable models
  - Bayesian hierarchical model (heights and gender)
  - Hidden Markov model (counting fish in a pond)
  - Election forecasting model
- EM algorithm

This time: finish EM, start on sampling algorithms

EM Algorithm

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

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

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

Gaussian example

$$p(x_1, z_1, \ldots \mid \pi, \mu_0, \mu_1, \sigma) = \prod_{i=1}^{n} p(z_i \mid \pi) p(x_i \mid z_i, \mu_0, \mu_1, \sigma)$$

$$= \prod_{i=1}^{n} \left[\pi^{z_i}(1-\pi)^{1-z_i}\right] \cdot \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2)\right]$$

Gaussian example

$$p(x_1, z_1, \ldots \mid \pi, \mu_0, \mu_1, \sigma) = \prod_{i=1}^{n} p(z_i \mid \pi) p(x_i \mid z_i, \mu_0, \mu_1, \sigma)$$

$$= \prod_{i=1}^{n} \left[\pi^{z_i}(1-\pi)^{1-z_i}\right] \cdot \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2)\right]$$

Want to maximize likelihood. Take log:

$$\log p(x_1, z_1, \ldots, x_n, z_n \mid \pi, \mu_0, \mu_1, \sigma) = \sum_{i=1}^{n} \underbrace{z_i \log(\pi) + (1-z_i)\log(1-\pi)}_{\text{log-likelihood of Bernoulli}}$$

$$+ \underbrace{\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2}_{\text{log-likelihood of (two) Gaussians}}$$

Gaussian example

$$p(x_1, z_1, \dots \mid \pi, \mu_0, \mu_1, \sigma) = \prod_{i=1}^{n} p(z_i \mid \pi) p(x_i \mid z_i, \mu_0, \mu_1, \sigma)$$

$$= \prod_{i=1}^{n} \left[\pi^{z_i}(1-\pi)^{1-z_i}\right] \cdot \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)\right]$$

Want to maximize likelihood. Take log:

$$\log p(x_1, z_1, \dots, x_n, z_n \mid \pi, \mu_0, \mu_1, \sigma) = \sum_{i=1}^{n} \underbrace{z_i \log(\pi) + (1 - z_i) \log(1 - \pi)}_{\text{log-likelihood of Bernoulli}}$$

$$+ \underbrace{\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2}_{\text{log-likelihood of (two) Gaussians}}$$

Note this is same as maximizing $\mathbb{E}[\log p(x, z \mid \pi, \mu_0, \mu_1, \sigma)]$, where expectation is over uniform distribution on x_i, z_i.

Maximizing Likelihood for Exponential Families

The following result is helpful for quickly computing MLE solutions without algebra:

Maximizing Likelihood for Exponential Families

The following result is helpful for quickly computing MLE solutions without algebra:

Theorem. Suppose that $p(x \mid \theta)$ is an "exponential family with sufficient statistics $g(x)$". Then for any $q(x)$, the solution to $\operatorname{argmax}_\theta \mathbb{E}_{x \sim q}[\log p(x \mid \theta)]$ is the parameters θ^* such that $\mathbb{E}_{x \sim q}[g(x)] = \mathbb{E}_{x \sim p(x \mid \theta^*)}[g(x)]$.

Maximizing Likelihood for Exponential Families

The following result is helpful for quickly computing MLE solutions without algebra:

Theorem. Suppose that $p(x \mid \theta)$ is an "exponential family with sufficient statistics $g(x)$". Then for any $q(x)$, the solution to $\operatorname{argmax}_\theta \mathbb{E}_{x\sim q}[\log p(x \mid \theta)]$ is the parameters θ^* such that $\mathbb{E}_{x\sim q}[g(x)] = \mathbb{E}_{x\sim p(x\mid\theta^*)}[g(x)]$.

Will say what exponential families are next, but for now note that $g(x) = (x, x^2)$ for Gaussians, and $g(x) = x$ for Bernoulli.

Maximizing Likelihood for Exponential Families

The following result is helpful for quickly computing MLE solutions without algebra:

Theorem. Suppose that $p(x \mid \theta)$ is an "exponential family with sufficient statistics $g(x)$". Then for any $q(x)$, the solution to $\operatorname{argmax}_\theta \mathbb{E}_{x \sim q}[\log p(x \mid \theta)]$ is the parameters θ^* such that $\mathbb{E}_{x \sim q}[g(x)] = \mathbb{E}_{x \sim p(x \mid \theta^*)}[g(x)]$.

Will say what exponential families are next, but for now note that $g(x) = (x, x^2)$ for Gaussians, and $g(x) = x$ for Bernoulli.

Tells us that MLE for Bernoulli is $\pi^* =$ fraction of 1's, MLE for Gaussian is $\mu^*, \sigma^* =$ empirical mean and stdev.

Working out Gaussian updates

[on board]

Proving the Exponential Family Result

[on board]

Proving the Exponential Family Result

[on board]

Next: sampling

Sampling: General Idea

Recall: two frameworks

- Maximize $\log p(x \mid \theta) = \log\left(\sum_z p(x, z \mid \theta)\right)$ (EM, last time)
- Place prior on θ, sample $p(\theta, z \mid x)$ (this time)

Sampling: General Idea

Recall: two frameworks

- Maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$ (EM, last time)
- Place prior on θ, sample $p(\theta, z \mid x)$ (this time)

Why samples?

- Interpretable, efficient way to represent a distribution

Sampling: General Idea

Recall: two frameworks

- Maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$ (EM, last time)
- Place prior on θ, sample $p(\theta, z \mid x)$ (this time)

Why samples?

- Interpretable, efficient way to represent a distribution
- Can approximate any statistic:

$$\mathbb{E}_{x \sim p}[f(x)] \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i), \tag{1}$$

where the x_i are n samples from p.

Sampling Algorithms

Eventual target: Metropolis-Hastings algorithm (MCMC)

- Named among the “top 10 algorithms of the 20th century”

Sampling Algorithms

Eventual target: Metropolis-Hastings algorithm (MCMC)

- Named among the “top 10 algorithms of the 20th century”

First, need some build-up:

- Rejection sampling
- Importance sampling

Rejection sampling

[board and Jupyter demo]

Importance sampling

[board]

Review: Markov chains

[board]