S&DS 365 / 665

Intermediate Machine Learning

Variational Inference

October 9

Yale

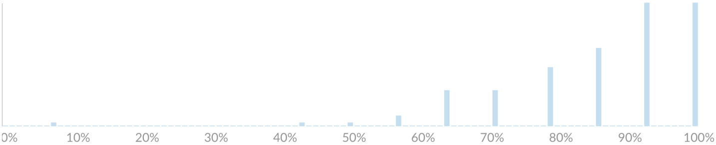
Reminders

- Assignment 2 due Wednesday
- Midterm week from today, in class
- Material up to and including today's class
- Cheat sheet: One side if 8.5x11 sheet, handwritten
- Multiple review sessions scheduled (see Canvas)
- Quiz 3 scores posted; discussed in review sessions
- Practice midterms posted

Quiz Summary

Section Filter ▾ | Student Analysis | Item Analysis

Average Score	High Score	Low Score	Standard Deviation	Average Time
86%	100%	7%	1.99	25:45



Review sessions

Chris:	Thursday, Oct 12, 7pm
Ruixiao:	Friday, Oct 13, 7pm
Zehao:	Saturday, Oct 14, 4pm

For Today

- Variational inference: The ELBO
- Derivations and examples
- Next time: Variational autoencoders (VAEs)

Recall: Inverting generative models

Template for generative model:

1. Choose Z
2. Given z, generate (sample) X

We often want to invert this:

1. Given x
2. What is Z that generated it?

Inverting models

Bayesian setup:

1. Choose θ
2. Given θ, generate (sample) X

Posterior inference:

1. Given x
2. What is θ that generated it?

Approximate inference

If we have a random vector $Z \sim p(Z \mid x)$, we might want to compute the following:

- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$
- maximum marginals $z_i^\star = \arg\max_{z_i} \mathbb{P}(Z_i = z_i \mid x)$
- joint probability $\mathbb{P}(Z \mid x)$
- joint mean $\mathbb{E}(Z \mid x)$

Each of these quantities is intractable to calculate exactly, in general.

Variational methods

- Gibbs sampling is *stochastic* approximation
- Variational methods iteratively refine *deterministic* approximations
- Variational and Markov chain approximations originated in physics

Example: Ising model

We have a graph with edges E and vertices V. Each node i has a random variable Z_i that can be "up" ($Z_i = 1$) or "down" ($Z_i = 0$)

$$\mathbb{P}_\beta(z_1, \ldots, z_n) \propto \exp\left(\sum_{s \in V} \beta_s z_s + \sum_{(s,t) \in E} \beta_{st} z_s \, z_t\right)$$

This is called an "Ising model" and is central to statistical physics.

Example: Ising model

We have a graph with edges E and vertices V. Each node i has a random variable Z_i that can be "up" ($Z_i = 1$) or "down" ($Z_i = 0$)

$$\mathbb{P}_\beta(z_1, \ldots, z_n) \propto \exp\left(\sum_{s \in V} \beta_s z_s + \sum_{(s,t) \in E} \beta_{st} z_s z_t\right)$$

E are the set of edges, V are the vertices. Imagine the Z_i are votes of politicians, and the edges encode the social network of party affiliations

Ising model—the "discrete Gaussian"

The nodes can take values in $\{0, 1\}$ or $\{-1, 1\}$—The two encodings are equivalent (exercise)

We can think of the Ising model as a "discrete Gaussian".

Why? The multivariate Gaussian and Ising models have exactly the same form, but the sample spaces are different.

The normalizing constant for the multivariate Gaussian has a closed form; not so for the Ising model.

Stochastic approximation

Gibbs sampler

Iterate until converged:

1. Choose vertex $s \in V$ at random
2. Sample z_s holding others fixed

$$\theta_s = \text{sigmoid}\left(\beta_s + \sum_{t \in N(s)} \beta_{st} z_t\right)$$
$$Z_s \,|\, \theta_s \sim \text{Bernoulli}(\theta_s)$$

You should verify that this is the correct conditional distribution of Z_s holding the others fixed.

$\text{sigmoid}(u) = 1/(1 + e^{-u})$

Deterministic approximation

Mean field variational algorithm

Iterate until converged:

1. Choose vertex $s \in V$ at random
2. Update mean μ_s holding others fixed

$$\mu_s = \text{sigmoid}\left(\beta_s + \sum_{t \in N(s)} \beta_{st}\, \mu_t\right)$$

sigmoid$(u) = 1/(1 + e^{-u})$

Determimnistic vs. stochastic approximation

- The z_s variables are random
- The μ_s variables are deterministic
- The Gibbs sampler convergence is in distribution
- The mean field convergence is numerical
- The Gibbs sampler approximates the full distribution
- The mean field algorithm approximates the mean of each node

Example 2: A finite mixture model

Fix two distributions F_0 and F_1, with densities $f_0(x)$ and $f_1(x)$, and form the mixture model

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ X \mid \theta &\sim \theta F_1 + (1-\theta) F_0. \end{aligned}$$

The likelihood for data $x_1, \ldots, x_n$ is

$$p(x_{1:n}) \;=\; \int_0^1 \text{Beta}(\theta \mid \alpha, \beta) \prod_{i=1}^{n} \left(\theta f_1(x_i) + (1-\theta) f_0(x_i)\right) \, d\theta.$$

Our goal is to approximate the posterior $p(\theta \mid x_{1:n})$

Stochastic approximation

Gibbs sampler

1. Sample $Z_i \,|\, \theta, x_{1:n}$ for $i = 1, \dots, n$
2. Sample $\theta \,|\, z_{1:n}, x_{1:n}$

The first step is carried out by sampling

$$Z_i = \begin{cases} 1 & \text{with probability} \propto \theta f_1(x_i) \\ 0 & \text{with probability} \propto (1-\theta) f_0(x_i) \end{cases}$$

Stochastic approximation

Gibbs sampler

1. Sample $Z_i \mid \theta, x_{1:n}$ for $i = 1, \ldots, n$
2. Sample $\theta \mid z_{1:n}, x_{1:n}$

The second step is carried out by sampling

$$\theta \sim \mathrm{Beta}\left(\sum_{i=1}^{n} z_i + \alpha, n - \sum_{i=1}^{n} z_i + \beta\right).$$

Posterior over θ is approximated as *mixture* of Beta distributions; number of components is $n + 1$

Variational inference: Strategy

- We'd like to compute $p(\theta, z \mid x)$, but it's too complicated.
- Strategy: Approximate as $q(\theta, z)$ that has a "nice" form
- q is a function of variational parameters, optimized for each x.
- Maximize a lower bound on $p(x)$.

Variational inference: The ELBO

The ELBO is the following lower bound on $\log p(x)$:

$$
\begin{aligned}
\log\ p(x) &= \int \sum_z q(z,\theta) \log p(x)\, d\theta \\
&= \sum_z \int q(z,\theta) \log \left(\frac{p(x,z,\theta)\, q(z,\theta)}{p(z,\theta \mid x)\, q(z,\theta)} \right) d\theta \\
&= \sum_z \int q(z,\theta) \log \left(\frac{p(x,z,\theta)}{q(z,\theta)} \right) d\theta + \sum_z \int q(z,\theta) \log \left(\frac{q(z,\theta)}{p(z,\theta \mid x)} \right) d\theta \\
&\geq \sum_z \int q(z,\theta) \log \left(\frac{p(x,z,\theta)}{q(z,\theta)} \right) d\theta \\
&= H(q) + \mathbb{E}_q\big(\log p(x,z,\theta)\big)
\end{aligned}
$$

We maximize this over the parameters of q

Variational inference: The ELBO

The inequality above uses concavity of the logarithm:

$$\log\left(\sum_\alpha w_\alpha x_\alpha\right) \geq \sum_\alpha w_\alpha \log x_\alpha$$

So, if $q_\alpha \geq 0$ and $p_\alpha \geq 0$ sum (or integrate) to one, then

$$0 = \log\left(\sum_\alpha p_\alpha\right) = \log\left(\sum_\alpha q_\alpha \frac{p_\alpha}{q_\alpha}\right) \geq \sum_\alpha q_\alpha \log\left(\frac{p_\alpha}{q_\alpha}\right)$$

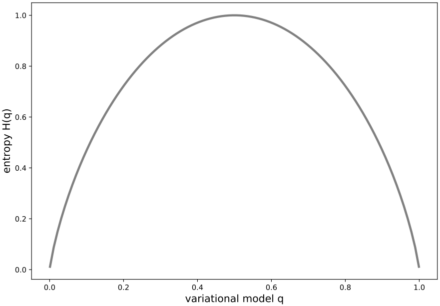
Therefore

$$\sum_\alpha q_\alpha \log\left(\frac{q_\alpha}{p_\alpha}\right) \geq 0$$

Variational inference: The ELBO

The ELBO is $H(q) + \mathbb{E}_q(\log p)$

The entropy term $H(q)$ encourages q to be spread out:



The cross-entropy $\mathbb{E}_q \log p$ tries to match q to p

Example 2: A finite mixture model

Fix two distributions F_0 and F_1, with densities $f_0(x)$ and $f_1(x)$, and form the mixture model

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ X \,|\, \theta &\sim \theta F_1 + (1-\theta) F_0. \end{aligned}$$

The likelihood for data $x_1, \ldots, x_n$ is

$$p(x_{1:n}) \;=\; \int_0^1 \text{Beta}(\theta \,|\, \alpha, \beta) \prod_{i=1}^{n} \left(\theta f_1(x_i) + (1-\theta) f_0(x_i)\right) \, d\theta.$$

Our goal is to approximate the posterior $p(\theta \,|\, x_{1:n})$

Variational approximation

Our variational approximation is

$$q(z, \theta) = q(\theta \,|\, \gamma_1, \gamma_2) \prod_{i=1}^{n} q_i^{z_i} (1 - q_i)^{(1-z_i)}$$

where $q(\theta \,|\, \gamma_1, \gamma_2)$ is a Beta(γ_1, γ_2) distribution, and $0 \leq q_i \leq 1$ are n free parameters.

Need to maximize ELBO $H(q) + \mathbb{E}_q \log p$

Let's sketch part of the calculation

Variational approximation

First, we have

$$\log p(x, \theta, z) = \log p(\theta \mid \alpha, \beta) + \sum_{i=1}^{n} \left\{ \log\left(\theta^{z_i} f_1(x_i)\right) + \log\left(\theta^{1-z_i} f_0(x_i)\right) \right\}$$

Next we use identities such as

$$\mathbb{E}_q \log \theta = \psi(\gamma_1) - \psi(\gamma_1 + \gamma_2)$$

for the digamma function $\psi(\cdot)$.

After some calculus and algebra, we end up with the following algorithm (see the notes on the course web page for more detail)

Variational algorithm for mixture

Variational inference

Iterate the following steps for variational parameters $q_{1:n}$ and (γ_1, γ_2):

1. Holding q_i fixed, set $\gamma = (\gamma_1, \gamma_2)$ to

$$\gamma_1 = \alpha + \sum_{i=1}^{n} q_i \qquad \gamma_2 = \beta + n - \sum_{i=1}^{n} q_i$$

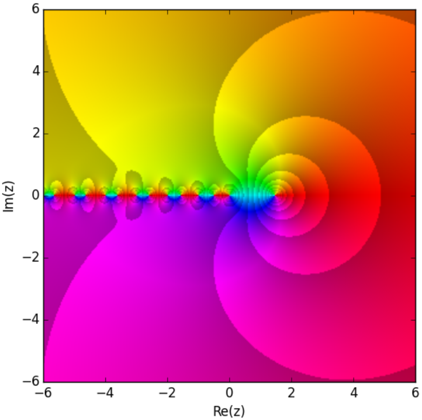
2. Holding γ_1 and γ_2 fixed, set q_i to

$$q_i = \frac{f_1(x_i) \exp \psi(\gamma_1)}{f_1(x_i) \exp \psi(\gamma_1) + f_0(x_i) \exp \psi(\gamma_2)}$$

After convergence, approximate posterior distribution over θ is

$$\widehat{p}(\theta \mid x_{1:n}) = \text{Beta}(\theta \mid \gamma_1, \gamma_2)$$

Digamma function



$\psi(x)$ is the *digamma function*

https://en.wikipedia.org/wiki/Digamma_function

Deterministic approximation

- Convergence is numerical, not stochastic
- Posterior is approximated as a *single* Beta
- Very similar algorithm is used for topic models

Example 3: More general mixtures

$$\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$$

$$X \mid \theta \sim \theta_1 F_1 + \cdots + \theta_k F_k$$

The likelihood for single data point x is

$$p(x) = \int \text{Dirichlet}(\theta \mid \alpha_1, \ldots, \alpha_k) \left(\sum_{j=1}^{k} \theta_j f_j(x) \right) d\theta.$$

When distributions F_j are learned, this is a “topic model.” Variational inference is one of the most useful ways of training topic models

- Notes on variational methods: Please read Sections 1–4
- The other sections are more advanced / specialized material
- Next up: Variational autoencoders
- VAEs not on midterm, but please review basics of variational methods discussed so far, including the ELBO

Some questions about variational inference

Variational inference

Q: What is the best q we could use?

A: The true posterior $q(z, \theta \mid x) = p(z, \theta \mid x)$

Why? Because this maximizes the ELBO. Mathematically,

$$\sum_z \int q(z, \theta) \log \left(\frac{q(z, \theta)}{p(z, \theta \mid x)} \right) d\theta = 0$$

in this case, so the ELBO inequality is an equality.

Variational inference

Q: How does the ELBO regularize?

A: The entropy term favors distributions that are “spread out”

Why? For discrete distributions the maximum entropy distribution is uniform. For Gaussian, the entropy is $\log \sigma^2$ which favors σ^2 large.

Variational inference

Q: Is the ELBO easy to maximize?

A: No.

Why? In general it is non-convex, and the solution depends on where we start an iterative algorithm. This is unlike the Gibbs sampler, which converges to the right thing if we wait long enough.

Next class: Variational autoencoders

- Variational autoencoders are generative models that are trained using variational inference
- The “decoder” is a neural net that generates from a latent variable
- The “encoder” approximates the posterior distribution with another neural network trained using variational inference

Summary

- Gibbs sampling makes stochastic approximations
- Variational methods make deterministic approximations
- General recipe: Maximize ELBO over variational parameters
- Gives a powerful approach to generative modeling