S&DS 365 / 665

Intermediate Machine Learning

Approximate Inference: Simulation and Variational Methods

March 7

Yale

Reminders

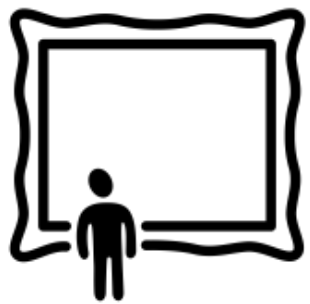
- Assignment 2 due Wednesday
- Midterm in class on March 16
  - practice exam posted Thursday
  - review sessions TBA
- Posted: Notes on Gibbs sampling for DPM

For Today

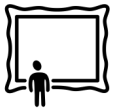
- Where are we headed? Road map for next few weeks
- Approximate inference: What's it all about?
- Approximate inference with Gibbs sampling (continued)
- Variational methods:
  - Mean field approximation
  - Two examples

Where have we gone, where are we headed?

Let's pause to discuss the "big picture"

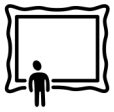


These are kind of the opposite of the technical details marked  that take us into the weeds...



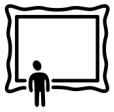
Different approaches

	Frequentist	**Bayesian**
Probability	limiting frequency	degree of subjective belief
Parameter θ:	fixed constant	random variable
Statements are about:	procedures	parameters
Frequency guarantees?	yes	no



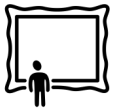
Correspondence

Statistical problem	Frequentist approach	Bayesian approach
regression	kernel smoother	Gaussian process
CDF estimation	empirical cdf	Dirichlet process
density estimation	kernel density estimator	Dirichlet process mixture
regression	wide neural network	Gaussian process



Bayesian computatation

- Computing Bayesian posteriors can be impractical
- Two approaches: Simulation and variational
- Simulation is the “right” way to do it — unbiased
- Variational methods are an alternative



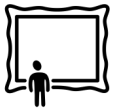
Inverting generative models

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?



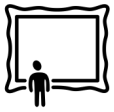
Inverting models

Bayesian setup:

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

Posterior inference:

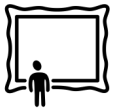
1. Given x
2. What is θ that generated it?



Approximate inference

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

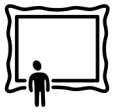
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$



Approximate inference

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

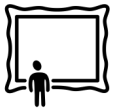
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$



Approximate inference

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

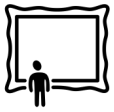
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$



Approximate inference

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

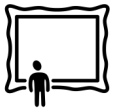
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$
- maximum marginals $z_i^\star = \arg\max_{z_i} \mathbb{P}(Z_i = z_i \mid x)$



Approximate inference

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

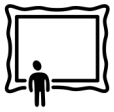
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$
- maximum marginals $z_i^\star = \arg\max_{z_i} \mathbb{P}(Z_i = z_i \mid x)$
- joint probability $\mathbb{P}(Z \mid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



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 $\mathbb{E}(Z \mid x)$



Approximate inference

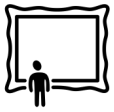
If we have a random vector $Z \sim p(Z \mid x)$, we might want to compute the following:

- 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.

Advanced generative models

https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html?



Advanced generative models

- Simulation and variational methods are two broad classes of approaches to inverting models
- We'll explore these in the next few classes
- First, let's finish our derivation of Gibbs sampling for DP mixtures

Nonparametric Bayesian mixture model

$$
\begin{aligned}
F &\sim DP(\alpha, F_0) \\
\theta_1, \ldots, \theta_n \,|\, F &\sim F \\
X_i \,|\, \theta_i &\sim f(x \,|\, \theta_i), \quad i = 1, \ldots, n.
\end{aligned}
$$

The posterior for a DPM

- The posterior distribution does not have a closed form — need to approximate it algorithmically
- The random variables Z_i here correspond to which cluster (table) the data point (customer) x_i is generated from (served at)

Gibbs sampling

Let's go to the board and finish our "rolled up sleeves" derivation of the Gibbs sampling algorithm for Dirichlet process mixtures, started last time

We've also posted notes to our course page so you can follow the arguments at your own pace...

Gibbs sampling for the DPM

For each point x_i:

(a) For every non-empty cluster j, compute

$$w_j = \frac{n_{j,-i}}{n - 1 + \alpha} \, p(x_i \mid x_{i'} \text{ in cluster } j \text{ for } i' \neq i)$$

For the empty cluster, compute

$$w_0 = \frac{\alpha}{n - 1 + \alpha} \, p(x_i \mid F_0)$$

(b) Normalize $w_j \leftarrow \dfrac{w_j}{\sum_k w_k}$ so the weights sum to one

(b) Reassign x_i to cluster j with probability w_j (possibly starting a new cluster)

Gibbs sampling for the DPM

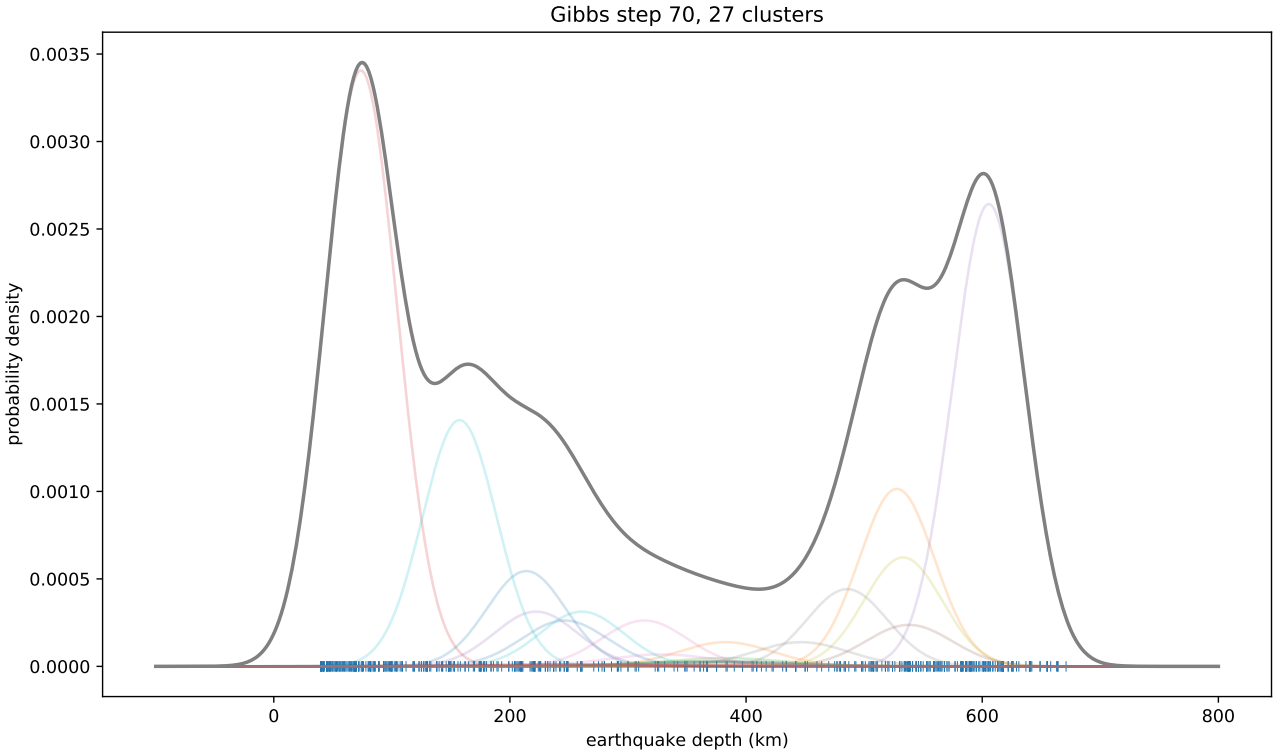
For each clustering $c^{(b)}$, get a predictive density

$$p(x \mid c_{1:n}^{(b)}, x_{1:n}) = \sum_j \frac{n_j^{(b)}}{n+\alpha}\, p\big(x \mid x_i \text{ in cluster } c_j^{(b)}\big) + \frac{\alpha}{n+\alpha}\, p(x \mid F_0).$$

The posterior mean is approximated as

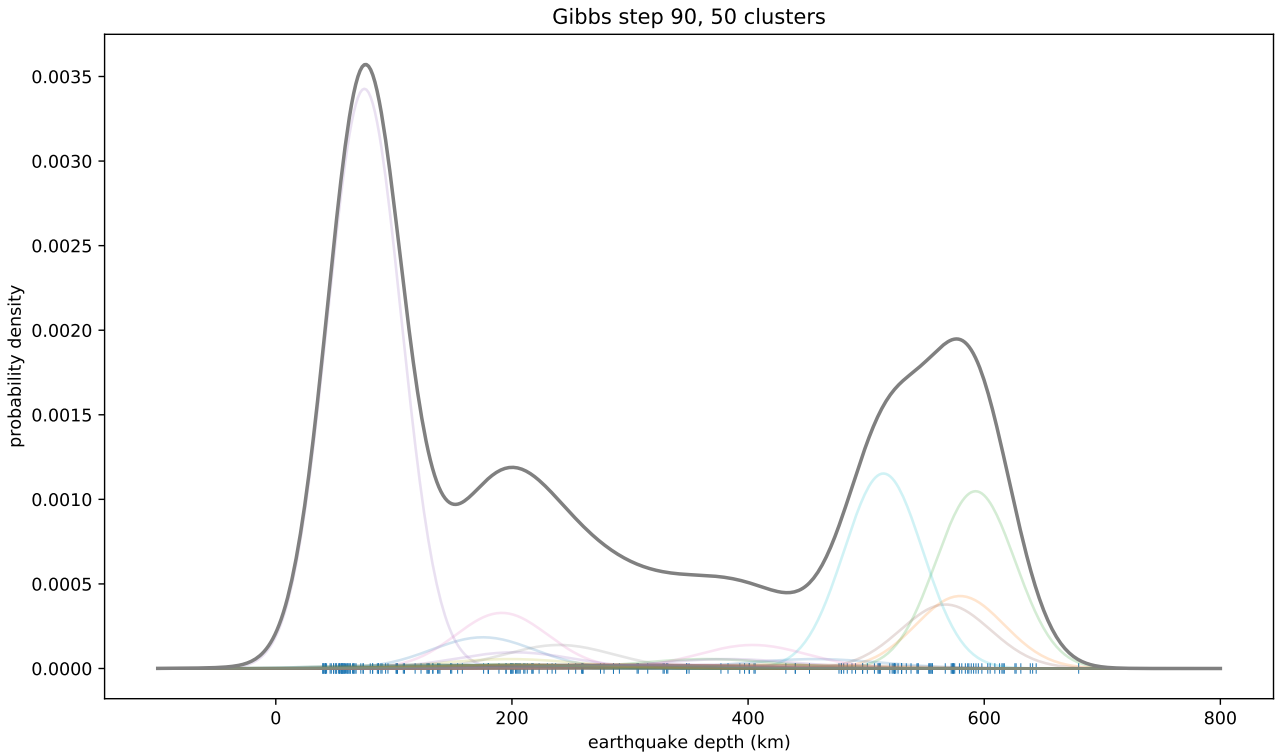
$$p(x_{n+1} \mid x_{1:n}) = \frac{1}{B}\sum_{b=1}^{B} p(x_{n+1} \mid c_{1:n}^{(b)}, x_{1:n})$$

Example: Fiji Earthquake data



Example: Fiji earthquake data. The prior over the Gaussian mean θ is a Dirichlet process $DP(\alpha, F_0)$ with $\alpha = 5$ and base model $F_0 = N(\mu_0, \tau_0^2)$ with $\mu_0 = 300$ and $\tau_0 = 70$. The density shown is a particular density sampled from the posterior after 70 Gibbs sampling steps; in this case the number of clusters is 27 (color curves).

Example: Fiji Earthquake data



You'll implement this for Assn 3—with plenty of help!

Variational methods

- Gibbs sampling is *stochastic* approximation
- Variational methods iteratively refine *deterministic* approximations
- We'll first discuss "mean field" approximations, originating in statistical physics

---

Stochastic simulation methods like Monte Carlo algorithms also originated in physics, notably during the "Manhattan project"

Example 1: Interacting particles

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) 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 1: Interacting particles

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) or "down" ($Z_i = 0$)

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

Imagine the Z_i are votes of politicians, and the edges encode the social network of party affiliations

Stochastic approximation

Gibbs sampler

1. Choose vertex $s \in V$ at random
2. Sample $u \sim \text{Uniform}(0, 1)$ and update

$$z_s = \begin{cases} 1 & u \leq \left(1 + \exp\left(-\beta_s - \sum_{t \in N(s)} \beta_{st} z_t\right)\right)^{-1} \\ 0 & \text{otherwise} \end{cases}$$

3. Iterate

Determimnistic approximation

Mean field variational algorithm

1. Choose vertex $s \in V$ at random
2. Update
$$\mu_s = \left(1 + \exp\left(-\beta_s - \sum_{t \in N(s)} \beta_{st}\, \mu_t\right)\right)^{-1}$$
3. Iterate

Determimnistic vs. stochastic approximation

- The z_i variables are random
- The μ_i 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

Think of how to interpret this with Z_i the vote of politician i

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

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

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

The first step is carried out by sampling $u_i \sim \text{Uniform}(0, 1)$, independently for each i, and selecting

$$z_i = \begin{cases} 1 & \text{if } u_i \leq \dfrac{\theta f_1(x_i)}{\theta f_1(x_i) + (1-\theta) f_0(x_i)} \\ 0 & \text{otherwise.} \end{cases}$$

Posterior is approximated as *mixture* of Beta distributions, number of components is $n + 1$

Stochastic approximation

Gibbs sampler

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

The second step is carried out by sampling

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

Posterior is approximated as *mixture* of Beta distributions, number of components is $n + 1$

Deterministic approximation

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

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

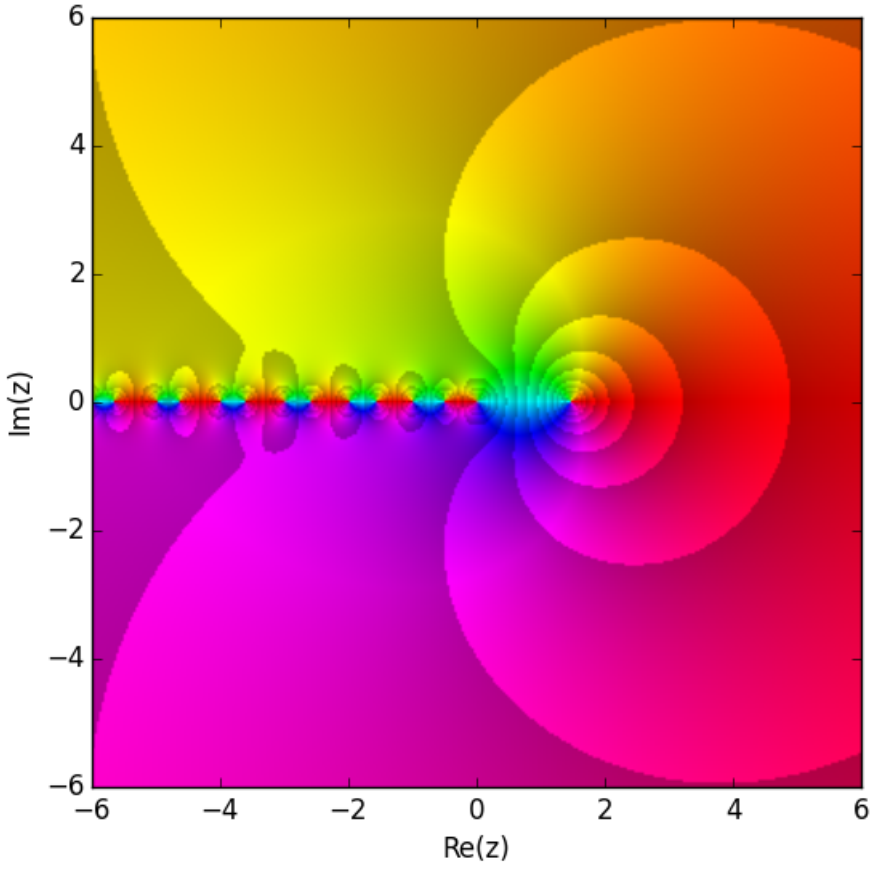
2. Holding γ_1 and γ_2 fixed, set q_i to

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

After convergence, approximate posterior distribution over θ is

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

Digamma



$\Psi(x)$ 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 distribution
- We'll see next time where this algorithm comes from

Summary

- Approximation is required for many types of models
- Two forms: Simulation (Gibbs sampling) and variational methods
- For DP mixtures, Gibbs sampling approximates posterior by iteratively re-clustering the data
- Gibbs sampling iteratively makes stochastic approximations
- Variational methods iteratively make deterministic approximations