S&DS 365 / 665

Intermediate Machine Learning

Approximate Inference: Simulation and Variational Methods

October 4

Yale

Reminders

- Quiz 3 available today at 10:30 am
  - Available for 48 hours
  - 30 minutes once started
  - DD, CNNs, GPs
- Assignment 2 due week from today
- Midterm in class on Monday October 16
  - practice midterms posted to Canvas
  - review sessions TBA

For Today

- Finish up Gaussian processes
- Where are we headed? Road map for next couple classes
- Approximate inference: What's it all about?
- Approximate inference with Gibbs sampling
- Variational methods:
  - Mean field approximation
  - Two examples

Gaussian processes

The key to computation in the Gaussian process is *conjugacy*: If the noise model is Gaussian, the posterior is another Gaussian process.

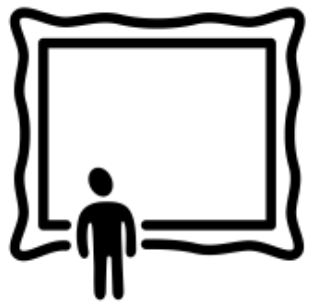
- Everything boils down to multivariate Gaussian calculations

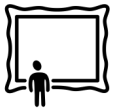
For classification, the likelihood is not Gaussian. We don't have conjugacy and the computations become more complex.

- We need to carry out some sort of *approximation*

Where have we gone, where are we headed?

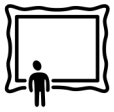
Let’s pause to discuss the “big picture”





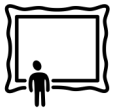
Different approaches

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



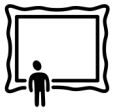
Correspondence

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 computation

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



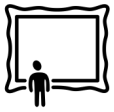
Inverting generative models

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?



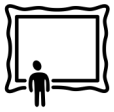
Inverting models

Bayesian setup:

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

Posterior inference:

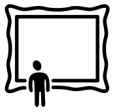
1. Given x
2. What is θ that generated it?



Approximate inference

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

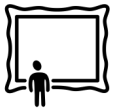
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$



Approximate inference

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

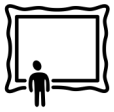
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$



Approximate inference

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

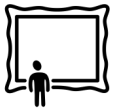
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$



Approximate inference

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

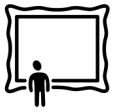
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$
- maximum marginals $z_i^\star = \arg\max_{z_i} \mathbb{P}(Z_i = z_i \mid x)$



Approximate inference

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

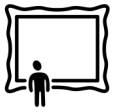
- marginal probabilities $\mathbb{P}(Z_i = z \mid x)$
- marginal means $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments $z^\star = \arg\max_z \mathbb{P}(\{Z_i = z_i\} \mid x)$
- maximum marginals $z_i^\star = \arg\max_{z_i} \mathbb{P}(Z_i = z_i \mid x)$
- joint probability $\mathbb{P}(Z \mid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



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

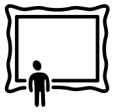
If we have a random vector $Z \sim p(Z \mid x)$, we might want to compute the following:

- 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 couple classes

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" ($Z_i = 1$) 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 1: Interacting particles

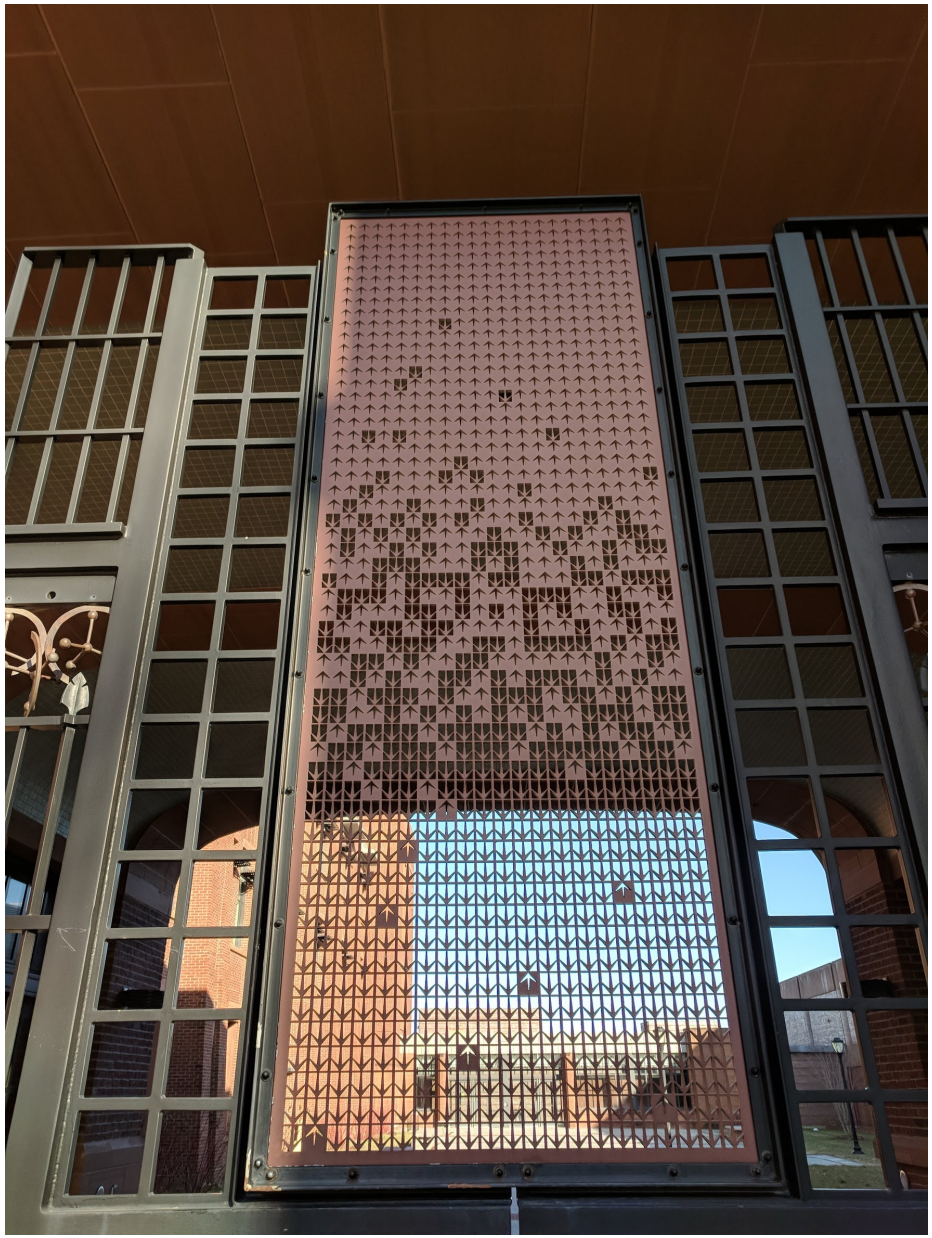
We have a graph with edges E and vertices V. Each node i has a random variable Z_i that can be "up" ($Z_i = 1$) or "down" ($Z_i = 0$)

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

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

Sterling Gate

Sterling Gate



Ising model

- Generative model for Z
- However, we can't sample from it directly
- Instead, we sample each component iteratively

Stochastic approximation

Gibbs sampler

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

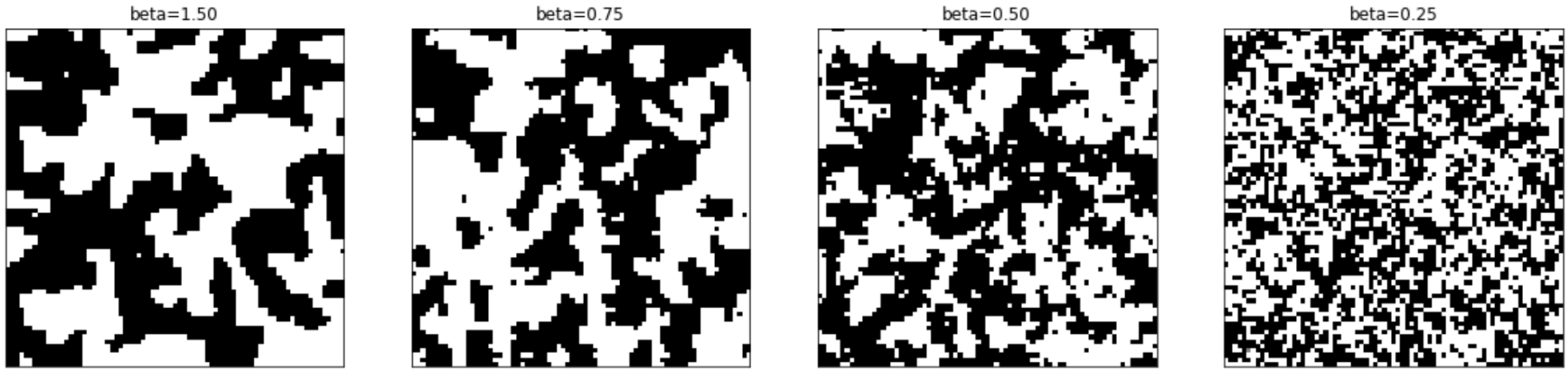
$$z_s = \begin{cases} 1 & u \leq \left(1 + \exp\left(-\beta_s - \sum_{t \in N(s)} \beta_{st} z_t\right)\right)^{-1} \\ 0 & \text{otherwise} \end{cases}$$

3. Iterate

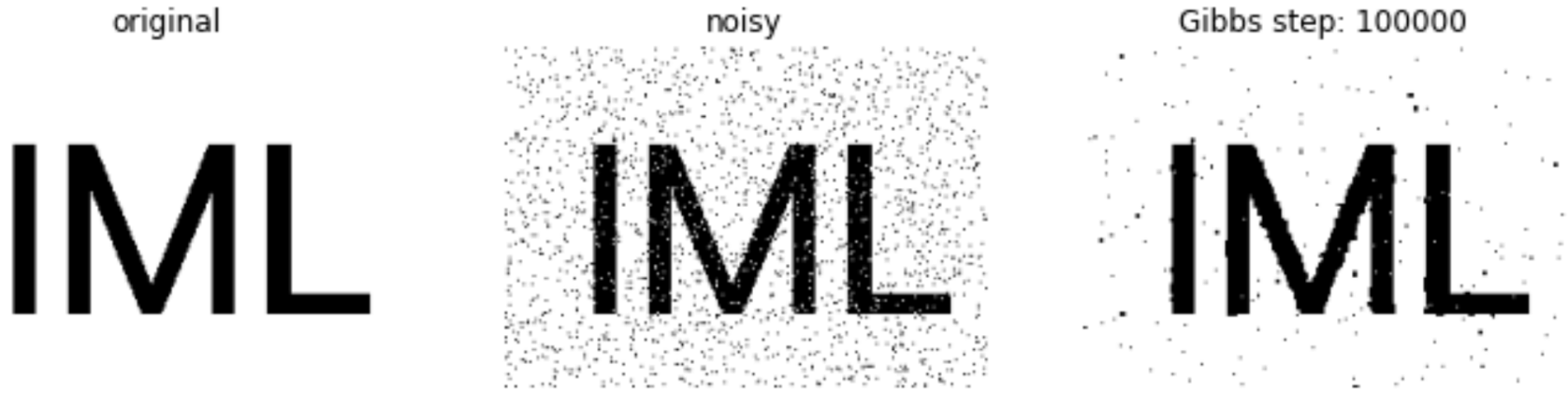
Also called“Glauber dynamics” in some communities.

Resting state

Demo



Demo



Demo

original

noisy

Gibbs step: 500000

Notes on Simulation

To learn more, see posted notes on simulation and Gibbs sampling

14.5 Why It Works

An understanding of why MCMC works requires elementary Markov chain theory, which is reviewed in an Appendix at the end of this chapter.

Recall that a distribution π satisfies *detailed balance* for a Markov chain if

$$p_{ij}\pi_i = p_{ji}\pi_j. \tag{14.57}$$

If π satisfies detailed balance, then it is a stationary distribution for the chain.

Because we are now dealing with continuous state Markov chains, we will change notation a little and write $p(x, y)$ for the probability of making a transition from x to y. Also, let's use $f(x)$ instead of π for a distribution. In this new notation, f is a stationary distribution if $f(x) = \int f(y)p(y, x)\, dy$ and detailed balance holds for f if

$$f(x)p(x, y) = f(y)p(y, x). \tag{14.58}$$

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"

Deterministic 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 $n + 1$

The "collapsed Gibbs sampler" integrates out θ.

Variational inference: Strategy

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

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

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

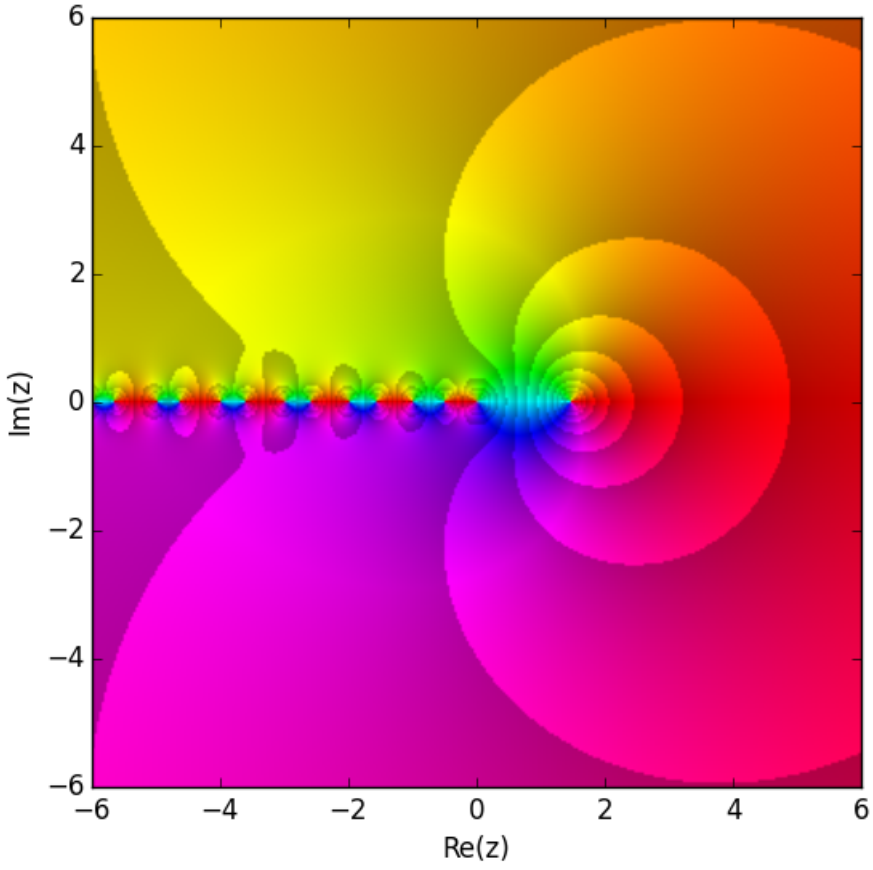
2. Holding γ_1 and γ_2 fixed, set q_i to

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

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 later where this algorithm comes from

Summary

- Approximation is required for many types of models
- Two forms: Simulation (Gibbs sampling) and variational methods
- Gibbs sampling iteratively makes stochastic approximations
- Variational methods iteratively make deterministic approximations