Variational Inference: The Basics

Philip Schulz and Wilker Aziz

Joint Distribution

Let X and Z be random variables. A generative model is any model that defines a joint distribution over these variables.

Joint Distribution

Let X and Z be random variables. A generative model is any model that defines a joint distribution over these variables.

2 Examples of Generative Models

- $p(x, z) = p(x)p(z|x)$
- $p(x, z) = p(z)p(x|z)$

Likelihood and prior

From here on, x is our observed data. On the other hand, z is an unobserved outcome.

- $p(x|z)$ is the **likelihood**
- $p(z)$ is the **prior** over Z

Notice: the prior may depend on a non-random quantity α (write $p(z|\alpha)$). In that case, we call α a hyperparameter.

Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{1}$$

Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$p(z|x) = \frac{\overbrace{p(x|z)}^{\text{likelihood}}\overbrace{p(z)}^{\textit{prior}}}{p(x)} \tag{2}$$

Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}}\,\overbrace{p(z)}^{\textit{prior}}}{p(x)} \tag{3}$$

Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}}\overbrace{p(z)}^{\text{prior}}}{\underbrace{p(x)}_{\text{marginal likelihood/evidence}}} \tag{4}$$

The Basic Problem

We want to compute the posterior over latent variables $p(z|x)$. This involves computing the marginal likelihood

$$p(x) = \int p(x, z) dz$$

which is often **intractable**. This problem motivates the use of **approximate inference** techniques.

Bayesian Inference

Under the Bayesian view, model parameters θ are also random. The generative model becomes

- $p(x, \theta)$ for fully observed data (supervised learning)
- $p(x, z, \theta)$ for observed and latent data (unsupervised learning)

Bayesian Inference

The evidence becomes even harder to compute. This is because θ is often high-dimensional (just think of neural nets!).

- $p(x) = \int p(x, \theta) d\theta$ (supervised learning)
- $p(x) = \int \int p(x, z, \theta) dz \; d\theta$ (unsupervised learning)

Bayesian Inference

The evidence becomes even harder to compute. This is because θ is often high-dimensional (just think of neural nets!).

- $p(x) = \int p(x, \theta) d\theta$ (supervised learning)
- $p(x) = \int \int p(x, z, \theta) dz \, d\theta$ (unsupervised learning)

Again, approximate inference is needed.