Variational Inference

Shubham Gupta

Department of Computer Science and Automation
Indian Institute of Science, Bangalore

April 29, 2017

Outline

Motivation

Motivation

Given:

1. $X = \{X_1, X_2, ..., X_m\}$: Set of **observed** random variables

Motivation

Given:

1. $X = \{X_1, X_2, ..., X_m\}$: Set of **observed** random variables
2. $Y = \{Y_1, Y_2, ..., Y_n\}$: Set of **unobserved** random variables

Motivation

Given:

1. $X = \{X_1, X_2, ..., X_m\}$: Set of **observed** random variables
2. $Y = \{Y_1, Y_2, ..., Y_n\}$: Set of **unobserved** random variables

Task: Find $P(Z|X)$ where $Z \subseteq Y$

Motivation

Given:

1. $X = \{X_1, X_2, ..., X_m\}$: Set of **observed** random variables
2. $Y = \{Y_1, Y_2, ..., Y_n\}$: Set of **unobserved** random variables

Task: Find $P(Z|X)$ where $Z \subseteq Y$

Bayes Theorem

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

Motivation

Given:

1. $X = \{X_1, X_2, ..., X_m\}$: Set of **observed** random variables
2. $Y = \{Y_1, Y_2, ..., Y_n\}$: Set of **unobserved** random variables

Task: Find $P(Z|X)$ where $Z \subseteq Y$

Bayes Theorem

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

Problem: Can not evaluate $P(X) = \int_Z P(X, z)\, dz$

Motivation

Given:

1. $X = \{X_1, X_2, \ldots, X_m\}$: Set of **observed** random variables
2. $Y = \{Y_1, Y_2, \ldots, Y_n\}$: Set of **unobserved** random variables

Task: Find $P(Z|X)$ where $Z \subseteq Y$

Bayes Theorem

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

Problem: Can not evaluate $P(X) = \int_Z P(X, z)\, dz$

Inference is NP-hard!

Variational Inference

Variational Inference

Goal: Approximate $P(Z|X)$ by $q_\theta(Z)$, $q_\theta(Z) \in \mathcal{D}$

Variational Inference

Goal: Approximate $P(Z|X)$ by $q_\theta(Z)$, $q_\theta(Z) \in \mathcal{D}$

where,

1. $\mathcal{D}$: A parametric class of distributions, eg. all normal distribution in d dimensions

Variational Inference

Goal: Approximate $P(Z|X)$ by $q_\theta(Z)$, $q_\theta(Z) \in \mathcal{D}$

where,

1. $\mathcal{D}$: A parametric class of distributions, eg. all normal distribution in d dimensions
2. $q_\theta(Z)$: A member of $\mathcal{D}$ parameterized by θ, eg. $\mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are parameters

Variational Inference

Goal: Approximate $P(Z|X)$ by $q_\theta(Z)$, $q_\theta(Z) \in \mathcal{D}$

where,

1. $\mathcal{D}$: A parametric class of distributions, eg. all normal distribution in d dimensions
2. $q_\theta(Z)$: A member of $\mathcal{D}$ parameterized by θ, eg. $\mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are parameters

Optimization Objective:

$$\max_\theta KL(q(Z;\theta) \parallel P(Z|X))$$

Evidence Lower Bound Objective (ELBO)

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Problem: $P(Z|X)$ is not known

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Problem: $P(Z|X)$ is not known

ELBO:

$$J(\theta) = \mathbb{E}_{z\sim q(Z;\theta)}\left[\log q(z;\theta) - \log P(z|X)\right]$$

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Problem: $P(Z|X)$ is not known

ELBO:

$$\begin{aligned} J(\theta) &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log q(z;\theta) - \log P(z|X)\big] \\ &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log q(z;\theta) - \log P(z,X) + \log P(X)\big] \end{aligned}$$

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Problem: $P(Z|X)$ is not known

ELBO:

$$\begin{aligned} J(\theta) &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log q(z;\theta) - \log P(z|X)\big] \\ &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log q(z;\theta) - \log P(z,X) + \log P(X)\big] \\ \Rightarrow \log P(X) &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log P(z,X) - \log q(z;\theta)\big] \\ &\quad + KL(q(Z;\theta) \parallel P(Z|X)) \end{aligned}$$

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Problem: $P(Z|X)$ is not known

ELBO:

$$\begin{aligned}
J(\theta) &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log q(z;\theta) - \log P(z|X)\big] \\
&= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log q(z;\theta) - \log P(z,X) + \log P(X)\big] \\
\Rightarrow \log P(X) &= \mathbb{E}_{z\sim q(Z;\theta)}\big[\log P(z,X) - \log q(z;\theta)\big] \\
&\quad + KL(q(Z;\theta) \parallel P(Z|X)) \\
&\geq \mathbb{E}_{z\sim q(Z;\theta)}\big[\log P(z,X) - \log q(z;\theta)\big]
\end{aligned}$$

Evidence Lower Bound Objective (ELBO)

Optimization Objective:

$$\max_{\theta} KL(q(Z;\theta) \parallel P(Z|X))$$

Problem: $P(Z|X)$ is not known

ELBO:

$$\begin{aligned}
J(\theta) &= \mathbb{E}_{z\sim q(Z;\theta)}\left[\log q(z;\theta) - \log P(z|X)\right] \\
&= \mathbb{E}_{z\sim q(Z;\theta)}\left[\log q(z;\theta) - \log P(z,X) + \log P(X)\right] \\
\Rightarrow \log P(X) &= \mathbb{E}_{z\sim q(Z;\theta)}\left[\log P(z,X) - \log q(z;\theta)\right] \\
&+ KL(q(Z;\theta) \parallel P(Z|X)) \\
&\geq \mathbb{E}_{z\sim q(Z;\theta)}\left[\log P(z,X) - \log q(z;\theta)\right] \\
&= ELBO(q)
\end{aligned}$$

Evidence Lower Bound Objective (ELBO) (Cont.)

$$\log P(X) \geq \mathbb{E}_{z\sim q(Z;\theta)}\big[\log P(z, X) - \log q(z;\theta)\big] = ELBO(q)$$

Evidence Lower Bound Objective (ELBO) (Cont.)

$$\log P(X) \geq \mathbb{E}_{z \sim q(Z;\theta)}\big[\log P(z, X) - \log q(z;\theta)\big] = ELBO(q)$$

1. Maximizing ELBO(q) $\equiv$ Minimizing $KL(q(Z;\theta) \parallel P(Z|X))$

Evidence Lower Bound Objective (ELBO) (Cont.)

$$\log P(X) \geq \mathbb{E}_{z \sim q(Z;\theta)}\big[\log P(z,X) - \log q(z;\theta)\big] = ELBO(q)$$

1. Maximizing ELBO(q) $\equiv$ Minimizing $KL(q(Z;\theta) \parallel P(Z|X))$
2. Usually it is *easy* to calculate $P(z,X)$

Evidence Lower Bound Objective (ELBO) (Cont.)

$$\log P(X) \geq \mathbb{E}_{z \sim q(Z;\theta)} \big[\log P(z, X) - \log q(z;\theta) \big] = ELBO(q)$$

1. Maximizing ELBO(q) $\equiv$ Minimizing $KL(q(Z;\theta) \parallel P(Z|X))$
2. Usually it is *easy* to calculate $P(z, X)$
3. **Problem:** Calculating expectation might be hard

Evidence Lower Bound Objective (ELBO) (Cont.)

$$\log P(X) \geq \mathbb{E}_{z\sim q(Z;\theta)}\big[\log P(z,X) - \log q(z;\theta)\big] = ELBO(q)$$

1. Maximizing ELBO(q) $\equiv$ Minimizing $KL(q(Z;\theta) \parallel P(Z|X))$
2. Usually it is *easy* to calculate $P(z,X)$
3. **Problem:** Calculating expectation might be hard

Mean-Field Assumption: $q(Z;\theta) = \prod_{i=1}^{n} q_i(Z_i;\theta)$

Evidence Lower Bound Objective (ELBO) (Cont.)

$$\log P(X) \geq \mathbb{E}_{z \sim q(Z;\theta)}\big[\log P(z, X) - \log q(z;\theta)\big] = ELBO(q)$$

1. Maximizing ELBO(q) $\equiv$ Minimizing $KL(q(Z;\theta) \parallel P(Z|X))$
2. Usually it is *easy* to calculate $P(z, X)$
3. **Problem:** Calculating expectation might be hard

Mean-Field Assumption: $q(Z;\theta) = \prod_{i=1}^{n} q_i(Z_i;\theta)$

Optimization Objective:

$$\max_{\theta} ELBO(q)$$

Univariate Gaussian

Univariate Gaussian

1. Need to estimate mean μ and variance σ^2

Univariate Gaussian

1. Need to estimate mean μ and variance σ^2
2. $q_1(\mu) = \mathcal{N}(\mu|m, s^2)$ and $q_2(\sigma^2) = \mathcal{G}(\sigma^2|\alpha, \beta)$

Univariate Gaussian

1. Need to estimate mean μ and variance σ^2
2. $q_1(\mu) = \mathcal{N}(\mu|m, s^2)$ and $q_2(\sigma^2) = \mathcal{G}(\sigma^2|\alpha, \beta)$
3. $P(\mu) = \mathcal{N}(\mu|m_p, \sigma_p^2)$ and $P(\sigma^2) = \mathcal{G}(\sigma^2|\alpha_p, \beta_p)$

Univariate Gaussian

1. Need to estimate mean μ and variance σ^2
2. $q_1(\mu) = \mathcal{N}(\mu|m, s^2)$ and $q_2(\sigma^2) = \mathcal{G}(\sigma^2|\alpha, \beta)$
3. $P(\mu) = \mathcal{N}(\mu|m_p, \sigma_p^2)$ and $P(\sigma^2) = \mathcal{G}(\sigma^2|\alpha_p, \beta_p)$
4. $P(X = x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$

Univariate Gaussian

1. Need to estimate mean μ and variance σ^2
2. $q_1(\mu) = \mathcal{N}(\mu|m, s^2)$ and $q_2(\sigma^2) = \mathcal{G}(\sigma^2|\alpha, \beta)$
3. $P(\mu) = \mathcal{N}(\mu|m_p, \sigma_p^2)$ and $P(\sigma^2) = \mathcal{G}(\sigma^2|\alpha_p, \beta_p)$
4. $P(X = x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$
5. m, s^2, α and β are variational parameters

Univariate Gaussian

1. Need to estimate mean μ and variance σ^2
2. $q_1(\mu) = \mathcal{N}(\mu|m, s^2)$ and $q_2(\sigma^2) = \mathcal{G}(\sigma^2|\alpha, \beta)$
3. $P(\mu) = \mathcal{N}(\mu|m_p, \sigma_p^2)$ and $P(\sigma^2) = \mathcal{G}(\sigma^2|\alpha_p, \beta_p)$
4. $P(X = x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$
5. m, s^2, α and β are variational parameters
6. m_p, s_p^2, α_p and β_p are hyper-parameters

Univariate Gaussian (Cont.)

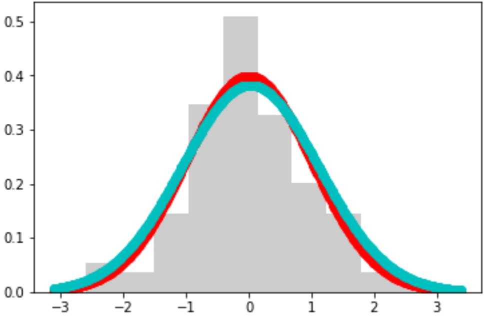

Figure 1: True (red) vs Approximated (blue) Distribution

Univariate Gaussian (Cont.)

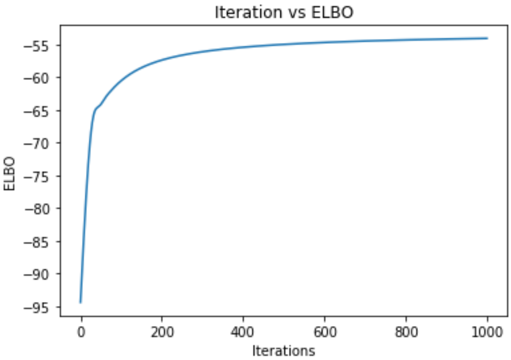

Figure 2: ELBO(q) vs Iterations

Gaussian Mixture Model

Gaussian Mixture Model

Same setup as the case of univariate Gaussian except:

Gaussian Mixture Model

Same setup as the case of univariate Gaussian except:

1. Log-Normal distribution is used as prior for variance

Gaussian Mixture Model

Same setup as the case of univariate Gaussian except:

1. Log-Normal distribution is used as prior for variance
2. Expectation is not computed analytically

Gaussian Mixture Model

Same setup as the case of univariate Gaussian except:

1. Log-Normal distribution is used as prior for variance
2. Expectation is not computed analytically

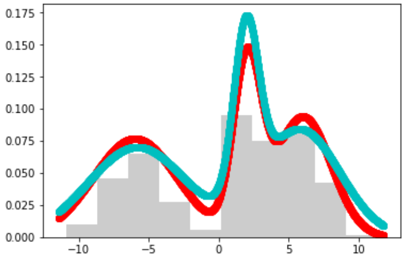

Figure 3: True (red) vs Approximated (blue) Distribution

Gaussian Mixture Model (Cont.)

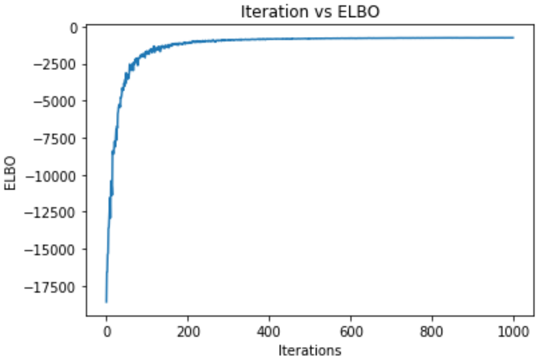

Figure 4: ELBO(q) vs Iterations

Introduction

Introduction

Goal: Generate images "similar" to images in the dataset

Introduction

Goal: Generate images "similar" to images in the dataset

1. Z: Low dimensional, latent representation

Introduction

Goal: Generate images "similar" to images in the dataset

1. Z: Low dimensional, latent representation
2. $P_\theta(X|Z)$: Model for data likelihood (Bernoulli)

Introduction

Goal: Generate images "similar" to images in the dataset

1. Z: Low dimensional, latent representation
2. $P_\theta(X|Z)$: Model for data likelihood (Bernoulli)
3. $q_\phi(Z)$: Approximation of $P(Z|X)$ (Gaussian)

Introduction

Goal: Generate images "similar" to images in the dataset

1. Z: Low dimensional, latent representation
2. $P_\theta(X|Z)$: Model for data likelihood (Bernoulli)
3. $q_\phi(Z)$: Approximation of $P(Z|X)$ (Gaussian)
4. $P(Z)$: Prior on Z (Gaussian)

Introduction

Goal: Generate images "similar" to images in the dataset

1. Z: Low dimensional, latent representation
2. $P_\theta(X|Z)$: Model for data likelihood (Bernoulli)
3. $q_\phi(Z)$: Approximation of $P(Z|X)$ (Gaussian)
4. $P(Z)$: Prior on Z (Gaussian)
5. θ and ϕ: Neural networks

Introduction

Goal: Generate images "similar" to images in the dataset

1. Z: Low dimensional, latent representation
2. $P_\theta(X|Z)$: Model for data likelihood (Bernoulli)
3. $q_\phi(Z)$: Approximation of $P(Z|X)$ (Gaussian)
4. $P(Z)$: Prior on Z (Gaussian)
5. θ and ϕ: Neural networks

Core Idea: Optimize ϕ and θ using variational inference

Alternative Expression for ELBO(q)

$$ELBO(q) = \mathbb{E}_{z\sim q}\big[\log P(z, X) - \log q(z;\theta)\big]$$

Alternative Expression for ELBO(q)

$$
\begin{aligned}
ELBO(q) &= \mathbb{E}_{z\sim q}\big[\log P(z,X) - \log q(z;\theta)\big] \\
&= \mathrm{E}_{z\sim q}\big[\log P(X|z) + \log P(z) - \log q(z;\theta)\big]
\end{aligned}
$$

Alternative Expression for ELBO(q)

$$
\begin{aligned}
ELBO(q) &= \mathbb{E}_{z\sim q}\big[\log P(z,X) - \log q(z;\theta)\big] \\
&= \mathrm{E}_{z\sim q}\big[\log P(X|z) + \log P(z) - \log q(z;\theta)\big] \\
&= \mathrm{E}_{z\sim q}\big[\log P(X|z)] - KL[q(Z;\theta)||P(Z)\big]
\end{aligned}
$$

Alternative Expression for ELBO(q)

$$
\begin{aligned}
ELBO(q) &= \mathbb{E}_{z\sim q}\big[\log P(z, X) - \log q(z;\theta)\big] \\
&= \mathrm{E}_{z\sim q}\big[\log P(X|z) + \log P(z) - \log q(z;\theta)\big] \\
&= \mathrm{E}_{z\sim q}\big[\log P(X|z)] - KL[q(Z;\theta)||P(Z)\big]
\end{aligned}
$$

This highlights the usual trade-off between prior and observations

Results on MNIST

Figure 5: Generated Images

Results on CIFAR-10

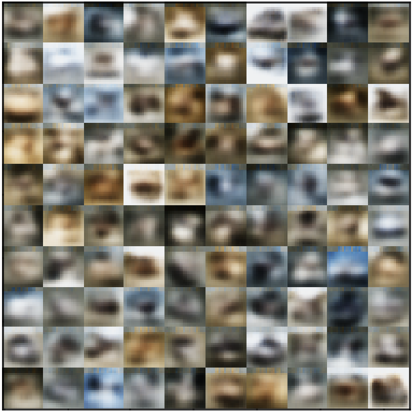

Figure 6: Generated Images

Other Topics

Other Topics

1. Comparison with maximum likelihood estimation for univariate Gaussian

Other Topics

1. Comparison with maximum likelihood estimation for univariate Gaussian
2. Comparison with Expectation Maximization algorithm for Gaussian Mixture Model

Other Topics

1. Comparison with maximum likelihood estimation for univariate Gaussian
2. Comparison with Expectation Maximization algorithm for Gaussian Mixture Model
3. Case study: Latent Dirichlet Allocation (LDA)

Other Topics

1. Comparison with maximum likelihood estimation for univariate Gaussian
2. Comparison with Expectation Maximization algorithm for Gaussian Mixture Model
3. Case study: Latent Dirichlet Allocation (LDA)
4. Comparison with MCMC methods

Other Topics

1. Comparison with maximum likelihood estimation for univariate Gaussian
2. Comparison with Expectation Maximization algorithm for Gaussian Mixture Model
3. Case study: Latent Dirichlet Allocation (LDA)
4. Comparison with MCMC methods
5. A survey of open research problems

Other Topics

1. Comparison with maximum likelihood estimation for univariate Gaussian
2. Comparison with Expectation Maximization algorithm for Gaussian Mixture Model
3. Case study: Latent Dirichlet Allocation (LDA)
4. Comparison with MCMC methods
5. A survey of open research problems

Tutorials and other resources are available at:
`https://github.com/sh-gupta/VariationalInference`

References

- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, *Variational Inference: A Review for Statisticians*, ArXiv e-prints (2016).
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent dirichlet allocation*, J. Mach. Learn. Res. **3** (2003), 993–1022.
- D. P Kingma and M. Welling, *Auto-Encoding Variational Bayes*, ArXiv e-prints (2013).

Thank You!