Auto-Encoding Variational Bayes

mingzailao

Sun Dec 11 15:03:06 2016

Outline

Variational Inference

Auto-Encoding Variational Bayes

Example: Variational Auto-Encoder

Auto-Encoding Variational Bayes

Variational Inference

Auto-Encoding Variational Bayes

Example: Variational Auto-Encoder

Variational Inference

Notations

1. latent variables : $\mathbf{z} = z_{1:m}$
2. obseevations : $\mathbf{x} = x_{1:n}$

Joint density

$$p(\mathbf{z}, \mathbf{x}) \;\; = \;\; p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$$

Variational Inference

Notes

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{z})}$$

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \int p(\mathbf{z},\mathbf{x})\,d\mathbf{z}$$

Variational Inference

One example

Bayesian mixture of unit-variance univariate Gaussians.

$$
\begin{aligned}
\mu_k &\sim \mathcal{N}(0, \sigma^2) \quad k = 1, \cdots, K \\
c_i &\sim Categorical(1/K, \cdots, 1/K) \quad i = 1, \cdots n \\
x_i | c_i, \mu &\sim \mathcal{N}(c_i^T \mu, 1) \quad i = 1, \cdots, n
\end{aligned}
$$

Variational Inference

One example

For a sample of size n, the joint density of latent and observed variables is:

$$p(\mu, \mathbf{c}, \mathbf{x}) = p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \mu)$$

then, the evidence is

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i | c_i, \mu) d\mu \\ &= \sum_{\mathbf{c}} p(\mathbf{c}) \int p(\mu) \prod_{i=1}^{n} p(x_i | c_i, \mu) d\mu \end{aligned}$$

Variational Inference

Notes

1. assume that $\mathcal{Z}$: a set of densities over the latent variables.
2. what we want :

$$q^*(\mathbf{z}) \quad = \quad arg \min_{q(\mathbf{z}) \in \mathcal{Z}} KL(q(\mathbf{z})|p(\mathbf{z}|\mathbf{x}))$$

3. Variational inference thus turns the inference problem into an optimization problem.

Variational Inference

recall KL

$$
\begin{aligned}
KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}|\mathbf{x})] \\
&= \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}] \\
&= \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z},\mathbf{x})] + \log p(\mathbf{x})
\end{aligned}
$$

Variational Inference

Define evidence lower bound(ELBO)

$$ELBO(q(\mathbf{z})) \;\; = \;\; \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z},\mathbf{x})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})]$$

then $\max ELBO(q(\mathbf{z})) \Leftrightarrow \min KL(q(\mathbf{z})|p(\mathbf{z}|\mathbf{x}))$

Another form of $ELBO$:

$$ELBO(q(\mathbf{z})) \;\; = \;\; \mathbb{E}_{q(\mathbf{z})}[p(\mathbf{x}|\mathbf{z})] - KL(q(\mathbf{z})||p(\mathbf{z}))$$

Auto-Encoding Variational Bayes

Variational Inference

Auto-Encoding Variational Bayes

Example: Variational Auto-Encoder

Auto-Encoding Variational Bayes

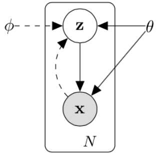

Figure 1: The type of directed graphical model under consideration. Solid lines denote the generative model $p_{\boldsymbol{\theta}}(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, dashed lines denote the variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to the intractable posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. The variational parameters $\boldsymbol{\phi}$ are learned jointly with the generative model parameters $\boldsymbol{\theta}$.

Auto-Encoding Variational Bayes

Define $q(\mathbf{z}|\mathbf{x}^{(i)})$ **use the information of** $\mathbf{x}^{(i)}$

Given dataset $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$,then :

$$\log p_\theta(\mathbf{x}^{(i)}) = KL(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z}|\mathbf{x}^{(i)})) + ELBO(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}))$$

ELBO

$$\begin{aligned} ELBO(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})] \\ &- KL(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[-\log q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) + \log p_\theta(\mathbf{x},\mathbf{z})] \end{aligned}$$

Auto-Encoding Variational Bayes

The usual (naive) Monte Carlo gradient estimator

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})] &= \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z}) \nabla_{q_\phi(\mathbf{z})} \log q_\phi(\mathbf{z})] \\ &\approx \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}) \nabla_{q_\phi(\mathbf{z}^{(l)})} \log q_\phi(\mathbf{z}^{(l)})\end{aligned}$$

where $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$

Auto-Encoding Variational Bayes

Chose $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\tilde{\mathbf{z}} \sim g_\phi(\epsilon, \mathbf{x}) \quad with \quad \epsilon \sim p(\epsilon)$$

Use Monte Carlo estimates of expectations

$$\begin{aligned} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[f(\mathbf{z})] &= \mathbb{E}_{p(\epsilon)}[f(g_\phi(\epsilon, \mathbf{x}^{(i)}))] \\ &\approx \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(\epsilon^{(l)}, \mathbf{x}^{(i)})) \quad where\ \epsilon^{(l)} \sim p(\epsilon) \end{aligned}$$

Auto-Encoding Variational Bayes

Apply the technique to the variational lower bound

Get generic Stochastic Gradient Variational Bayes (SGVB) estimator $\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) \approx \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$:

$$
\begin{aligned}
\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) &= \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}^{(i)}, \mathbf{z}^{(i,l)}) \\
&- \log q_\phi(\mathbf{z}^{(i,l)}|\mathbf{x}^{(i)})
\end{aligned}
$$

where $\mathbf{z}^{i,l} = g_\phi(\epsilon_{i,l}, \mathbf{x}^{(i)})$, and $\epsilon^{(i,l)} \sim p(\epsilon)$ for $\forall\ i, l$

Auto-Encoding Variational Bayes

Anther

$$\tilde{\mathcal{L}}^{(B)}(\theta, \phi; \mathbf{x}^{(i)}) \quad = \quad -KL(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L}(\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}$$

where $\mathbf{z}^{i,l} = g_\phi(\epsilon_{i,l}, \mathbf{x}^{(i)})$, and $\epsilon^{(i,l)} \sim p(\epsilon)$ for $\forall\ i, l$

Batch version

$$\tilde{\mathcal{L}}(\theta, \phi; \mathbf{X}) \approx \tilde{\mathcal{L}}^M(\theta, \phi; \mathbf{X}^M) \quad = \quad \frac{N}{M}\sum_{i=1}^{M}\tilde{\mathcal{L}}^M(\theta, \phi; \mathbf{x}^{(i)})$$

Auto-Encoding Variational Bayes

Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\boldsymbol{\epsilon} \leftarrow$ Random samples from noise distribution $p(\boldsymbol{\epsilon})$
  $\mathbf{g} \leftarrow \nabla_{\boldsymbol{\theta}, \boldsymbol{\phi}} \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M, \boldsymbol{\epsilon})$ (Gradients of minibatch estimator (8))
  $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Update parameters using gradients $\mathbf{g}$ (e.g. SGD or Adagrad [DHS10])
until convergence of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$
return $\boldsymbol{\theta}, \boldsymbol{\phi}$

Auto-Encoding Variational Bayes

Variational Inference

Auto-Encoding Variational Bayes

Example: Variational Auto-Encoder

Example : Variational Auto-Encoder

Idea:

Use a neural network for the probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ the approximation to the posterior of the generative model $p_\theta(\mathbf{x}, \mathbf{z})$, and where the parameters θ and ϕ are optimized jointly with the AEVB algorithm.

Example : Variational Auto-Encoder

Set $p_\theta(\mathbf{z})$:

$$p_\theta(\mathbf{z}) \quad = \quad \mathcal{N}(\mathbf{z}; 0, \mathbf{I})$$

Notes

- In this case, the prior lacks parameters.

Example : Variational Auto-Encoder

Set $p_\theta(\mathbf{x}|\mathbf{z})$:

1. Multivariate Gaussian (in case of real-valued data)
2. Bernoulli (in case of binary data)

Both parameters of the distributions are computed from $\mathbf{z}$ with a MLP(a fully-connected neural network with a single hidden layer)

Example : Variational Auto-Encoder

Set $q_\phi(\mathbf{z}|\mathbf{x})$

$$\log q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \quad = \quad \log \mathcal{N}(\mathbf{z}; \mu^{(i)}, (\sigma^{(i)})^2 \mathbf{I})$$

where the $\mu^{(i)}$ and $\sigma^{(i)}$ are the output of the encoding MLP

Example : Variational Auto-Encoder

DO IT IN PRACTICE

$$\begin{aligned} \mathbf{z}^{(i,l)} &= g_\phi(\mathbf{x}^{(i)}, \epsilon^{(l)}) \\ &= \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)} \end{aligned}$$

where $\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

1. In this model both both $p_\theta(\mathbf{z})$ (the prior) and $q_\phi(\mathbf{z}|\mathbf{x})$ are Gaussian;

Example : Variational Auto-Encoder

Solution of $-KL(q_\phi(\mathbf{z})||p_\theta(\mathbf{z}))$, Gaussian case

$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$, and $q_\phi(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \sigma^2\mathbf{I})$

$$-KL(q_\phi(\mathbf{z})||p_\theta(\mathbf{z})) \quad = \quad \frac{1}{2}\sum_{j=1}^{J}(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2)$$

Example : Variational Auto-Encoder

The resulting estimator for this model and datapoint $\mathbf{x}^{(i)}$

$$
\begin{aligned}
\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) &\approx \frac{1}{2}\sum_{j=1}^{J}(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2) \\
&+ \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)})
\end{aligned}
$$

where $\mathbf{z}^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)}$, and $\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

- the decoding term $p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)})$ is a Bernoulli or Gaussian MLP, depending on the type of data we are modelling.

Example : Variational Auto-Encoder

MLP's as probabilistic encoders and decoders

1. encoder : MLP with Gaussian output;
2. decoder : MLPs with either Gaussian or Bernoulli outputs, depending on the type of data.

Example : Variational Auto-Encoder

Bernoulli MLP as decoder

- Recall Bernoulli : $p^k(1-p)^{1-k}$

$$\log p(\mathbf{x}|\mathbf{z}) \;=\; \sum_{i=1}^{D} x_i \log y_i + (1-x_i) \cdot \log(1-y_i)$$

where

$$\mathbf{y} \;=\; f_\sigma(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2)$$

Example : Variational Auto-Encoder

Gaussian MLP as decoder

$$\log p(\mathbf{x}|\mathbf{z}) = \log \mathcal{N}(\mathbf{x}; \mu, \sigma^2 \mathbf{I})$$

where

$$\begin{aligned} \mu &= \mathbf{W}_4 \mathbf{h} + \mathbf{b}_4 \\ \log \sigma^2 &= \mathbf{W}_5 \mathbf{h} + \mathbf{b}_5 \\ \mathbf{h} &= \tanh(\mathbf{W}_3 \mathbf{z} + b_3) \end{aligned}$$

Example : Variational Auto-Encoder

Gaussian MLP as encoder

$$\log p(\mathbf{z}|\mathbf{x}) \quad = \quad \log \mathcal{N}(\mathbf{z}; \mu, \sigma^2\mathbf{I})$$

where

$$\begin{aligned} \mu &= \mathbf{W}_7\mathbf{h} + \mathbf{b}_7 \\ \log \sigma^2 &= \mathbf{W}_8\mathbf{h} + \mathbf{b}_8 \\ \mathbf{h} &= \tanh(\mathbf{W}_6\mathbf{z} + b_6) \end{aligned}$$