Probability & Neural Networks

by Charles Blundell

Google DeepMind
27th August 2015

Notation:

θ, ϕ	Parameters of a network
x,y,z	Inputs and outputs of a network
$\mathcal{N}(x\|\mu,\Sigma)$	Gaussian (also known as Normal) distribution with mean μ and variance Σ.
$\mathbb{E}_{p(x)}[f(x)]$	Expectation of f with respect to the distribution p. The same as: $\sum_x p(x)f(x)$ or $\int f(x)p(x)\,\mathrm{d}x$.
$d\cdot\epsilon$	Pointwise multiplication of two vectors the ith element of the resulting vector is the product of the ith element of d with the ith element of ϵ.

1 Neural Autoregressive Density Estimator

"The Neural Autoregressive Distribution Estimator", Larochelle & Murray, ICML 2011.

http://jmlr.csail.mit.edu/proceedings/papers/v15/larochelle11a/larochelle11a.pdf

1. Download the MNIST digits data set from http://yann.lecun.com/exdb/mnist/. Binarize the images of the data set by thresholding.

2. Implement NADE. See Algorithm 1 in the above paper for details.

3. Train NADE on the MNIST digit images and then generate examples.

2 Variational Weight Uncertainty

"Practical Variational Inference for Neural Networks", Graves, NIPS 2011.

http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf

"Weight Uncertainty in Neural networks", Blundell et al, ICML 2015

http://jmlr.org/proceedings/papers/v37/blundell15.pdf

1. Use Jensen's inequality to show that, if $q(\theta)\neq 0$,

$$\log p(y|x) \;\geqslant\; \mathbb{E}_{q(\theta)}[\log p(y|x,\theta)p(\theta) - \log q(\theta)].$$

2. Now show that if

$$q(\theta) \;=\; \mathcal{N}(\theta|\mu, D)$$

where μ is the mean vector and D is a diagonal covariance matrix, then

$$\begin{aligned}\mathbb{E}_{q(\theta)}[\log p(y|x,\theta)p(\theta) - \log q(\theta)] \;&=\; \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}[\log p(y|x,\theta)p(\theta) - \log q(\theta)]\\ \text{where}\;\; \theta \;&=\; \mu + D^{\frac{1}{2}}\epsilon.\end{aligned}$$

3. Show that if

$$
\begin{aligned}
p(\theta) &= \mathcal{N}(\theta|0, \sigma^2 I) \\
q(\theta) &= \mathcal{N}(\theta|\mu, D) \\
D &= \text{diag}([e^{2d_1}, e^{2d_2}, ..., e^{2d_p}]) \\
\mathcal{F}(\mu, D) &= \mathbb{E}_{q(\theta)}[\log p(x|\theta)p(\theta) - \log q(\theta)]
\end{aligned}
$$

then if $\theta = \mu + D^{\frac{1}{2}}\epsilon$

$$
\begin{aligned}
\frac{\partial\theta}{\partial\mu_i} &= 1 \\
\frac{\partial\theta}{\partial d_i} &= e^{d_i}\epsilon_i
\end{aligned}
$$

and so

$$
\begin{aligned}
\frac{\partial\mathcal{F}(\mu, D)}{\partial\mu_i} &= \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\frac{\partial\log p(y|x,\theta)}{\partial\theta} + \frac{\partial\log p(\theta)}{\partial\theta} - \frac{\partial\log q(\theta)}{\partial\theta}\right] \\
\frac{\partial\mathcal{F}(\mu, D)}{\partial d_i} &= \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\left(\frac{\partial\log p(y|x,\theta)}{\partial\theta} + \frac{\partial\log p(\theta)}{\partial\theta} - \frac{\partial\log q(\theta)}{\partial\theta}\right)e^{d_i}\epsilon_i\right].
\end{aligned}
$$

4. What is the fewest number of times backpropagation be run to compute unbiased estimates both $\frac{\partial\mathcal{F}(\mu, D)}{\partial\mu_i}$ and $\frac{\partial\mathcal{F}(\mu, D)}{\partial d_i}$?

5. Implement Bayes by Backprop

   a) Generate a simple regression data set by sampling points from the curve:

$$
\begin{aligned}
y &= x + 0.3\sin(2\pi(x+\upsilon)) + 0.3\sin(4\pi(x+\upsilon)) + \upsilon \\
\upsilon &\sim \mathcal{N}(0, 0.02).
\end{aligned}
$$

   b) Construct a neural network that represents $p(y|x,\theta)$.

      Architecture 1-50-50-1.

      Output will be the mean of the Gaussian likelihood.

   c) Train a plain neural network on these data, plot the results.

   d) Find or write another network that represents the Gaussian likelihood (for $p(\theta)$ and $q(\theta)$).

   e) Construct the loss function $\mathcal{F}(\mu, D)$ and is derivative as above.

   f) Train a variational Bayesian neural network on these data. During training plot:

      i. An estimate of the loss function $\mathcal{F}(\mu, D)$.

      ii. The likelihood $\log p(y|x,\theta)$.

      iii. The surprise $\log p(\theta) - \log q(\theta)$

      iv. Predictions on a test set (say 1000 values uniformly spaced between 0 and 1). For each prediction, sample the weights multiple times to build an estimate of the mean and variance. Plot both as well as the true value.

3 Variational Autoencoders

"Stochastic Backpropagation and Approximate Inference in Deep Generative Models", Rezende et al, ICML 2014.

http://arxiv.org/pdf/1401.4082v3.pdf

"Auto-Encoding Variational Bayes", Kingma & Welling, ICLR 2014.

http://arxiv.org/pdf/1312.6114v10.pdf

1. Use Jensen's inequality to show that, if $q(z|x) \neq 0$,

$$\log p(x) \geqslant \mathbb{E}_{q(z|x)}[\log p(x|z)p(z) - \log q(z|x)]$$

2. Show that if $\mu_\theta(x)$ and $\sigma_\phi(x)$ are the encoding neural networks with parameters θ and ϕ, respectively, taking as input x and

$$\begin{aligned} p(z) &= \mathcal{N}(z|0, I) \\ q(z|x) &= \mathcal{N}(z|\mu_\theta(x), \sigma_\phi(x)I) \\ \mathcal{F}(\theta, \phi) &= \mathbb{E}_{q(z|x)}[\log p(x|z)p(z) - \log q(z|x)] \end{aligned}$$

then

$$\begin{aligned} \mathcal{F}(\theta, \phi) &= \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}[\log p(x|z)p(z) - \log q(z|x)] \\ z &= \mu_\theta(x) + \sigma_\phi(x) \cdot \epsilon \\ \frac{\partial z}{\partial \theta} &= \frac{\partial \mu_\theta(x)}{\partial \theta} \\ \frac{\partial z}{\partial \phi} &= \frac{\partial \sigma_\phi(x)}{\partial \phi}\epsilon \end{aligned}$$

3. Show that the gradients of the free energy $F(\theta, \phi)$ are then:

$$\begin{aligned} \frac{\partial \mathcal{F}(\theta, \phi)}{\partial \theta_i} &= \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\left(\frac{\partial \log p(x|z)}{\partial z} + \frac{\partial \log p(z)}{\partial z} - \frac{\partial \log q(z|x)}{\partial z}\right)\frac{\partial \mu_\theta(x)}{\partial \theta_i}\right] \\ \frac{\partial \mathcal{F}(\theta, \phi)}{\partial \phi_i} &= \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\left(\frac{\partial \log p(x|z)}{\partial z} + \frac{\partial \log p(z)}{\partial z} - \frac{\partial \log q(z|x)}{\partial z}\right)\frac{\partial \sigma_\phi(x)}{\partial \phi_i}\epsilon_i\right]. \end{aligned}$$

4. What is the fewest number of times backpropagation be run to compute unbiased estimates both $\frac{\partial \mathcal{F}(\theta, \phi)}{\partial \theta_i}$ and $\frac{\partial \mathcal{F}(\theta, \phi)}{\partial \phi_i}$ on each of the networks?

5. Implement a variational autoencoder

   a) Download the MNIST digits data set from http://yann.lecun.com/exdb/mnist/. Binarize the images of the data set by thresholding.

   b) Implement the encoding network that maps the input digit x to the mean $\mu_\theta(x)$ and standard deviation $\sigma_\phi(x)$. The architecture should have two hidden layers of 200 ReLU units and an output size of 200.

   c) Implement the decoding network, $p(x|z)$, that maps the encoded digit z onto the generated digit x. Let's make it have the same architecture as the encoding network, in reverse (although this is not necessary in general).

d) Find or write another network that represents the Gaussian prior $p(z) = \mathcal{N}(z|0, \mathcal{I})$, and also the encoding distribution $q(z|x) = \mathcal{N}(z|\mu_\theta(x), \sigma_\phi(x))$.

e) Construct the loss function $\mathcal{F}(\theta, \phi)$ and is derivative as above.

f) Train the variational autoencoder on the MNIST digits. During training plot

   i. An estimate of the loss function $\mathcal{F}(\mu, D)$.

   ii. The likelihood of the current encoding $\log p(x|z)$.

   iii. The surprise $\log p(z) - \log q(z|x)$.

   iv. Also show training images x along with their reconstruction after being encoded as z. Note you may only want to do this from time-to-time as it could slow down training.