Macro NN architecture

Machine Learning and Data Mining

Maxim Borisyak

National Research University Higher School of Economics (HSE)

November 14, 2017

Generative models

Generative models

- Informally, given samples we wish to learn generative procedure.
- Formally, given samples of a random variable X, we wish to find X', so that:

$$P(X) \approx P(X')$$

Sampling generative models

- direct sampling procedure, usually in form:

$$\begin{aligned} X &= f(Z); \\ \\ Z &\sim U^n[0,1] \\ &\text{or} \\ Z &\sim \mathcal{N}^n[0,1]; \end{aligned}$$

- density is usually unknown, since:

$$p(x) = \sum_{z \mid f(z) = x} p(z) \left| \frac{\partial}{\partial z} f(z) \right|^{-1}$$

Density generative models

- density function $P(x)$ or unnormalized density function:

$$P(x) = \frac{1}{C}\rho(x)$$

- sampling is usually done via some kind of Monte-Carlo Markov Chains (possible for unnormalized density).

Boltzmann machines

Energy models

$$P(x) = \frac{1}{Z} \exp(-E(x))$$

where:

- $E(x)$ - **energy function**;
- $Z = \sum_x \exp(-E(x))$ - normalization constant, **partition function**.

Latent variables

- one of the simplest way to model a complex distribution is via hidden or *latent* variables h:

$$
\begin{aligned}
P(x,h) &= \frac{1}{Z}\exp(-E(x,h));\\
P(x) &= \frac{1}{Z}\exp(-E(x));\\
E(x) &= \mathrm{FreeEnergy}(x) = -\log\sum_h \exp(-E(x,h));\\
Z &= \sum_x \exp(-\mathrm{FreeEnergy}(x)).
\end{aligned}
$$

Maximum Likelihood fit

$$\mathcal{L} = \sum_i \log P(x_i) \to \max;$$

$$\begin{aligned}
&\frac{\partial}{\partial\theta} \log P(x) = \\
&\qquad \frac{\partial}{\partial\theta} \log\left[\frac{1}{Z}\exp(-\mathrm{FreeEnergy}(x))\right] = \\
&\qquad -\frac{\partial}{\partial\theta}\log Z - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x) = \\
&\qquad -\frac{1}{Z}\frac{\partial}{\partial\theta}Z - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x) = \\
&-\frac{1}{Z}\frac{\partial}{\partial\theta}\left[\sum_\chi \exp(-\mathrm{FreeEnergy}(\chi))\right] - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x)
\end{aligned}$$

Maximum Likelihood fit

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\log P(x) = \\
&-\frac{1}{Z}\frac{\partial}{\partial\theta}\left[\sum_{\chi}\exp(-\mathrm{FreeEnergy}(\chi))\right] - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x) = \\
&\sum_{\chi}\frac{1}{Z}\exp(-\mathrm{FreeEnergy}(\chi))\frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(\chi) - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x) = \\
&\sum_{\chi}P(\chi)\frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(\chi) - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x)
\end{aligned}
$$

Maximum Likelihood fit

$$\frac{\partial}{\partial \theta} \log P(x) = \sum_{\chi} P(\chi) \frac{\partial}{\partial \theta} \mathrm{FreeEnergy}(\chi) - \frac{\partial}{\partial \theta} \mathrm{FreeEnergy}(x)$$

$$\mathbb{E}_x \left[\frac{\partial}{\partial \theta} \log P(x) \right] = \mathbb{E}_\chi \left[\frac{\partial}{\partial \theta} \mathrm{FreeEnergy}(\chi) \right] - \mathbb{E}_x \left[\frac{\partial}{\partial \theta} \mathrm{FreeEnergy}(x) \right]$$

where:

- x - sampled from `real' data;
- χ - sampled from current model.

Maximum Likelihood fit

$$\Delta\theta \sim \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(\chi) - \frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x)$$

Energy model can be trained by:

- sampling x from given data;
- sampling χ from the current model;
- following difference between deriviatives of $\mathrm{FreeEnergy}$.

This is known as **contrastive divergence**.

Latent variables

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\mathrm{FreeEnergy}(x) = \\
&\qquad -\frac{\partial}{\partial\theta}\left[\log\sum_h \exp(-E(x,h))\right] = \\
&\qquad \frac{1}{\sum_h \exp(-E(x,h))}\left[\sum_h \exp(-E(x,h))\frac{\partial}{\partial\theta}E(x,h)\right] = \\
&\qquad \frac{1}{\frac{1}{Z}\sum_h \exp(-E(x,h))}\left[\frac{1}{Z}\sum_h \exp(-E(x,h))\frac{\partial}{\partial\theta}E(x,h)\right] = \\
&\qquad \frac{1}{\sum_h P(x,h)}\left[\sum_h P(x,h)\frac{\partial}{\partial\theta}E(x,h)\right] = \\
&\qquad \mathbb{E}_h\left[\frac{\partial}{\partial\theta}E(x,h)\,\middle|\,x\right]
\end{aligned}
$$

Maximum Likelihood fit

$$\Delta\theta \sim \frac{\partial}{\partial\theta}\mathrm{Energy}(\chi, h') - \frac{\partial}{\partial\theta}\mathrm{Energy}(x, h)$$

Energy model can be trained by:

- sampling x from given data and sampling h from $P(h \mid x)$;
- sampling χ from the current model and sampling h' from $P(h \mid \chi)$;
- following difference between deriviatives of Energy.

Gibbs sampling

Sampling $x = (x^1, x^2, \ldots, x^n) \in \mathbb{R}^n$ *from* $P(x)$.

Repeat until the end of the time:

- for i in $1, \ldots, n$:
  - $x^i := \text{sample from } P(X^i \mid X^{-i} = x^{-i})$

where:

- x^{-i} - all components of x except i-th.

Boltzmann machine

Model with energy function:

$$E(x,h) \quad = \quad -b^T x - c^T h - h^T W x - x^T U x - h^t V h;$$

is called **Boltzmann machine**.

If $\mathrm{diag}(U) = 0$ and $\mathrm{diag}(V) = 0$ then x and h are binomial: $x_i, h_j \in \{0, 1\}$.

Training Boltzmann machine

Let $s = (x, h)$, then:

$$E(s) = -d^T s - s^T A s$$

then for binomial units:

$$P\left(s^i = 1 \mid S^{-i} = s^{-i}\right) =$$

$$\frac{\exp\left(-E\left(s^i = 1, s^{-i}\right)\right)}{\exp\left(-E\left(s^i = 1, s^{-i}\right)\right) + \exp\left(-E\left(s^i = 0, s^{-i}\right)\right)} =$$

$$\sigma\left(d_i + 2a^{-i} s^{-i}\right)$$

where:

- a^{-i} - i-th row without i-th element;
- $\sigma(x)$ - sigmoid function.

Training Boltzmann machine

Positive phase:

- sample x from real data;
- perform Gibbs sampling of h under fixed x;

Negative phase:

- init Gibbs chain with x;
- sample both χ and h' from the model.

$$\Delta\theta = \frac{\partial}{\partial\theta}\mathrm{Energy}(x, h) - \frac{\partial}{\partial\theta}\mathrm{Energy}(\chi, h')$$

Boltzmann machine: discussion

- two MCMC chains (positive and negative) for each step of SGD;
- training is slow...

Restricted Boltzmann machine

Product of experts

Consider energy function in form of **product of experts**:

$$E(x,h) = -\beta(x) + \sum_i \gamma(x,h_i)$$

$$\begin{aligned} P(X) = \frac{1}{Z}\sum_h \exp(-E(x,h)) = \\ \frac{1}{Z}\sum_h \exp(\beta(x))\exp(-\sum_i \gamma(x,h_i)) = \\ \frac{1}{Z}\exp(\beta(x))\sum_h \prod_i \exp(-\gamma(x,h_i)) = \\ \frac{1}{Z}\exp(\beta(x))\prod_i \sum_{h_i} \exp(-\gamma(x,h_i)). \end{aligned}$$

Product of experts

Consider energy function in form of **product of experts**:

$$\begin{aligned} E(x,h) &= -\beta(x) + \sum_i \gamma(x,h_i); \\ \mathrm{FreeEnergy}(x) &= -\beta(x) - \sum_i \log \sum_{h_i} \exp(-\gamma(x,h_i)). \end{aligned}$$