TTIC 31230, Fundamentals of Deep Learning

David McAllester, April 2017

Variational Autoencoders

In a variational autoencoder a distribution on x is modeled by

$$P_{\Theta,\Psi}(x,z) \;=\; P^{\mathrm{gen}}_{\Theta}(z)\; P^{\mathrm{dec}}_{\Psi}(x|z)$$

$$P_{\Theta,\Psi}(x) \;=\; \mathrm{E}_{z\sim P^{\mathrm{gen}}_{\Theta}}\left[P^{\mathrm{dec}}_{\Psi}(x|z)\right]$$

We would like

$$\Theta^*,\Psi^* = \operatorname*{argmin}_{\Theta,\Psi}\; \mathrm{E}_{x\sim D}\left[\log\frac{1}{P_{\Theta,\Psi}(x)}\right] = \operatorname*{argmin}_{\Theta,\Psi} H(D, P_{\Theta,\Psi})$$

However, computing $P_{\Theta,\Psi}(x)$ by sampling z from $P^{\mathrm{gen}}_{\Theta}$ is very inefficient as z is rarely compatible with x.

The Variational Lower Bound

It would be more efficient to sample z from $P_{\Theta,\Psi}(z|x)$. We approximate this marginal by an encoding distribution $P^{\mathrm{enc}}_{\Phi}(z|x)$.

$$\Phi^*, \Theta^*, \Psi^* = \operatorname*{argmin}_{\Phi,\Theta,\Psi} \mathrm{E}_{x\sim D}\left[-\mathcal{L}(x, \Phi, \Theta, \Psi)\right]$$

$$\log \frac{1}{P_{\Theta,\Psi}(x)} \leq -\mathcal{L}(x, \Phi, \Theta, \Psi)$$

$$\mathcal{L}(x, \Phi, \Theta, \Psi) = \mathrm{E}_{z\sim P^{\mathrm{enc}}_{\Phi}(\cdot|x)}\left[\log P_{\Theta,\Psi}(x, z)\right] + H(P^{\mathrm{enc}}_{\Phi}(\cdot|x))$$

For Gaussian distributions gradient estimation through sampling is now feasible.

The Variational Lower Bound

$$
\begin{aligned}
\mathcal{L}(x, \Phi, \Theta, \Psi) &= \mathrm{E}_{z \sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\log P_{\Theta,\Psi}(x, z)\right] + H(P_\Phi^{\mathrm{enc}}(\cdot|x)) \\
&= \mathrm{E}_{z \sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\log P_{\Theta,\Psi}(x) P_{\Theta,\Psi}(z|x)\right] + H(P_\Phi^{\mathrm{enc}}(\cdot|x)) \\
&= \mathrm{E}_{z \sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\log \; P_{\Theta,\Psi}(x)\right] + \mathrm{E}_{z \sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\log P_{\Theta,\Psi}(z|x)\right] + H(P_\Phi^{\mathrm{enc}}(\cdot|x)) \\
&= \log P_{\Theta,\Psi}(x) + \mathrm{E}_{z \sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\log P_{\Theta,\Psi}(z|x) - \log(P_\Phi^{\mathrm{enc}}(z|x))\right] \\
&= \log P_{\Theta,\Psi}(x) - KL(P_\Phi^{\mathrm{enc}}(\cdot|x), P_{\Theta,\Psi}(\cdot|x)) \\
&\leq \log P_{\Theta,\Psi}(x)
\end{aligned}
$$

Consistency Theorem

$$
\begin{aligned}
\Phi^*, \Theta^*, \Psi^* &= \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}}\ \mathrm{E}_{x \sim D}\left[-\mathcal{L}(x, \Phi, \Theta, \Psi)\right] \\
&= \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}}\ \mathrm{E}_{x \sim D}\left[\log \frac{1}{P_{\Theta,\Psi}(x)}\right] + KL(P_\Phi^{\mathrm{enc}}(\cdot|x), P_{\Theta,\Psi}(\cdot|x)) \\
&= \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}}\ H(D, P_{\Theta,\Psi}) + KL(P_\Phi^{\mathrm{enc}}(\cdot|x), P_{\Theta,\Psi}(\cdot|x))
\end{aligned}
$$

Consistency Theorem: If for all Θ and Ψ there exist Φ such that $P_\Phi^{\mathrm{enc}}(z|x) = P_{\Theta,\Psi}(z|x)$ then

$$
\Theta^*, \Psi^* = \underset{\Theta,\Psi}{\operatorname{argmin}}\ H(D, P_{\Theta,\Psi})
$$

Sampling from Variational Autoencoders

For Gaussian distributions samples from a variational autoencoder appear blurry.

[Alec Radford]

END