TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

The Evidence Lower Bound (ELBO)

and Variational Auto Encoders (VAEs)

Modeling y

We would like to use the fundamental equation

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{y \sim \mathrm{Pop}} \; - \ln P_\Phi(y)$$

But even when $P_\Phi(z)$ and $P_\Phi(y|z)$ are samplable and computable we cannot typically compute $P_\Phi(y)$.

Specifically, for $P_\Phi(y)$ defined by a generator we cannot compute $P_\Phi(y)$ for a test image y.

Interpretable Latent Variables

$$P_\Phi(y) = \sum_z P_\Phi(z) P_\Phi(y|z) = E_{z \sim P_\Phi(z)} P_\Phi(y|z)$$

$P_\Phi(z)$ is called the prior.

Given an observation of y (the evidence) $P_\Phi(z|y)$ is called the posterior.

Variational Bayesian inference involves approximating the posterior.

The Evidence Lower Bound (The ELBO)

To model y (the evidence about z) we introduce a samplable and computable model $\hat{P}_\Phi(z|y)$ to approximate $P_\Phi(z|y)$.

$$
\begin{aligned}
\ln P_\Phi(y) &= E_{z\sim\hat{P}_\Phi(z|y)} \ln \frac{P_\Phi(y)P_\Phi(z|y)}{P_\Phi(z|y)} \\
&= E_{z\sim\hat{P}_\Phi(z|y)} \left(\ln \frac{P_\Phi(z,y)}{\hat{P}_\Phi(z|y)} + \ln \frac{\hat{P}_\Phi(z|y)}{P_\Phi(z|y)} \right) \\
&= \left(E_{z\sim\hat{P}_\Phi(z|y)} \ln \frac{P_\Phi(z,y)}{\hat{P}_\Phi(z|y)} \right) + KL(\hat{P}_\Phi(z|y), P_\Phi(z|y)) \\
&\geq E_{z\sim\hat{P}_\Phi(z|y)} \ln \frac{P_\Phi(z,y)}{\hat{P}_\Phi(z|y)} \quad \text{The ELBO}
\end{aligned}
$$

The Variational Auto-Encoder (VAE)

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; E_{y \sim \mathrm{Pop},\, z \sim \hat{P}_\Phi(z|y)} \; - \ln \frac{P_\Phi(z) P_\Phi(y|z)}{\hat{P}_\Phi(z|y)}$$

EM is Alternating Optimization of the ELBO

Expectation Maximimization (EM) applies in the (highly special) case where the exact posterior $P_\Phi(z|y)$ is samplable and computable. EM alternates exact optimization of $\hat{P}_\Phi$ and P_Φ.

$$\text{VAE: } \Phi^* = \text{argmin}_\Phi \, E_{y\sim\text{Train},\, z\sim\hat{P}_\Phi(z|y)} \quad -\ln \frac{P_\Phi(z,y)}{\hat{P}_\Phi(z|y)}$$

$$\text{EM: } \Phi^{t+1} = \text{argmin}_\Phi \;\; E_{y\sim\text{Train}} \, E_{z\sim P_{\Phi^t}(z|y)} \;\; -\ln P_\Phi(z,y)$$

Update (M Step)	Inference (E Step)
Hold $\hat{P}$ fixed	$\hat{P}(z\|y) = P_{\Phi^t}(z\|y)$

RDAs vs. VAEs

Noisy Channel RDA:

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{y,z\sim P_\Phi(z|y)} \; - \ln \frac{\hat{P}_\Phi(z)}{P_\Phi(z|y)} + \lambda \mathrm{Dist}(y, y_\Phi(z))$$

VAE:

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{y,z\sim \hat{P}_\Phi(z|y)} \; - \ln \frac{P_\Phi(z)}{\hat{P}_\Phi(z|y)} \; - \; \ln P_\Phi(y|z)$$

The RDA is a bi-criterion (rate and distortion) while the VAE has a single objective.

But otherwise they are extremely similar.

$-\ln p_\Phi(y|z)$ as Distortion

For $p_{\Phi,\sigma}(y|z) \propto \exp(-||y - y_\Phi(z)||^2/(2\sigma^2))$ we get

$$\Phi^*, \sigma^* = \underset{\Phi,\sigma}{\operatorname{argmin}} E_{y,\epsilon} \; -\ln \frac{p_\Phi(z)}{q_\Phi(z|y)} - \ln p_{\Phi,\sigma}(y|z)$$

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} E_{y,\epsilon} \; -\ln \frac{p_\Phi(z)}{q_\Phi(z|y)} + \left(\frac{1}{2\sigma^{*2}}\right) ||y - y_\Phi(z)||^2 + d \ln \sigma^*$$

where

d is the dimension of y and $\sigma^* = \sqrt{\frac{1}{d} E_{y,\epsilon} \, ||y - y_\Phi(z)||^2}$

Here we have L_2 distortion but no rate-distortion bi-criterion.

END