Variational Autoencoders Derivation

Jacob Yeung

August 28, 2020

Note detailing derivation of variational autoencoders (VAEs). Also first attempt at writing notes in LaTeX.

1 Brief Background

Autoencoders are a technique for dimensionality reduction, akin to PCA, but allow for more complex features because of non-linearities that can be utilized. We can use a neural network to reduce dimensionality to prevent memorization of input (identity mapping). There are several types of autoencoders - basic, sparse, denoising. This note focuses on VAEs.

2 Derivation

2.1 Basic Probability

- Information := $-\log(P(x))$
  - This makes intuitive sense - consider x describes probability of my dog crying and not crying. Probability of x occurring is 1, which gives me no useful information.
  - Low prob. x gives lots info
- Entropy := $-\sum P(x)\log P(x)$
- Kullback–Leibler (KL) divergence $\mathrm{KL}(p \parallel q) = -\sum p(x)\log\frac{p(x)}{q(x)}$
  - KL divergence tells us how similar two probability distributions are w.r.t. first distr. - similar to measuring difference between two distr.
  - Intuitive first step: $-\sum q(x)\log q(x) + \sum p(x)\log p(x)$ (incorrect)
  - Tweak $-\sum p(x)\log q(x) + \sum p(x)\log p(x)$ (correct since distr. q w.r.t. p)
  - Properties
    - $\mathrm{KL} \geq 0$
    - $\mathrm{KL}(p \parallel q) \neq \mathrm{KL}(q \parallel p)$

2.2 Variational Inference

Suppose we have observation x from hidden variable z. We want to know more about z so we want

$$\begin{aligned} P(z|x) &= \frac{P(x|z)P(z)}{P(x)} \\ &= \frac{P(x,z)}{P(x)} \end{aligned}$$

However, in most cases $P(x)$ is intractable, so we want to approximate $P(z|x)$ with $q(z)$, a known tractable distribution.

2.3 Minimize KL divergence

We want to minimize KL divergence of below eq since this creates the best approximation.

$$\begin{aligned} \mathrm{KL}(q(z) \parallel p(z|x)) &= -\sum_z q(z) \log \frac{p(z|x)}{q(z)} \\ &= -\sum_z q(z) \log \frac{\frac{p(x,z)}{p(x)}}{q(z)} \\ &= -\sum_z q(z) \log \frac{p(x,z)}{q(z)} \frac{1}{p(x)} \\ &= -\sum_z q(z) \log \frac{p(x,z)}{q(z)} + \sum_z q(z) \log p(x) \\ &= -\sum_z q(z) \log \frac{p(x,z)}{q(z)} + \log p(x) \end{aligned}$$

Rearrange to find constant in terms of distr. dependents

$$\begin{aligned} \log p(x) &= \mathrm{KL}(q(z) \parallel p(z|x)) + \sum_z q(z) \log \frac{p(x,z)}{q(z)} \\ &= \mathrm{KL}(q(z) \parallel p(z|x)) + \mathcal{L} \end{aligned}$$

We define our second term as the variational lower bound since $\mathcal{L} \leq \log p(x)$.

2.4 Maximizing Variational Lower Bound

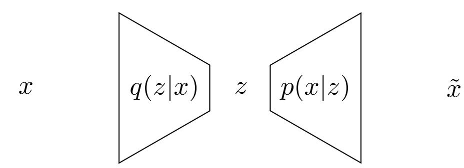
Since $p(x)$ is a constant, minimzing our KL divergence is equivalent to maximizing $\mathcal{L}$.

$$\begin{aligned}
\mathcal{L} &= \sum q(z) \log \frac{p(x,z)}{q(z)} \\
&= \sum q(z) \log \frac{p(x|z)p(x)}{q(z)} \\
&= \sum q(z) \log p(x|z) + \sum q(z) \log \frac{p(z)}{q(z)} \\
&= \mathbb{E}_{q(z)} p(x|z) - KL(q(z) \parallel p(z))
\end{aligned} \tag{1}$$

Thus, we want to maximize the expectation and minimize the KL divergence.

3 Application to VAEs

We can treat our decoder as $q(z|x)$ mapping x to z, and our encoder as $p(x|z)$ mapping z to $\tilde{x}$ where z is a vector of the latent variables.



The encoder is deterministic, so $p(x|z) \approx p(x|\tilde{x})$.

3.1 Gaussian Example

Let us assume $p(x)$ is roughly Gaussian. Then

$$p(x|\tilde{x} = e^{-|x-\tilde{x}|^2}$$

The reconstruction error is as follows.

$$\mathbb{E}_{q(z)} = -|x - \tilde{x}|^2$$

Now we substitute the reconstruction error back into our lower bound and multiply by -1 to minimize it.

$$\min \mathcal{L} = |x - \tilde{x}|^2 + KL(q(z) \parallel \mathcal{N}(\mu, \Sigma))$$

Now, instead of learning the hidden features directly, the decoder network learns the mean and variance of each hidden feature.

4 Sources

- Ali Ghodsi lecture: https://www.youtube.com/watch?v=uaaqyVS9-rMfeature=youtu.bet=19m42s
- Jeremy Jordan: https://www.jeremyjordan.me/variational-autoencoders/
- Deep Learning Book http://www.deeplearningbook.org/contents/autoencoders.html