再看变分自动编码器 VAE

赖泽强

<laizeqiang@outlook.com>

November 25, 2020

再看变分自动编码器 VAE

赖泽强

2020 年 10 月 21 日

目录

1 Introduction

VAE 时 Variational Autoencoder 的缩写，中文名通常翻译为变分自动编码器。这个模型自从论文 Auto-Encoding Variational Bayes[1] 提出以来，受到了广泛的讨论和应用，并产生了许多变种，如 Conditional VAE 等等。虽然 VAE 中有一个 autoencoder，但 VAE 最初并不是从 autoencoder 引出的，而是推导结果中恰好有类似 autoencoder 的形式。

网络上关于 VAE 的教程很多，不过各个教程侧重的点比较分散，有些点可能这个教程有涉及，但另一个就没有。本文主要是基于 Tutorial on Variational Autoencoders[2] 这篇文章，对 VAE 再次进行一个比较详细的总结。

VAE - Background

- 我们有一堆数据 $X = x_1, x_2, \ldots, x_n$ (训练数据)
- 我们想要**学习这些数据满足的分布**

VAE - Background

- 我们希望找到一个分布使得下面的似然概率最大

$$l(\theta) = P(x_1)P(x_2)\dots P(x_n)$$

- 可以直接对 P(x) 进行建模
- 但我们更希望在一个更紧凑的低维空间 Z 对 x 进行建模
- 学习一个函数 f 将 Z 中隐变量映射到 X 空间中 (以一个分布的形式)

$$P(x) = \int P(x|z)P(z)dz$$

VAE - Background

$$P(x) = \int P(x|z)P(z)dz$$

- $P(x|z)$ 的分布形式是什么?
- $P(z)$ 分布形式如何选取? (VAE 要解决的问题)
- 如何计算这个积分? (VAE 要解决的问题)

VAE - $P(x|z;\theta)$ 的分布

- 分布形式是人工定义的
- 可以随意选取，一般根据输出 X 的形式进行选择，例如输出 X 的取值范围是离散的，则取伯努利分布，连续则取正态分布。
- 分布必须要有确定形式，在分布参数的空间连续可导（这样我们才能梯度下降求解)。
- VAE 取的是正态分布 $P(X \mid z;\theta) = \mathcal{N}\left(X \mid f(z;\theta), \sigma^2 * I\right)$，使用神经网络实现（输出均值即可，方差不使用)

VAE - $P(z)$ 的分布

- z 这个隐变量的真实分布是很复杂的
- 不想要去手工定义 z 的分布形式

VAE 的做法

先从一个简单的分布中对 z 进行取样，比如说标准正态分布 $\mathcal{N}(0, I)$，然后通过一个足够复杂的函数 (可以用神经网络表示) 将这个正态分布映射到其他任何分布

$$P(x \mid z; \theta) = \mathcal{N}\left(x \mid f(z; \theta), \sigma^2 * I\right)$$

- 神经网络实现
- 前几层：将 z 映射到真实分布上
- 后几层：从 z 的真实分布映射到 x 的真实分布上。

VAE - 计算积分 - Intractablility

$$P(x) = \int P(x|z;\theta)P(z)dz$$

- 没有解析解
- z 的隐空间太大，近似算法如蒙特卡洛采样法同样不可行。

VAE 的做法

尝试从更可能产生 X 的 z 中进行 sample，为此，VAE 添加了一个新的函数 Q(z|X)，可以输入一个 x，输出一个可能产生 x 的 z 的分布

VAE - 解决 Intractablility

- 我们现在至少可以使用近似算法算出下面这个期望

$$E_{z\sim Q}P(X|z) = \int P(X|z)Q(z|X)$$

- 但我们希望算的是下面这个

$$P(x) = E_{z\sim P(z)}P(x|z)$$

- 需要想办法把两个式子联系起来

VAE - 核心公式推导

- 首先考虑近似分布 Q(z|X) 和真实分布 P(z|X) 的 KL divergence:

$$\mathcal{D}[Q(z|X)\|P(z \mid X)] = E_{z\sim Q}[\log Q(z|X) - \log P(z \mid X)]$$

- 使用贝叶斯法则引入 $P(X|z;\theta)$ 和 $P(X)$:

$$\mathcal{D}[Q(z|X)\|P(z \mid X)] = E_{z\sim Q}[\log Q(z|X) - \log P(X \mid z) - \log P(z)] + \log P(X)$$

- 重新组合 $\log Q(z|X)$ 和 $\log P(z)$ 成 $\mathcal{D}[Q(z|X)\|P(z)]$，移项得

$$\log P(X) - \mathcal{D}[Q(z \mid X)\|P(z \mid X)] = E_{z\sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X)\|P(z)]$$

Loss 的代码实现

```python
def vae_loss(x, x_gt, mu, log_var):
    reconstruct_loss = F.mse_loss(x, x_gt, reduction='sum')
    kl_divergence = torch.sum(0.5 * (-log_var + mu ** 2 + torch.exp(log_var) - 1))
    return reconstruct_loss + kl_divergence
```

Python

Loss 代码与公式的关系 - KL 散度

$$\log P(X) - \mathcal{D}[Q(z \mid X) \| P(z \mid X)] = E_{z \sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X) \| P(z)]$$

```python
def vae_loss(x, x_gt, mu, log_var):
    reconstruct_loss = F.mse_loss(x, x_gt, reduction='sum')
    kl_divergence = torch.sum(0.5 * (-log_var + mu ** 2 + torch.exp(log_var) - 1))
    return reconstruct_loss + kl_divergence
```

Python

$$KL\left(N\left(\mu, \sigma^2\right) \| N(0,1)\right) = \frac{1}{2}\left(-\log \sigma^2 + \mu^2 + \sigma^2 - 1\right)$$

Loss 代码与公式的关系 - KL 散度

$$
\begin{aligned}
& KL\left(N\left(\mu,\sigma^2\right)\|N(0,1)\right)\\
=& \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\left(\log\frac{e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}}{e^{-x^2/2/\sqrt{2\pi}}}\right)dx\\
=& \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\log\left\{\frac{1}{\sqrt{\sigma^2}}\exp\left\{\frac{1}{2}\left[x^2-(x-\mu)^2/\sigma^2\right]\right\}\right\}dx\\
=& \frac{1}{2}\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\left[-\log\sigma^2+x^2-(x-\mu)^2/\sigma^2\right]dx
\end{aligned}
$$

Loss 代码与公式的关系 - 重构误差

$$\log P(X) - \mathcal{D}[Q(z \mid X) \| P(z \mid X)] = E_{z \sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X) \| P(z)]$$

```python
def vae_loss(x, x_gt, mu, log_var):
    reconstruct_loss = F.mse_loss(x, x_gt, reduction='sum')
    kl_divergence = torch.sum(0.5 * (-log_var + mu ** 2 + torch.exp(log_var) - 1))
    return reconstruct_loss + kl_divergence
```

Python

$$E_{z \sim Q}[\log P(X \mid z)] = \frac{1}{N} \sum_{i=1}^{N} \log P(X_i \mid z)$$

$$logP(x|z) = log(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma}}) = -\frac{(x-\mu)^2}{2\sigma} - log(\sigma\sqrt{2\pi}) \propto -(x-\mu)^2$$