

Introduction

一. Generative model

- 对于训练数据，假设所有样本点均来自一个相同的分布 P_{data} 即 $X_i \sim P_{\text{data}}, i = 1, 2, \dots, n$ ，那么 generative model 的任务在于近似概率分布 P_{data} ，并从其中采样，以取得新的数据
- 因此，定义模型族，即一组不同的概率分布 $P_\theta, \theta \in \Theta$ ，之后，目标就成为了从 P_θ 中寻找 P_{data} 的良好近似
- 定义损失函数 $d(P_{\text{data}}, P_\theta)$ ，这是需要最小化的目标

Q1. How to represent $p(x)$?

① Bernoulli distribution:

- $D = \{A, B\}$

- 令 $P(X=A) = p$ ，则 $P(X=B) = 1-p$

- 记作 $X \sim B(p)$

日期： /

② Categorical distribution :

$$- D = \{1, 2, \dots, m\}$$

$$- \sum P(X=i) = p_i, \text{ 则 } \sum p_i = 1$$

$$- 记作 X \sim \text{Cat}(p_1, p_2, \dots, p_m)$$

[joint distribution]

X_1, X_2, \dots, X_n 的联合概率分布表示为：

$$P(X_1, X_2, \dots, X_n)$$

当他们彼此独立时， $P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2) \cdots P(X_n)$

当 $|Val(S_i)|=2$ 时，表示概率分布
 $P(S_1, S_2, \dots, S_n)$ 需要 $\frac{2^n}{2} = 2^{n-1}$ 个参数

[Chain rule] $\underbrace{P(S_1 \cap S_2 \cap \dots \cap S_n)}_{\text{(自回归模型)}} = P(S_1)P(S_2|S_1)P(S_3|S_1, S_2) \cdots P(S_n|S_1, S_2, \dots, S_{n-1})$

(自回归模型) $P(S_n|S_{n-1}, S_{n-2}, \dots, S_1)$

而当 $X_{n+1} \perp X_{n-1} \perp X_{n-2} \dots \perp X_1$ 时， $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2) \cdots P(X_n|X_{n-1})$

此时参数数量缩小为 $2n-1$ ，但因此假设过强，一个推广方法为使用贝叶斯网络

贝叶斯网将使用条件概率分布代替联合概率分布，对于 X_1 ，假设有 $P(X_1|X_{A_1}), X_{A_1}$ 另一维随机变量，称为父节点，则

$$P(X_1, \dots, X_n) = \prod_i P(X_i|X_{A_i})$$

[Bayes' rule] $P(S_1|S_2) = \frac{P(S_2|S_1)P(S_1)}{P(S_2)}$

日期： /

[Baye's network] 贝叶斯网络的底层是一个有向无环图(DAG),

$G = (V, E)$, 同时有:

① 图中节点 $i \in V$ 对应一个随机变量 X_i ;

② 对于图中每个节点有条件概率分布 $P(X_i | X_{\text{Par}(i)})$,
 $X_{\text{Par}(i)}$ 为 X_i 的父节点

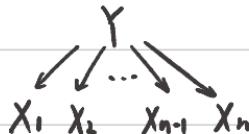
则有 $P(X_1 | X_2 \dots X_n) = \prod_{i \in V} P(X_i | X_{\text{Par}(i)})$

贝叶斯网络在图概率模型(PGM)中有应用

[使用生成式模型的思想进行分类任务]

假设有-一个邮件分类任务,有一系列变量 $X_1, X_2 \dots X_n$, 每个变量表示某个词是否出现在邮件中 ($X_i = 0$ 或 1), 输出为判断
邮件是否为垃圾邮件 ($Y = 0$ 或 1)

如果假设 $X_i \perp X_j$ ($i \neq j$), 则可有 DAG



由贝叶斯网络的思想, $P(y, X_1, X_2 \dots X_n) = P(y) \prod_{i=1}^n P(X_i | Y=y)$

日期： /

根据全概率公式：

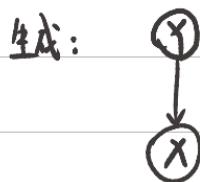
$$P(Y_2 | X_1 \dots X_n) = \frac{P(Y_2 | X_1) \prod_{i=1}^n P(X_i | Y_2)}{\sum_{y_2 \in \text{all}} P(Y_2 | X_1) \prod_{i=1}^n P(X_i | Y_2)}$$

根据贝叶斯公式：

$$P(Y, X) = P(X|Y) P(Y) = P(Y|X) P(X)$$

生成模型和判别模型捕获相同的联合概率分布：

但是其表示不同：



$$P(X|Y) = P(X|Y) P(Y)$$



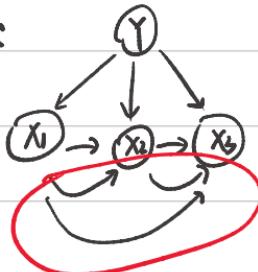
$$P(X|Y) = P(Y|X) P(X)$$

然而，对于左侧，我们只需关注 $P(Y|X)$ ，对于判别任务，了解 $P(X)$ 无意义

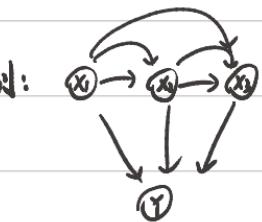
对于左侧，我们需关注 $P(X|Y)$ 和 $P(Y)$
并根据 Bayes' rule 建模 $P(Y|X)$

日期： /

生成：



判别：



判别模型的简化

生成式模型的简化方式是

削弱X人间的依赖关系

方式是假设依赖关系

以函数方式出现

$$Pr(Y_2=1|X; \alpha) = f(X; \alpha)$$

[α 为参数]，即回归

$f(X; \alpha)$ 的确定依赖于假设：

· 线性假设： $Z(\alpha, x) = \alpha_0 + \sum_{i=1}^n \alpha_i x_i$:

$$Pr(Y_2=1|X; \alpha) = \sigma(Z(\alpha, X)), \text{ 其中}$$
$$\sigma(z) = 1/(1+e^{-z}) \quad [\text{logistic function}]$$

由线性假设并不显式地声明 X_i 之间独立，所以其优于朴素贝叶斯

· 非线性假设： $h(A, b, x) = f(Ax + b)$

h 是一个非线性转换，这就是神经网络中的激活函数

对生成模型：

$$P(X_1, X_2, X_3, X_4)$$

$$\approx P(X_1) P(X_1|X_2) P_{\text{Neural}}(X_3|X_1, X_2)$$

$$P_{\text{Neural}}(X_4|X_1, X_2, X_3)$$

日期：

对于连续变量，以概率密度函数描述其分布

$P_x: R \rightarrow R^+$, 固常考虑：

• Gaussian: $X \sim N(\mu, \sigma)$ If $p_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

• Uniform: $X \sim U(a, b)$ If $p_x(x) = \frac{1}{b-a}$ [$a \leq x \leq b$]

若 X 为连续与离散变量的混合，使用联合概率密度函数：

$$\text{Gaussian: } P_x(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Chain rule, Bayes rule 仍适用：

例如：Bayes net $Z \rightarrow X$ 有 $P_{Z,X}(z,x) = P_Z(z)P_{X|Z}(x|z)$

- 当 $Z \sim B(p)$, $X|Z=z \sim N(\mu_0, \sigma_0)$, $X|Z=1 \sim N(\mu_1, \sigma_1)$

时，参数有 $p, \mu_0, \sigma_0, \mu_1, \sigma_1$

- 当 $Z \sim U(a, b)$, $X|Z=z \sim N(\mu_z, \sigma_z)$ 时，参数为 a, b, σ_z

- VAE: $Z \sim N(0, 1)$

$$X|Z=z \sim N(\mu_{\theta}(z), e^{\phi(z)})$$

其中 $\mu_{\theta}: R \rightarrow R$ 和 ϕ 为带有权重 θ, ϕ
的神经网络