

日期： /

## Evaluation

### 1. Density Estimation or Compression

(对数)仍然是密度估计的 metrics:

$$E_{P_{\text{data}}} [\log p_{\theta}(x)]$$

这可以评估模型对数据的压缩能力:

#### [Shannon coding]

Shannon coding 为  $x$  分配长为  $\lceil \log \frac{1}{p_{\theta}(x)} \rceil$  的编码

故平均编码长度为:

$$E_{x \sim P_{\text{data}}} \left[ \lceil \log \frac{1}{p_{\theta}(x)} \rceil \right] \approx E_{P_{\text{data}}} \left[ \log \frac{1}{p_{\theta}(x)} \right] = -E_{P_{\text{data}}} [\log p_{\theta}(x)]$$

Shannon/Huffman 难以实现, 可以使用 Arithmetic Coding 代替

而对于生成式语言模型, 常用的指标为困惑度(perplexity)

$$\text{perplexity} = 2^{-\frac{1}{N} E_{P_{\text{data}}} [\log p_{\theta}(x)]} \quad \text{for } x \in \mathcal{X}$$

这种评估方式的问题在于将所有信息(bit)视为同等重要的。

此外, 部分模型不具有建模似然的能力, 如

极密度估计  
VAEs, GANs, EBMs, SBMs 等。

WS ELBO 代替, 或退火重要性采样

日期： /

## [Kernel Density Estimation] for GANs

$$\hat{p}(x) = \frac{1}{n} \sum_{x^{(i)} \in S} K\left(\frac{x - x^{(i)}}{\sigma}\right)$$

核带宽，控制核的平滑程度  
核函数

例如，对于 Gauss 核， $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$

K 应当是一个非负的归一化函数：

- Normalization :  $\int_{-\infty}^{\infty} K(u) du = 1$
- Symmetric :  $K(u) = K(-u)$

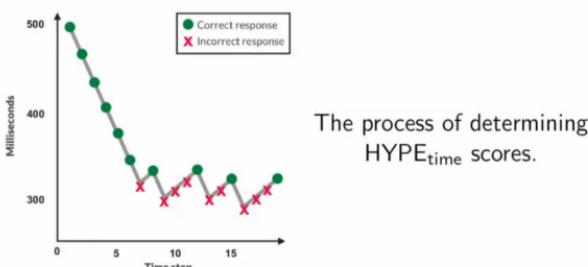
$\sigma$  的选择通常通过交叉验证进行，避免欠平滑/过度平滑

## 2. Sample quality

- 人类评估是金标准（成本大，复现性差，易受干扰，无法评估以化能力）

- HYPE<sub>time</sub>：人类辨认样本真伪的最短时间

- HYPE<sub>∞</sub>：在无限时间内人类能区分的生成样本的比例



日期:

这是因为通常使用 Inception net 作为分类器

## [ Inception scores ]

用于处理有标签数据集。通过观察一个充分优化的分类器在生成数据上的表现评价模型。

Inception scores 关注:

① 清晰度 (Sharpness)

$$S = \exp \left( E_{x \sim p} [ \int c(y|x) \log c(y|x) dy ] \right)$$

② 多样性 (Diversity)

$$D = \exp \left( - E_{x \sim p} [ \int c(y|x) \log c(y|x) dy ] \right)$$

$$c(y) = E_{x \sim p} [ c(y|x) ]$$

$$IS = D \times S$$

## [ Frechet Inception Distance ]

比较由预训练模型学习的特征分布。

- 由模型采样  $G$ , 由数据集采样  $T$
- 由预训练模型 (Inception net 等) 计算表征  $F_G, F_T$
- 对两个分布分别拟合多元高斯分布  $(\mu_G, \Sigma_G), (\mu_T, \Sigma_T)$

日期： /

$$\cdot \text{FID} = \| \mu_T - \mu_G \|^2 + \text{Tr}(\Sigma_T + \Sigma_G - 2(\Sigma_T \Sigma_G)^{-1})$$

### [ Kernel Inception Distance (KID) ]

通过 Maximum Mean Discrepancy (MMD) 进行两样本比较。MMD

通过计算不同阶矩的差异比较分布的不同。

$$\text{MMD}(p, q) = E_{x, x' \sim p}[K(x, x')] + E_{x, x' \sim q}[K(x, x')] - 2 E_{x \sim p, x' \sim q}[K(x, x')]$$

KID并不对原始像素计算KID，而是对预训练分类器隐变量计算。

KID为无偏估计，计算复杂度为  $O(n^2)$

FID为有偏估计 ( $> 0$ )，计算复杂度为  $O(m)$

## 3. Clustering

可以使用 k-means 或任何聚类算法对生成式模型隐空间进行聚类。

对于有标签数据，可以使用标签评价聚类效果。

评价指标有 completeness score (完整性得分)、homogeneity score (同质性得分)、

V measures (V 测度)

日期： /

### [Completeness score (between [0,1])]

用于衡量一个簇中的样本是否属于同一类别，得分越高结果越好。

$$C = 1 - \frac{H(C|k)}{H(c)}$$

其中， $H(C) = -\sum_{i=1}^{n_c} P(c_i) \log P(c_i)$

$$H(C|k) = -\sum_{j=1}^{n_k} p(k_j) \sum_{i=1}^{n_c} P(c_i|k_j) \log P(c_i|k_j)$$

### [Homogeneity score (between [0,1])]

用于衡量某一类别的所有样本是否属于同一族，得分越高聚类效果越好。

$$H = 1 - \frac{H(k|c)}{H(k)}$$

其中， $H(k) = -\sum_{j=1}^{n_k} p(k_j) \log p(k_j)$

$$H(k|c) = -\sum_{j=1}^{n_k} p(k_j) \sum_{i=1}^{n_c} P(c_i|k_j) \log P(c_i|k_j)$$

### [V measure score (between [0,1])]

对同质性、完整性得分的综合。

日期： /

$$V = 2 \cdot \frac{H \cdot C}{H+C}$$

#### 4. Lossy Compression or Reconstruction

这方面的指标衡量模型从隐空间恢复数据的能力。常见指标为 MSE, RMSE, SSIM, PSNR 等

#### 5. Disentanglement (解耦性)

我们希望隐变量的各维度相互独立且可解释。

对未标注数据而言，学习解耦的模型是不可能的。