

Metrics for Generative Models

How to judge quality of generative model samples?

— want $p_{\text{gen}}(x) = p_{\text{data}}(x)$

— 1d summary statistics (histograms of $H(LF_0)$)

— Frechet distance

— Wasserstein - 2 dist b/w P & Q

$$d_F(P, Q) = \min_{\gamma \in \Gamma(P, Q)} \int \|x_{\text{gen}}\|^2 d\gamma(x_{\text{gen}})$$



tractable when P & Q gaussian

Frechet Inception Distance:

~~to~~ Train classifier (Incept-)
on Ingeret

Use last hidden layer as HLFs

Assume HLFs Gaussian (μ, Σ)

↓

FID score btw gen & ref

- Binary classifier test

CIF (gen, ref)

pros: "ultimate test"

cons: classif could be undertrained

if 100% separable not useful

diff classif for each gen model

— hard to use to compare
gen models

- "log posterior
test"

22.11.1765

Train clf on $gen_1, gen_2, \dots, gen_N$

clf score: $p(i|x)$

$$\textcircled{b} \quad \frac{\sum_{j=1}^N p(j|x)}{N}$$

$$LP = \sum_i \sum_{x \in ref} \log p(i|x) - \text{const}$$

↓
product of posteriors for each $x \in ref$
which gen model is it more similar to

note: $\textcircled{LP}$ $\log p(i|x) = \log \frac{p(x|i) p(i)}{p(x)}$

So $LP_i = LL_i + \text{const}$

LP is also LL up to const!

Can use to compare models ✓

Example: 2211.11765