Wednesday, March 31, 2021 10:11 AM

Lecture 18

Generative Adversarial Networks

Idea: train 2 networks — G generator: noise z $\longrightarrow$ data space x

$z \sim P_{simple}(z)$ like uniform

$z \in \mathbb{R}^n$

n is size of latent space

WANT: $x \sim P_{data}(x)$

$x \in \mathbb{R}^d$

D discriminator: $D(x \text{ or } G(z)) \longrightarrow$ class probability

classifier

tries to distinguish real from fake.

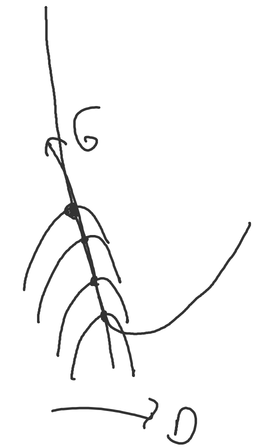
want G to fool D

and D to be as good as possible

Wednesday, March 31, 2021 10:31 AM

loss function: $L = \sum_{i \in \text{data}} \log D(x_i) + \sum_{i \in \text{gen}} \log\left(1 - D(G(z_i))\right)$ ($x_i \in$ generated sample)

usual BCE but negative



Want: $\min_G \max_D L$ ← saddle pt optimization
unstable
could be multiple saddles
convergence is tricky

generator fools discriminator (G) — discriminator is a good classifier (D)

in principle optimality: $\max_D L\big|_G \longrightarrow \hat{D}_G(x) = \dfrac{p_{data}(x)}{p_{data}(x) + p_G(x)}$ (NP lemma)

$p_G(x)$ ← prob distr'n of x coming from generator

$\sum_{x \in data} \approx \int dx\, p_{data}(x)$

$\sum_{x \in gen} \approx \int dx\, p_G(x)$

↓ plug back into L

$$L\Big|_{D=\hat{D}_G} = \sum_{x \in data} \log \frac{p_{data}(x)}{p_{data}(x) + p_G(x)} + \sum_{x \in gen} \log \frac{p_G(x)}{p_{data}(x) + p_G(x)}$$

Wednesday, March 31, 2021 10:40 AM

$$\max_D L_G \rightarrow \int p_{data}(x) \log \frac{p_{data}(x)}{p_{data}(x)+p_G(x)} + \int p_G(x) \log \frac{p_G(x)}{p_{data}(x)+p_G(x)}$$

$$= 2\,JSD(p_{data}, p_G) + \text{const}$$

$$\min_G (\quad) \rightarrow \left(\begin{array}{l} 0 \le JSD \le 1 \\ JSD(P,Q) = 0 \text{ iff } P = Q. \end{array} \right) \rightarrow \boxed{p_{data} = p_G}\,!$$

GAN learns $p_{data}(x)$ implicitly!

We don't ever get $p_{data}(x)$ itself

given x we can't know value of $p_{data}(x)$.

we only get samples.

example: "likelihood ratio trick"

"likelihood-free inference learning"

Wednesday, March 31, 2021 10:40 AM

Traing a GAN:

(freeze generator)

take a batch from real data $\{x\}$

" " " gen. data $\{x'\}$

train discriminator

(freeze discriminator)

take sample from gen. data

train generator.

repeat