Lecture 15

Autoencoders

Sol 4

HW 5 - due yesterday

Sol 5 - Mon 8am

Q2 Tue 11:30

LVI t.b. posted

Autoencoders

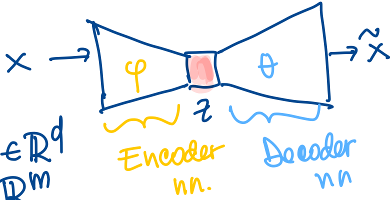
estion How to learn from data without outputs y?
This is unsupervised learning, not prediction

Idea Learn a low dimensional/sparse representation $h(x)$ of data $x \in \mathbb{R}^d$

$$h(x) \in \mathbb{R}^m, \text{ with } m < d \quad f(h(x)) \approx x! \tag{13}$$

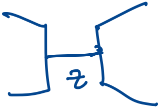
- Optimize $L(x, f(h(x)))$

$d_{\varphi,\theta}(x, \tilde{x}) = \|\tilde{x} - x\|^2$



- train by back prop.

$x, \tilde{x} \in \mathbb{R}^d$
$z \in \mathbb{R}^m$
$m < d$



Variations

- If f linear, L_{LS}, then we "learn" PCA
- Denoising autoencoder
  - Add noise to x input, predict true x

$$\tilde{x} \sim C(\,|x), \quad \min L(x, f(h(\tilde{x}))). \tag{14}$$

$x^{noisy} = x + noise \quad L(x, Dec(Enc(x^{noisy})))$

- Sparse autoencoder

$$\min L(x, f(h(x)) + \Omega(h) \tag{15}$$

φ

sparse Encoder
do not encode noise

Ω is regularization that makes h sparse

- Bias for sparsity

$h(x) = Enc(x) = z$

$\lambda \|z\|_1 = \Omega(h)$

sparsity inducing

many $z_j = 0$

$j = 1:m$

regularization parameter

For each x, only few $z_j(x) \neq 0$

Autoencoders and Variational Autoencoders

VAE

Kingma & Welling

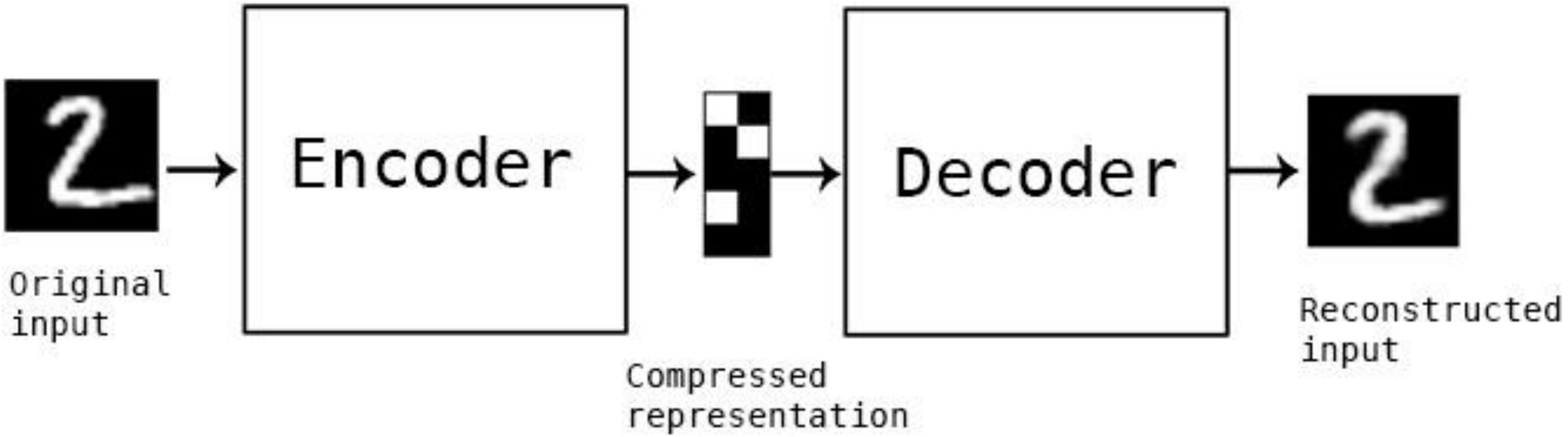
Gautam Kamath

THANKS

Autoencoder

- A type of "compression"
- Input: $x \in \mathbf{R}^d$
- Encoder: $f: \mathbf{R}^d \to \mathbf{R}^m$, Decoder: $g: \mathbf{R}^m \to \mathbf{R}^d$
- Goal: $g\big(f(x)\big) = x$
- Trivial if $m = d$: just let $f(x) = x$ and $g(x) = x$
- Interesting when $m \ll d$ (e.g., $d = 1000, m = 10$)

Autoencoder



Linear Autoencoder

- (Draw simple autoencoder, label weights W_f and W_g and bottleneck)
- Output: $W_g W_f x$
- How to optimize? Use objective

$$\min_{W_f, W_g} \sum_i \frac{1}{2} \left\| W_g W_f x_i - x_i \right\|_2^2$$

- $W_f x$ is a compression of x
- With linear autoencoder, similar to principal component analysis (PCA) (draw)

Nonlinear Autoencoder

- f and g are non-linear (draw non-linear auto-encoder, label W_f, W_g)
- $\min_{W_f, W_g} \sum_i \frac{1}{2} \|g(f(x_i)) - x_i\|_2^2$
- Deep autoencoder (draw)

Other Autoencoders

- Sparse autoencoders
  - Encourage a sparse encoding of input
  - May have wider bottleneck layer (draw)
  - $\min_{W_f, W_g} \sum_i \frac{1}{2} \|g(f(x_i)) - x_i\|_2^2 + \lambda \|f(x_i)\|_1$
- Denoising autoencoders
  - Given noised input $\tilde{x}$, produce denoised x as output (draw)
  - $\min_{W_f, W_g} \sum_i \frac{1}{2} \|g(f(\tilde{x}_i)) - x_i\|_2^2$

Uses of Autoencoders

Decoder = generate new samples

- Can "detach" input and output, use separately Enc.
- Can compress data to a smaller dimension ← dim reduction
- Can find interesting representations of data Encode data
- Generally, finds some underlying structure of the dataset
- However, is not useful to understand *distribution* of dataset
  - In particular, can't necessarily generate new images

Generative Modelling

- Given $X_1, \ldots, X_n \sim D$, can we generate $X_{n+1}, X_{n+2}, \ldots$?
  - Ideally from D, but actually from something *close* to D
- D may be more complex than a GMM
  - E.g., the distribution of all handwritten numbers, or ImageNet (draw)
- Solution: use a neural network to do the work
- Draw a sample from $N(0, I)$, use an NN to map it to a sample from D
- (Draw NN version, where low d Gaussian mapped to high d output)
- Actually: use variational autoencoder (VAE)

Variational Autoencoder

- (Draw encoder, from $x \in \mathbf{R}^d$ to $\mu(x), \sigma(x) \in \mathbf{R}^m$, decoder from $z \sim N\left(\mu(x), \text{diag}\big(\sigma(x)\big)\right) \in R^m$ to $\tilde{x} \in \mathbf{R}^d$)

VAE

Idea 1

$x \mapsto Enc_\varphi(x) = p_\varphi(Z|x)$ Gaussian ---> sample zs

$z \mapsto Dec_\theta(z) = p_\theta(X|z) = [p_\theta(z)]_{j=1:m}$ ← Bernoulli ↓ sample $\tilde{x}$'s

$p = $ Prob (density)

$x \in \{0,1\}^d$

Z, X = random variable

z, x = actual values

$x \mapsto \mu_\varphi(x), \ln\sigma_\varphi(x) \in \mathbb{R}^m$

$p_\varphi(Z|x)$ variational

$x \to [\varphi] \; \mu(x), \ln\sigma(x) \to z = \mu + \sigma\varepsilon \longmapsto [\theta] \; p_\theta(X|z)$ → sample $\tilde{x}$

sample $\varepsilon \sim N(0, I_m)$

$p(x)$ = likelihood of data point x

$p(X)$ = model for X distribution

why? Bayesian → Monte Carlo

Integral $\int dz$ → samples

Gradient → stoch. gradient

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

$m < d$

Step 2 "Likelihoods

$p_\theta(X, Z) = p_\theta(Z) \cdot p_\theta(X|Z)$

(X: $\tilde{x}$; $p_\theta(Z)$: Gaussian $\approx N(0, I)$; $p_\theta(X|Z)$: decoder)

$[p_\theta(X) = \int_{\mathbb{R}^m} p(X, z)\, dz]$

$p_\theta(X) = \frac{p_\theta(X, z)}{p_\theta(z|X)}$, $p_\theta(z|X) \simeq p_\varphi$ ← Ex

Idea 3.1

Idea 3.2 Variational approx

Max Likelihood

$x \in \mathcal{D}$

$$\ln p_\theta(x) = E_\varphi\left[\ln p_\theta(x)\right] = E_\varphi\left[\ln \frac{p_\theta(x,Z)}{p_\theta(Z|x)} \cdot \frac{p_\varphi(Z|x)}{p_\varphi(Z|x)}\right] \quad E_\varphi = E_{z\sim p_\varphi(z|x)}$$

log-Likelihood

$$= E_\varphi\left[\ln p_\theta(x,Z) - \ln p_\varphi(Z|x)\right] + E_\varphi\left[\ln \frac{p_\varphi(Z|x)}{p_\theta(Z|x)}\right]$$

$KL(p_\varphi(Z|x) \,\|\, p_\theta(Z|x)) \geq 0$

ELBO

max θ, φ

samples $z \sim p_\varphi$

min

$KL(p \,\|\, q) = \int \ln \frac{p}{q} p \, dz$

ELBO = Evidence Lower Bound

$\int p_\theta(x,Z)\, dz = \text{evidence}$

Decoder

max θ avg z $\ln p_\theta(x|z)$ -

$$\text{avg}_z \sum_j \left[x_j \ln p_j(z,\theta) + (1-x_j)\ln(1-p_j(z,\theta))\right]$$

max φ

Variational Autoencoder (VAE)

- Some notation: x's live in the *data* space (in $\mathbf{R}^d$), while z's live in the *latent* space (in $\mathbf{R}^m$). p_θ is the decoder network's distribution, q_ϕ is the encoder network's distribution
- E.g., $p_\theta(x)$ is density of decoder network's outputs. $p_\theta(x|z)$ is density of decoder network's outputs, *conditioned on* some latent vector input z. $p_\theta(z)$ is density of decoder network's latent vector input. $q_\phi(z|x)$ is distribution of encoder network's outputs, *conditioned on* some data input x
    - $p_\theta(z)$ generally chosen to be $N(0, I)$
    - Why does $p_\theta(x|z)$ have a distribution? Isn't it deterministic? For loss calculation, we assume the output of the network is fed into a Gaussian sampler. Will revisit shortly.
    - (Draw mapping from data space to latent space and back)

VAE Goals

- Ensure that input image distribution maps to latent distribution $N(0, I)$ (draw)
  - Minimize $KL\left(q_\phi(z|x)||p_\theta(z)\right) = KL\left(q_\phi(z|x)||N(0, I)\right)$ (draw lines)
- Similar to autoencoder, ensure that an input gets encoded and mapped back to itself
  - Maximize $E_{z\sim q_\phi(\cdot|x)}[\log p_\theta(x|z)]$
- Claim: $\log p_\theta(x) \geq E_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL\left(q_\phi(z|x)||N(0, I)\right)$
  - Similar to the inequality when doing EM
  - Bigger picture: variational inference

Optimizing: Minimize KL divergence

- $\log p_\theta(x) \geq E_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - \boldsymbol{KL}\big(\boldsymbol{q_\phi}(\boldsymbol{z}|\boldsymbol{x})||\boldsymbol{N}(\boldsymbol{0},\boldsymbol{I})\big)$
- $KL\big(q_\phi(z|x)||N(0,I)\big) = KL\big(N(\mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x)))||N(0,I)\big)$
- For two Gaussians, this KL divergence has a simple expression

$$= \frac{1}{2}\left(\|\mu_\phi(x)\|_2^2 - m + \sum_{j=1}^{m}\big(\sigma_\phi^2(x)_j - \log\big(\sigma_\phi^2(x)_j\big)\big)\right)$$

- Sanity check: what if $\mu_\phi(x) = 0$ and $\sigma_\phi^2(x)_j = 1$ for all j?

Optimizing: Autoencoding points

- $\log p_\theta(x) \geq \boldsymbol{E_{z\sim q_\phi(z|x)}}[\mathbf{\log p_\theta(x|z)}] - KL\big(q_\phi(z|x)||N(0,I)\big)$
- We imagine the density $p_\theta(x|z)$ is that of $N(\mu_\theta(z), I)$ where μ_θ is the decoder network
  - When sampling, can instead just output $\mu_\theta(z)$ rather than additional sampling
    - Analogy: when we run softmax on outputs of an NN, we output the max index, we don't sample from it
- $E_{z\sim q_\phi(z|x)}[||x - \mu_\theta(z)||_2^2]$ ~~$- d \log \sqrt{2\pi}$~~ (essentially same as AE)
- Given sampling capability, can draw $z \sim q_\phi(z|x)$ to optimize
- Reparameterization trick (Draw how to sample $Z \sim N(\mu, \sigma^2)$ as $\mu + \sigma G$ where $G \sim N(0,1)$)

Summary

- Solve generative modelling
- Use neural network to map Gaussian samples to data distribution
- Do it by using variational autoencoder: tries to map original distribution to a Gaussian, and also maps back to original distribution. Each is encoded in the loss function.

Samples from a VAE

(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Interpolation using VAEs (Explain how)