Autoencoders

Machine Learning: Jordan Boyd-Graber
University of Maryland
SLIDES ADAPTED FROM IAN GOODFELLOW

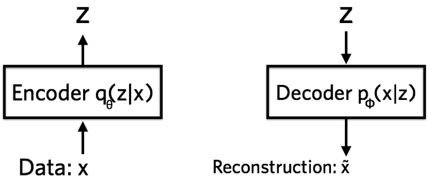
Problems of Autoencoders

- Unsupervised
  - Lots of data
  - Need priors / regularization
- Probabilistic loss function
  - sampling too slow
  - hard to explain hidden layer probabilistically

Why autoencoders

- Discover hidden structure
  - Unlike clustering or admixtures, continuous
  - Not always interpretable
- Reconstruct data
- Features for downstream model (a la word2vec)

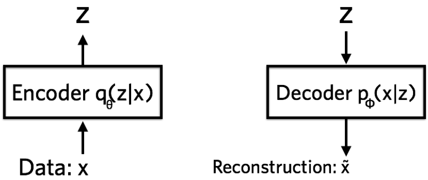
Loss Function



$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z|x_i)}\left[\log p_\phi(x_i|z)\right] + \mathrm{KL}(q_\theta(z|x_i)||p(z)) \tag{1}$$

- Reconstruction error
- Variational representation distribution
- Regularization

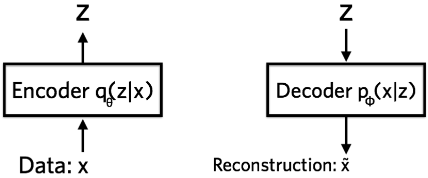
Loss Function



$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{1}$$

- Reconstruction error
- Variational representation distribution
- Regularization

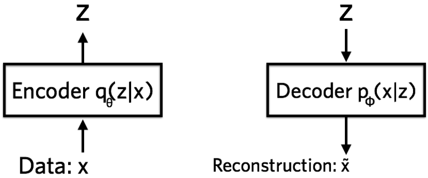
Loss Function



$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{1}$$

- Reconstruction error
- Variational representation distribution
- Regularization

Loss Function



$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{1}$$

- Reconstruction error
- Variational representation distribution
- Regularization

Interpretation

- Lower bound on reconstruction of decoder
- Keep representation constrained
- Probabilistic parameterization

Make this Concrete

- $p(z)$: standard normal distribution
- $q(z \mid x_i)$: normal distribution with output of NN as mean [variational distribution]
- $\mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z))$
- Decoder $p_\phi(x \mid z)$ depends on model / data:
  - Grayscale Image? Bernoulli distribution for each pixel
  - Words? Multinomial over vocabulary

Make this Concrete

- $p(z)$: standard normal distribution
- $q(z \mid x_i)$: normal distribution with output of NN as mean [variational distribution]
- $\text{KL}(q_\theta(z \mid x_i) \,||\, p(z))$
- Decoder $p_\phi(x \mid z)$ depends on model / data:
    - Grayscale Image? Bernoulli distribution for each pixel
    - Words? Multinomial over vocabulary

Make this Concrete

- $p(z)$: standard normal distribution
- $q(z \,|\, x_i)$: normal distribution with output of NN as mean [variational distribution]
- $\mathrm{KL}(q_\theta(z \,|\, x_i) \,||\, p(z))$
- Decoder $p_\phi(x \,|\, z)$ depends on model / data:
    - Grayscale Image? Bernoulli distribution for each pixel
    - Words? Multinomial over vocabulary

Variational Inference Story

$$\ell_i(\lambda) = \mathbb{E}_{q_\lambda(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] - \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{2}$$

- Want to optimize $p_\phi(x \mid z)$ (likelihood)
- ELBO remains lower bound
- Difference is KL between variational distribution and $p(z)$

Variational Inference Story

$$\ell_i(\lambda) = \mathbb{E}_{q_\lambda(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] - \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \quad (2)$$

- Want to optimize $p_\phi(x \mid z)$ (likelihood)
- ELBO remains lower bound
- Difference is KL between variational distribution and $p(z)$
- Actually simpler than LDA
  - No global latent variables (only z)
  - Can minibatch the data

Variational Inference Story

$$\ell_i(\lambda) = \mathbb{E}_{q_\lambda(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] - \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{2}$$

- Want to optimize $p_\phi(x \mid z)$ (likelihood)
- ELBO remains lower bound
- Difference is KL between variational distribution and $p(z)$
- Actually simpler than LDA
  - No global latent variables (only z)
  - Can minibatch the data
  - But what about ϕ? (encoder)

Variational EM

- Learn variational parameters
- Update ϕ using supervised backprop
- (Depends on data model)