Notes on Autoencoder

1 Autoencoder

Encoder: $f()$
Decoder: $g()$
Autoencoder: $g(f(x)) = x$
f and g are linear each represented by weight matrices w_f and w_g

$$h = w_f X \tag{1}$$

$$\tilde{X} = w_g h \tag{2}$$

2 Linear Autoencoder

2.1 Objective

$$\min_W \frac{1}{2} \sum_n ||W_g W_f x_n - x_n||_2^2 \tag{3}$$

2.2 Algorithm

Gradient descent
when using Euclidean norm (squared loss), solution is the same as principal component analysis

3 PCA

Project data into lower dimension hyperplane using a linear transformation

4 Nonlinear Autoencoder

f and g are non-linear functions:

$$\min_W \frac{1}{2} \sum_n ||g(f(x_n; W_f); W_g) - x_n||_2^2 \tag{4}$$

where hidden nodes correspond to non-linear manifold

5 Deep Autoencoders

f and g often consist of multiple layers

6 Sparse Representations

6.1 Force hidden nodes to be sparse

$$\min_W \frac{1}{2} \sum_n ||g(f(x_n; W_f); W_g) - x_n||_2^2 + c \quad nnz(f(x_n; W_f)) \tag{5}$$

in which $nnz(f(x_n; W_f))$ is the number of non-zero entries in the vector produced by f

6.2 Approximate objective: L1 regularization

$$\min_W \frac{1}{2} \sum_n ||g(f(x_n; W_f); W_g) - x_n||_2^2 + c \quad nnz(f(x_n; W_f)) \tag{6}$$

6.3 L1 regularization

$$\min_W \frac{1}{2} \sum_n ||g(f(x_n; W_f); W_g) - x_n||_2^2 + c||f(x_n; W_f)|| \tag{7}$$

7 Probabilistic Autoencoder

Let f and g represent conditional distributions: $f : Pr(h|x; W_f)$ and $g : Pr(x|h; W_g)$ by using sigmoid, softmax or linear units at the hidden and output layers

- sigmoid: Bernoulli distribution: binary hidden nodes and outputs
- softmax: Categorical distribution
- linear: Gaussian: $N(h|x, \mu, \sigma^2; W_f)$ and $N(x|h, \mu, \sigma^2; W_g)$