We need a title

Justin Johnson
jcjohns@stanford.edu

Bharath Ramsundar
rbharath@stanford.edu

1 Introduction

Some recent applications of deep learning have utilized unsupervised pretraining to greatly improve their performance on classification tasks [1]. One of the fundamental building blocks of unsupervised pretraining is the sparse autoencoder. In this project we aim to develop a theoretical and practical understanding of different varieties of autoencoders.

2 Definitions

A single-layer autoencoder is a neural network with a single hidden layer that attempts to learn the identity function. The hidden layer activations of a trained autoencoder can then be used as a feature vector for the original data.

More precisely, let $W^{(1)} \in \mathbb{R}^{p\times n}$ and $W^{(2)} \in \mathbb{R}^{n\times p}$ be matrices of weights, let $b^{(1)} \in \mathbb{R}^p$ and $b^{(2)} \in \mathbb{R}^n$ be bias vectors, and let $f : \mathbb{R} \to \mathbb{R}$ be an activation function; a common choice is the sigmoid $f(z) = 1/(1 + e^{-z})$. These parameters define a neural network with a single hidden layer. For an input $x \in \mathbb{R}^n$ the output of the network is $h_{W,b}(x) = f(W^{(2)}f(W^{(1)}x+b^{(1)})+b^{(2)})$ where f is applied componentwise. The term $f(W^{(1)}x + b^{(1)})$ represents the activations of the input x on the hidden layer of the network. To train an autoencoder, we are given data $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ and we must find W and b to minimize the reconstruction error $\sum_i \|h_{W,b}(x^{(i)}) - x^{(i)}\|$ under some norm; additional constraints such as regularization or sparsity may also be imposed.

A sparse autoencoder imposes additional constraints on the hidden layer activations averaged over the training data which is given by $\hat{\rho} = \frac{1}{m}\sum_i f(W^{(1)}x^{(i)} + b^{(1)})$. Typically we want to force $\hat{\rho}$ to be close to some desired activation level ρ.

3 Equivalence with PCA

In the case

4 Linearization

We considered a linearization of a neural net. That is, we defined nonlinearity f to be the identity function. Then the function $h_{W,b}(x) = W^{(2)}(W^{(1)}x + b^{(1)}) + b^{(2)}$. We also add a ℓ^2 sparsity constraint $\|\rho - \hat{\rho}\|_2^2$. The minimization problem then becomes

$$\min_W \sum_{i=1}^{m} \|W^T W x^{(i)} - x^{(i)}\|_2^2 + \beta\|\rho - \hat{\rho}\|_2^2$$

We explicitly derived the gradient of the above formula and implemented gradient descent.

References

[1] Quoc V Le, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Marc'Aurelio Ranzato, Jeff Dean, and Andrew Y Ng. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*, 2011.