Denoising Autoencoders における確率的勾配降下法
Stochastic Gradient Descent for Denoising Autoencoders

Yusuke Sugomori
http://yusugomori.com

February 2013

概要

Denoising Autoencoder を実装する際に必要となる数式の導出過程をまとめた.

Denoising Autoencoder (DA) [Vincent 2008] では，入力データを $\boldsymbol{x}$，入力データの一部を損傷させたデータを $\tilde{\boldsymbol{x}} \sim q_{\mathcal{D}}(\tilde{\boldsymbol{x}}|\boldsymbol{x})$ としたとき，$\tilde{\boldsymbol{x}}$ から $\boldsymbol{x}$ を復元するように訓練を行う．$\boldsymbol{x}$が 2 値のベクトル，すなわち $\boldsymbol{x} \in (0,1)^d$ であるとき，DA における各写像は

$$\text{Encode}: \quad \boldsymbol{y} = f_\theta(\tilde{\boldsymbol{x}}) = s(W\tilde{\boldsymbol{x}} + \boldsymbol{b}) \tag{1}$$

$$\text{Decode}: \quad \boldsymbol{z} = g_{\theta'}(\boldsymbol{y}) = s(W^{'}\boldsymbol{y} + \boldsymbol{b}^{'}) \tag{2}$$

で与えられる．ただし，$s(\cdot)$ はシグモイド関数である．また，$W^{'} = W^T$ として考える．式 (2) における $\boldsymbol{z}$ が，$\tilde{\boldsymbol{x}}$ から復元されたデータに相当する．

DA の実装には，確率的勾配降下法（stochastic gradient descent, SGD）が用いられる．誤差関数は，交差エントロピー誤差関数

$$L_H(\boldsymbol{x}, \boldsymbol{z}) = -\boldsymbol{x}\log\boldsymbol{z} - (\boldsymbol{1}-\boldsymbol{x})\log(\boldsymbol{1}-\boldsymbol{z}) \tag{3}$$

で与えられるため，$\theta = (W, \boldsymbol{b}, \boldsymbol{b}^{'})$ に対しての L_H の勾配を求めればよい．以下，符号の煩わしさを防ぐため，$L := -L_H = \boldsymbol{x}\log\boldsymbol{z} + (\boldsymbol{1}-\boldsymbol{x})\log(\boldsymbol{1}-\boldsymbol{z})$ とし，L の勾配を導出する．導出に際し，以下の式で表される $\boldsymbol{h}_1 = \boldsymbol{h}_1(\tilde{\boldsymbol{x}})$, $\boldsymbol{h}_2 = \boldsymbol{h}_2(\boldsymbol{y})$ を定義する．

$$\boldsymbol{h}_1(\tilde{\boldsymbol{x}}) = W\tilde{\boldsymbol{x}} + \boldsymbol{b} \tag{4}$$

$$\boldsymbol{h}_2(\boldsymbol{y}) = W^T\boldsymbol{y} + \boldsymbol{b}^{'} \tag{5}$$

$\boldsymbol{h}_1, \boldsymbol{h}_2$ を用いると，$\boldsymbol{y} = s(\boldsymbol{h}_1)$, $\boldsymbol{z} = s(\boldsymbol{h}_2)$ と表せる．

W, $\boldsymbol{b}$, $\boldsymbol{b}^{'}$ それぞれによる微分は

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \boldsymbol{h}_1}\frac{\partial \boldsymbol{h}_1}{\partial W} + \frac{\partial L}{\partial \boldsymbol{h}_2}\frac{\partial \boldsymbol{h}_2}{\partial W} \tag{6}$$

$$\frac{\partial L}{\partial \boldsymbol{b}} = \frac{\partial L}{\partial \boldsymbol{h}_1}\frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{b}} = \frac{\partial L}{\partial \boldsymbol{h}_1} \tag{7}$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{'}} = \frac{\partial L}{\partial \boldsymbol{h}_2}\frac{\partial \boldsymbol{h}_2}{\partial \boldsymbol{b}^{'}} = \frac{\partial L}{\partial \boldsymbol{h}_2} \tag{8}$$

であり，シグモイド関数の微分 $s^{'}(x) = s(x)(1-s(x))$ を利用すると，

$$\frac{\partial L}{\partial \boldsymbol{h}_1} = \frac{\partial L}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}_1} = \frac{\partial L}{\partial \boldsymbol{y}} * \boldsymbol{y} * (\boldsymbol{1} - \boldsymbol{y}) \tag{9}$$

$$\frac{\partial L}{\partial \boldsymbol{h}_2} = \boldsymbol{x} - \boldsymbol{z} \tag{10}$$

が得られるため，式 (9),(10) を式 (6),(7),(8) に代入すればよい．ただし，$*$ はベクトルの要素積を表している．これは，一般に実装の際に用いられる記号に合わせて記述した．また，

$$\frac{\partial L}{\partial \boldsymbol{y}} = W(\boldsymbol{x} - \boldsymbol{z}) \tag{11}$$

となるため，最終的に，

$$\frac{\partial L}{\partial W} = \Big(W(\boldsymbol{x} - \boldsymbol{z}) * \boldsymbol{y} * (\boldsymbol{1} - \boldsymbol{y})\Big)\, \tilde{\boldsymbol{x}}^T + \Big((\boldsymbol{x} - \boldsymbol{z})\, \boldsymbol{y}^T\Big)^T \tag{12}$$

$$\frac{\partial L}{\partial \boldsymbol{b}} = W(\boldsymbol{x} - \boldsymbol{z}) * \boldsymbol{y} * (\boldsymbol{1} - \boldsymbol{y}) \tag{13}$$

$$\frac{\partial L}{\partial \boldsymbol{b}'} = \boldsymbol{x} - \boldsymbol{z} \tag{14}$$

なる勾配が得られる．

N 個の入力データ群 $X = \{\, \boldsymbol{x}_1,\ \ldots,\ \boldsymbol{x}_N \,\}$ が与えられたとすると，$L = -L_H$ であるから，符号に注意すると，DA における確率的勾配降下法によるパラメータの更新は以下の式で与えられる．ただし，ϵ は学習率である．

$$W^{new} = W^{old} + \frac{\epsilon}{N} \sum_{n=1}^{N} \frac{\partial L}{\partial W} \tag{15}$$

$$\boldsymbol{b}^{new} = \boldsymbol{b}^{old} + \frac{\epsilon}{N} \sum_{n=1}^{N} \frac{\partial L}{\partial \boldsymbol{b}} \tag{16}$$

$$\boldsymbol{b}'^{new} = \boldsymbol{b}'^{old} + \frac{\epsilon}{N} \sum_{n=1}^{N} \frac{\partial L}{\partial \boldsymbol{b}'} \tag{17}$$

参考文献

[Vincent 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol: Extracting and Composing Robust Features with Denoising Autoencoders, *Proceedings of the 25th International Conference on Machine Learning*, (2008), pp.1096-1103