Deep Learning (Homework 1)

Due date : 4/13/2018

- Any tools with automatic differentiation are forbidden in this homework, such as Tensorflow, Pytorch, Keras, MXNet, etc. You should implement backpropagation algorithm by yourself.
- Homework submission – Please zip each of your source code and report into a single compress file and name the file using this format : HW1_StudentID_StudentName.zip (rar, 7z, tar.gz, ... etc are *not* acceptable)

Deep Neural Network for Regression and Classification

1. **Regression**: Buildings are the largest energy consumers in the world. They account for around 30% of global carbon emissions, and over one-third of final energy use. This share could double or triple by 2050 if we do not act, as buildings have a long life-cycle that locks in their energy use. According to the International Energy Agency, energy consumption in buildings needs to be reduced by 80% by 2050 if we want to limit global temperature rise to under 2°C. Our goal is to analyze the energy efficiency of buildings by using the dataset energy_efficiency_data.csv which contains the simulation energy loads of different buildings using 8 different features. 8 different features and two simulation energy loads are listed below
    - Relative compactness R_c
        - measurement of compactness of closure of the building
    - Surface area
        - surface area of the building
    - Wall area
        - area of the building covered by width of the wall
    - Roof area
        - area covered under roofs
    - Overall height
        - overall height of the building
    - Orientation
        - orientation of building based on direction
            * 2: north
            * 3: east
            * 4: south
            * 5: west
    - Glazing area

  - ratio of total area of the wall which is glass
- Glazing area distribution
  - distribution of glazing area within the whole building
    * 1: uniform
    * 2: north
    * 3: east
    * 4: south
    * 5: west
- Heating load
  - amount of heating load required to heat the building
- Cooling load
  - amount of cooling load required to cool the building

Related link is accessible at https://www.slideshare.net/NitinAgarwal53/exploratory-data-analysis-for-energy-efficiency. Heating load is the target. You need to encode the categorical features (orientation, glazing area distribution) into one-hot vectors. Shuffle the dataset and spilt it into the first 576 samples and the remaining 192 samples as training and test data, respectively.

i. Please construct a deep neural network (DNN) for regression by using the sum-of-squares error function

$$E(\mathbf{w}) = \sum_{n=1}^{N} (t_n - y(\mathbf{x}_n; \mathbf{w}))^2.$$

Minimize this error function $E(\mathbf{w})$ by running the error backpropagation algorithm using the stochastic gradient descent

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)})$$

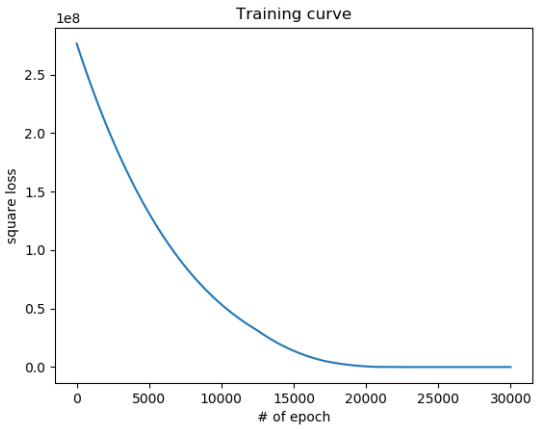
Measure the performance by root mean square (RMS) error

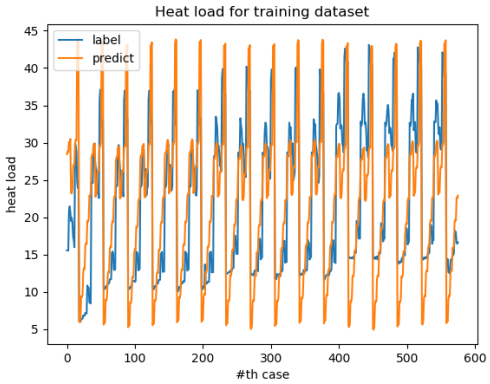
$$E_{\mathrm{RMS}}(\mathbf{w}) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (t_n - y(\mathbf{x}_n; \mathbf{w}))^2)}$$

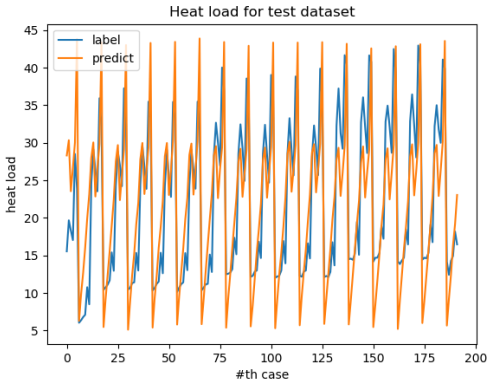
You need to design a feature selection procedure to verify the input features which significantly affects the heating load. You can refer to the result from the website https://goo.gl/Vh8zVv, which visualizes how input features affect the energy load. Each feature selection estimation should contain the network architecture, the features you choose, the learning curve, the training RMS error, the test RMS error, the regression result with training labels and the regression result with test labels.

ii. Discuss which input features significantly affect the heating load and explain in details why your procedure works.

Network architecture	$15 - 10 - 10 - 1$
Selected features	$[0, \ldots, 7]$
Training E_{RMS}	5.94988
Test E_{RMS}	5.99459







2. **Classification**: In this exercise, you will implement a DNN model for binary classification using the spam dataset. This database contains the attributes obtained from spam emails and normal emails together with their corresponding labels: "spam" or "not spam", which are coded as +1 and -1. The objective in this exercise is to create and train a neural network to identify spam on email automatically, based on the attributes obtained from the emails. The detail of this dataset is accessible at https://archive.ics.uci.edu/ml/datasets/spambase. Dataset file spam_data.mat contains 4 subsets for training and test, test_x.mat (600×40), test_y.mat (600×2), train_x.mat (3000×40), train_y.mat (3000×2). All the labels are expressed by two dimensional one-hot vectors. You can load the data by command load. Matlab : load('spam_data.mat') / Python : scipy.io.load('spam_data.mat')

   i. Please construct a DNN for binary classification according to the cross-entropy error function

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \log y_k(\mathbf{x}_n, \mathbf{w})$$

   Minimize the error function $E(\mathrm{w})$ by running the error backpropagation algorithm using the stochastic gradient descent

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)})$$

   You should decide the following variables: number of hidden layers, number of hidden units, learning rate, number of iterations and mini-batch size. You have to show your (a) learning curve, (b) training error rate and (c) test error rate in the report. You can design the network architecture by yourself.

   ii. Design your network architecture with the layer of 2 nodes or 3 nodes before the output layer. Plot the distributions of latent features for the cases of (a) 2 nodes and (b) 3 nodes at different training stages. For example, you may show the results when running at 5^{th} and 10^{th} learning epochs.

iii. Please discuss the evolution of latent features at different training stage.