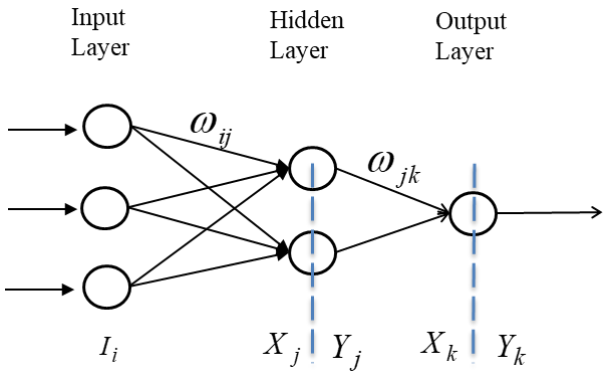
Chair of Transportation Systems Engineering
Department of Civil, Geo and Environmental Engineering
Technical University of Munich

Transportation Systems M.Sc.	Statistical Learning and Data Analytics for Transportation Systems
Problem set 2	Due Date: June 30, 2020, 15:00

Neural Network and Regression

Problem 1 (30 points)

(a) Write a computer program using R code to implement the forward and backward propagation algorithms of a multilayer perceptron (MLP) that you learned from the course. Use it to fit a single hidden layer MLP that is illustrated as follows.



Use the same data from the class example to verify your program, i.e. the input is an observation of three features X = [5.0, 0.5, 2.0], the target value is Y=1.0. The initial value is 0.1 for the weight matrix between the input and hidden layer, 0.5 for the weight matrix between the hidden and output layer. The learning rate is α=5.0. The activation function is the sigmoid function for both hidden and output layers. Run 2 epochs to make sure the program can produce the same results as the example illustrated in the lecture. Do not use any existing neural network packages in R! Submit your R code.

(b) Generate a sample data set of size 1000 from the following model

$$Y = f\left(a_1 X\right) + \left(a_2 X\right)^2 + 0.30Z$$

where

f is the sigmoid function;

$Z^T=(Z_1, Z_2)$ is a random vector, Z_1 and Z_2 are $N(0,1)$ independent standard normal variables;

$X^T = (X_1, X_2, X_3)$ is a random vector, $X_j, j = 1,2,3.$ are $N(0,1)$ independent standard normal variables;

the parameters

$$a_1 = \begin{bmatrix} 3 & 2 & 3 \\ 5 & 7 & 8 \end{bmatrix} \quad \text{and} \quad a_2 = \begin{bmatrix} 3 & -2 & 3 \\ 5 & -7 & 8 \end{bmatrix}$$

Divide the sample data that you generated into training dataset (i.e. 70%) and test dataset (i.e. 30%). Use the program code you developed from (a) to train a MLP neural network with the training dataset and validate the network with the test dataset. Use four neurons in the hidden layer. The weight matrices are initialized with random numbers that follow $N(0,1)$. Plot the training error curve and test error curve as a function of the number of epochs for two different learning rate (i.e. α=0.5, 5.0). Two curves should be shown together in one figure per learning rate. Discuss the overfitting behavior in each case. Submit your R code.

Problem 2 (30 points)

The highway management and control center in Muniville monitors the traffic state of a highway segment shown as Figure 1. Traffic volume and speed data were automatically and continuously collected from location 1, 2 and 3 when the loop detectors work well. However, the loop detectors are sometimes offline due to either a regular maintenance or malfunction. Usually, it takes about three business days for the loop detectors to be back online. For that reason, you are asked to use supervised learning method to estimate traffic state (i.e. volume and speed) for a location when its loop detector is absent.

Given one-month historical data from location-1, 2, and 3 (the data can be found in a separate spreadsheet.), you are asked to design a Deep Neural Network (DNN) for the purpose of traffic state estimation for location-3 when its loop detector is absent/malfunctional for a 3-day period. It is suggested that you split the historical data into three data sets; the first set is used to train the Deep Neural Network; and the second set (e.g. 15-20%) is used to validate your training; the third set which contains the last 3 days of data is used to test prediction accuracy of your neural network. The time interval is 60 minutes.

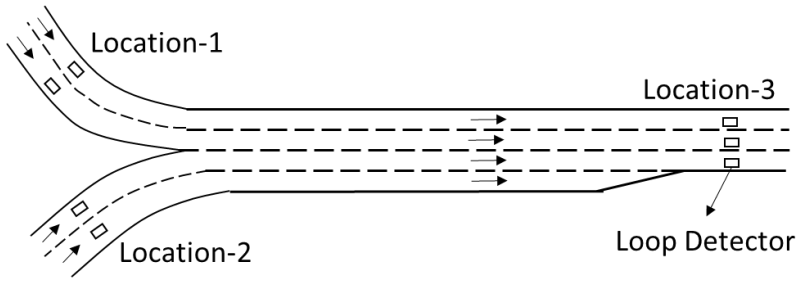


Figure 1: A highway segment in Muniville

Use one of the existing R packages to establish a Deep Neural Network (DNN), train the network, and perform the forecast for location-3 for 3 days of traffic state.

(a) Identify an optimal DNN structure from the following combinations that produces the best forecast:

- the number of hidden layers is between 1 to 3
- the number of neurons in each hidden layer is between 5 to 10

Describe your procedures how you determine the number of neurons for the input layer, hidden layer, and output layer, and how do you determine the number of hidden layers?

(b) Submit your R code. Plot your 3 days of forecast vs observation curves together in one figure according to time sequence. Choose at least three different metrics to measure the accuracy of your forecast and explain in detail with tables and figures. Plot the architecture of your neural network.

Problem 3 (6 points)

You have fit a multiple linear regression model and the $(X'X)^{-1}$ matrix is:

$$(X'X)^{-1} = \begin{bmatrix} 0\ 893758 & -0.0282448 & -0.0175641 \\ -0.0282448 & 0.0013329 & 0.0001547 \\ -0.0175641 & 0.0001547 & 0.0009108 \end{bmatrix}$$

(a) How many regressor variables are in this model?

(b) If the error sum of squares is 307 and there are 15 observations, what is the estimate of σ^2?

(c) What is the standard error of the regression coefficient $\hat{\beta}_1$?

Problem 4 (7 points)

Consider the linear regression model

$$Y_i = \beta_0^{'} + \beta_1 \left(x_{i1} - \bar{x}_1 \right) + \beta_2 \left(x_{i2} - \bar{x}_2 \right) + \varepsilon_i$$

Where $\bar{x}_1 = \frac{1}{n} \sum x_{i1}$ and $\bar{x}_2 = \frac{1}{n} \sum x_{i2}$

(a) Write out the least squares normal equations for this model.

(b) Verify that the least squares estimate of the intercept in this model is $\beta_0^{'} = \frac{1}{n} \sum y_i = \bar{y}$.

(c) Suppose that we use $y_i - \bar{y}$ as the response variable in this model. What effect will this have on the least squares estimate of the intercept?

Problem 5 (7 points)

Assume that we are given a dataset, where each sample x_i and regression target y_i is generated according to the following process

$$x_i \sim Uniform(-10,10)$$
$$y_i = ax_i^3 + bx_i^2 + cx_i + d + \varepsilon_i \quad where \quad \varepsilon_i \sim \mathrm{N}(0,1) \; and \; a,b,c,d \in \mathbb{R}.$$

The 3 regression algorithms below are applied to the given data. Your task is to say what the bias and variance of these models are (low or high). Provide a 1-2 sentence explanation to each of your answers.

(a) Linear regression

(b) Polynomial regression with degree 3

(c) Polynomial regression with degree 10

Problem 6 (20 points)

The data which can be found in a separate spreadsheet provides the highway gasoline mileage test results for 2005 model year vehicles from DaimlerChrysler.

(1) Fit a multiple linear regression model to these data to estimate gasoline mileage that uses the following regressors: cid, rhp, etw, cmp, axle, n/v

(2) Estimate σ^2 and the standard errors of the regression coefficients.

(3) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(4) Find the *t*-test statistic for each regressor. Using $\alpha = 0.05$, what conclusions can you draw? Does each regressor contribute to the model?

(5) Find 99% confidence intervals on the regression coefficients.

(6) Plot residuals versus $\hat{y}$ and versus each regressor. Discuss these residual plots.