Introduction to Neural Networks

Johns Hopkins University—Engineering for Professionals Program
605.447/625.438

Classification Programming Assignment

This handout describes the last assignment of the course and involves applying your knowledge of neural networks to create, train and evaluate a neural network based on data (see below). Below is a scenario and problem motivation along with data with which you will train and evaluate your neural network.

1. Data/Motivation/Scenario

Advertising companies, organizations and associations as well as credit providers have an interest in determining the best groups of consumers that should be exposed to advertising campaigns. These groups of consumers can be induced to engage in financial transactions such as purchases of goods and services by specially *targeted* advertising campaigns and other inducements. Advertising groups are motivated to generate as much revenue as possible per dollar spent on advertising that includes these targeted advertising campaigns. In other words, these groups want to get the 'biggest bang per buck' from their advertising expenditures.

Targeted advertising campaigns are especially profitable because groups to which these campaigns are directed are most likely to respond to advertising campaigns by engaging in financial transactions. It is therefore important to determine which of these groups and/or individuals should be targeted for these advertising campaigns. There are several factors used to determine which groups are most profitable to target. The effort/cost of determining these factors varies.

The first metric used to assess whether a targeted campaign is warranted is referred to as the household's *Size of Wallet* (SOW) and is an estimate of a household's disposable income or money that is likely to be spent on non-essential goods and services. This metric is associated with or based on a number of complex socio-economic factors, but is principally based on a household's estimated *gross income* (GI). When the GI is over $200,000, there is no further change in the SOW for these households.

Another important metric is the *local affluent code* (LAC). This is a very complicated metric to accurately determine, but rough estimates are obtainable by looking at neighborhood housing sales, comparable housing prices and other factors. This ranges from 0 – 3 where numbers near 0 indicate a very low level of affluence typically associated with low-grade building structures, low housing prices and some degree of neighborhood blight and crime. A grade of near 1 corresponds to higher housing prices, typically in the range of state average prices per square foot, higher quality housing structures and low crime rates. A grade near 2 or above (where 3 is the highest possible LAC score), corresponds to the highest affluence score typified by housing prices over $500k, newer constructions (although this rating can encompass very old, stable and high priced housing) and little or no crime.

Another value used to determine whether a targeted campaign is warranted is a *targeted advertising code assignment* (TACA). The TACA score ranges from 0 – 2 where a 0 indicate a negative return on investment (ROI), something that signals a group to avoid. A score of 1 indicates an approximately even return on investment, *e.g.,* a dollar spent targeting these households tends to return a dollar in revenue and again, represents a group advertisers want to avoid. A score of 2 however represents a distinctly positive return on investment and indicates a household that advertisers want to target in their advertising campaigns. The data below is based on historical information and can be used to train a neural network so that using inputs for Gross Income, SOW and LAC, a value for the TACA can be output.

2. Neural Network Data

Accurately determining the estimated Gross Income, SOW, LAC and TACA values involves collecting and analyzing a host of data and is relatively expensive. The whole point of using a neural network in this context is to serve as a decision support system that can quickly and accurately identify households with a high TACA score. Consequently, the TACA score can serve as an output of a neural network while the SOW and LAC can serve as inputs.

The following data shows the associations of the four values. Use this data to train and evaluate two neural networks: one to produce outputs of the SOW and the other to output a TACA value.

Data Item	LAC	SOW	TACA
1	1.98	10K	0
2	1.80	10K	1
3	1.05	160K	2
4	1.45	180K	1
5	1.8	80K	1
6	1.96	110K	1
7	0.4	40K	2
8	2.05	130K	1
9	0.90	10K	1
10	2.5	60K	0
11	1.6	105K	2
12	1.05	196K	1
13	0.52	105K	2
14	1.80	32K	1
15	2.3	106K	0
16	2.4	151K	1
17	2.5	170K	1
18	0.50	150K	2
19	1.1	35K	1
20	.85	70K	2

3. Methodology – General Considerations

Your neural network will have an input layer with two inputs and upto two hidden layers and a single output node that outputs a value of the TACA code value. Because the output ranges from 0 – 2 the output node should use the *ramp*

activation function: $y = \ln(1 + e^x)$. Note that just as the Sigmoid activation function is a differentiable version of the step function, the ramp function above is a differentiable version of the linear ramp function as in Fig. 1b. Also, note that the integrals of two functions in the left column are their corresponding functions in the right column.

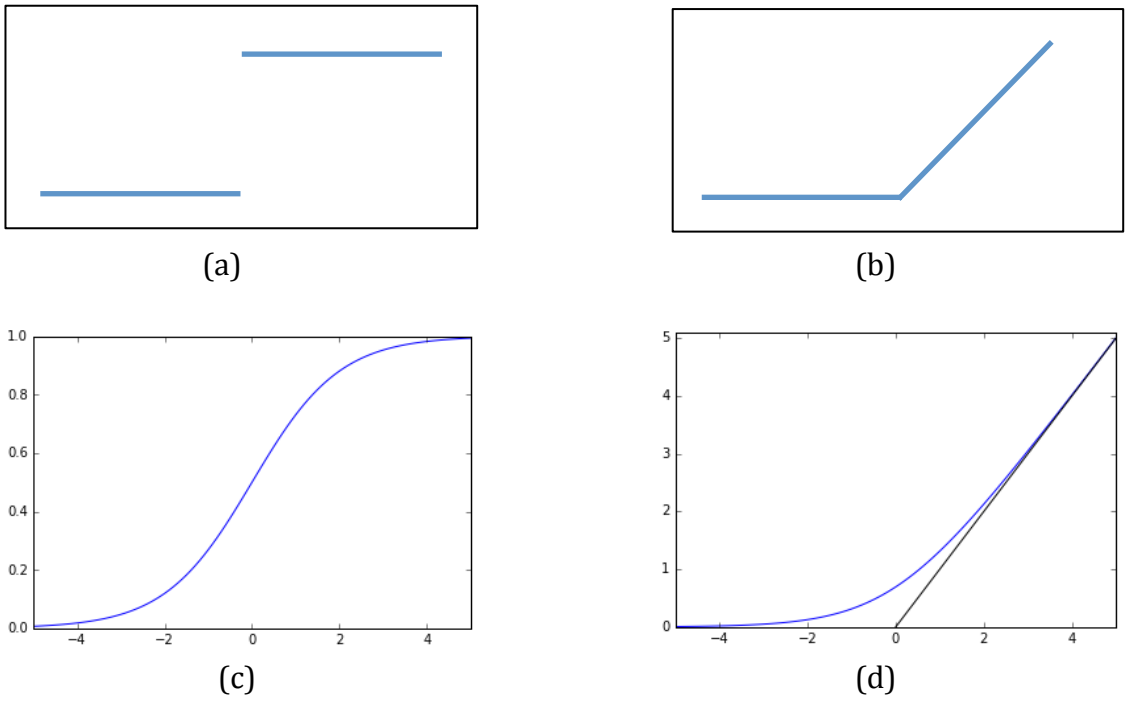

Fig. 1: The step function (upper left) and its integral, the linear ramp function. The continuous analogs of these function in the second row.

Consequently, for this network you will have to use a modified back-propagation algorithm that takes into account this activation function in the output layer node (recall, from your mid-term exam that this will necessitate modifying the Perceptron Delta Function for the output layer node). You will need to do the math.

The neural network that outputs the TACA code value can be a standard feed-forward, back-propagation algorithm. Also note, that because the LAC varies from 0 – 3, using the Sigmoid activation function in the hidden layers can be effective when the LAC value is normalized. Similarly, one can normalize the SOW since the upper bound is $200,000.

where the output is or maybe modified by a threshold logic function *after it has been trained* but before it is evaluated.

3.1 Methodology – Training

Exercise 1:
For training purposes, use the second 10 data items (items 11 – 20) indicated in the table above. Use the online form for training (similar to Method 1 in the FFBP assignment). Train the network for 1000 cycles.

For evaluation purposes, use the first 10 data items. During the evaluation phase, you will find it necessary to map the output value to an integer value that corresponds to the TACA value. This requires you to use some sort of threshold logic function. Once you decide on the best thresholds to use (how would you do that?), then determine the performance of your network using Receiver Operating Characteristics.

Exercise 2:
For this exercise, use the same methodology as in Exercise 1 including applying the same approach you used to set threshold values only this time, run the training phase for 5000 iterations. Calculate the ROCs again and compare them to those obtained from Exercise 1.

4. Written Report

The written report will indicate the weights for each perceptron after final training, and the measures of performance such as the mean squared error for the TACA value and the data corresponding to the *receiver operating characteristics* for the classification problem such as the sensitivity, the specificity and you *may* include the *negative predictive probability* and the *positive predictive probability* if you desire.

In your report, indicate any particulars you think is relevant towards using the neural network as a decision support tool and limit your report to no more than 10 pages. You may include your programming code as an appendix.