Mathematical Statistics
Stockholm University
Bachelor Thesis **2019:6**
http://www.math.su.se

A Comparison of Logistic Regression and Neural Networks for Binary Classification Problems

Anna Basetti*

June 2019

Abstract

Binary classification is the task of classifying data points into either one of two groups, based on their coordinates (explaining variables). Logistic regression and neural networks are two of the most widely used classification models. The former has its roots in traditional statistics, while the latter originates from the younger field of machine learning. Neural networks have seen a renaissance during the past years, and have been surrounded by a great deal of hype and an aura of mystery, while logistic regression kept being regarded as a more robust, down-to-earth method. The scope of this thesis is to dissect the two models, highlighting their similarities and differences, both theoretically and practically.

The first part of this work presents the theory behind logistic regression and neural networks. A close look at the structure of a vanilla neural network reveals that such a model is a rather natural expansion of logistic regression: in fact, a vanilla neural network with the sigmoid as its activation and output function and the cross-entropy error function is exactly a logistic regression in the hidden nodes, as well as each hidden node is a logistic regression in the explanatory variables. However, the flexibility gained through the addition of nonlinearity makes the neural network a more powerful method when the data at hand is not linearly separable.

The second part of this work consists of three experiments, performed on three different simulated data sets exhibiting different geometrical properties. For each experiment, one logistic regression and three neural networks with varying numbers of nodes in the hidden layer were fitted, and the resulting decision boundaries were plotted. The performance of each model was then tested on 50 simulated validation data sets. The results coincided with the expected performances, proving that neural networks do serve as a flexible and powerful extension of logistic regression, in areas where predictive power, rather than interpretability, is the priority.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: anna.basetti@outlook.com. Supervisor: Ola Hssjer.