Analytics

Silke Meiner, Rafaela Neff

10 3 2021

Comparison of NN configurations on the Breast Cancer Wisconsin (Diagnostic) Data Set

Beuth University for Applied Sciences, Machine Learning 2, Prof. Tim Downie, WS2020/2021

Abstract

I, Introduction:

Motivation: Deathrates: In 2016 there were 68,950 women and 710 men suffering from Breast Cancer (ICD-10 C50) in Germany only. In this year they've counted 18,570 deaths in females. Respectivly only 1.1 men died in the same year. In addition the relative 5-year survival rate decreases to 87% for Women, were the realtive 10-yearsurvival rate is just 82%. For men we see a 77% relative 5-year survival rate and a 72%.

Data Information: For our project we will be working with the Breast Cancer Wisconsin (Diagnostic) Data Set which consists of 569 entries with 357 being benign and 212 malignant. As you can see, this dataset is a bit unbalanced with 37% of the entries belonging to the positive class and 67% belonging to the negative class.

The data was obtained from a digitized image of a fine needle aspirate (FNA) of a breast mass. The features were computed from the FNA. They describe characteristics of the cell nuclei present in the image.

II, Model

III, Methods

We will investigate the performances of our proposed neural network architectures on this dataset.

For this project we've used the R-Package nnet, as well as the R-Interface to Java's H2o-Package. The R package nnet comes with 1 hidden layer by default. This is not supposed to be changed through this project.

The neural networks were trained on a 80% split of the dataset, keeping the proprtions of benign and malignant diagnoses as they were in the original dataset.

For fitting a network with the nnet R-Package we've used theh library caTools to split the training set again 80/20, where 20 results un the validation split.

For fitting our neural network with the h2o R-pakcage, we have choosen cross-validation using a stratified fold_assignment.

The test data split is used for benchmarking only.

model_nnet: One hidden layer wit 20 nodes, a weight decay (L2-Norm) of 0.01, a range of 0.6 (??? what'S that) and trace = TRUE (???? explain). The maximum iterations to be done are set to 200. Correct classification: error rate:

Interesting findings