James Taylor

Assignment 4

University of Maryland University College

DATA630-9041 Spring 2020

Dr. Gates

james.taylor.ps3@gmail.com

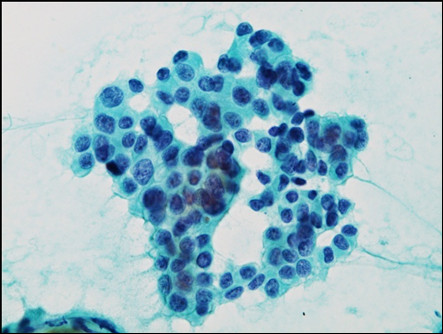
**Introduction**

Approximately 270,000 women will be diagnosed with breast cancer this year, and 42,000 will die from it (American Society of Clinical Oncology, 2020). Breast cancer is the second most common cause of death from cancer for women. The lethality dramatically increases when the cancer spreads and remains untreated.

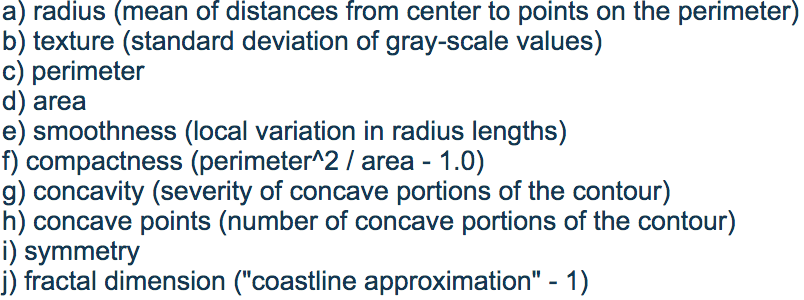
Cancer has become more of an issue as people live longer and don’t die from other causes less (like car accidents and disease). The survival rate of breast cancer has increased over time as medicine has improved. The modern cancer treatments can only be used to treat someone if physicians agree they actually have it.

The goal of this paper is to predict which unseen observations are malign as accurately as possible using a neural network. The model should particularly minimize false negatives, which can allow patients who have cancer to progressively get worse as it is left untreated. The lessons and model used in this paper will likely be useful in detecting malignant cells in other parts of the body.

**Analysis and Model Development**

About Dataset

The dataset used for this paper is from the University of Wisconsin. The data are from an image on a breast mass collected by a fine needle aspirate (example on right). The independent variables all describe the cell nucleus. The independent variable is if the cell is malignant or normal. These are listed below with an example image similar to those used in this dataset. There are values for three nuclei in each observation because multiple cells are extracted. That means there are three iterations of the below variables for each observation.



Preprocessing

The dataset had no missing values. The only variable removed was a unique identifier for each observation because the goal is to best predict the dependent variable regardless of theory. There were 569 observations on 31 variables. All variables were standardized in an attempt to speed up the model’s gradient decent and to help the weights since they all will have the similar magnitudes. The dependent variable also had to be converted into a numerical variable because neural networks can only handle numerical values.

The data was split into training and test sets. The training set was 80% of the original dataset and randomly selected, making it 456 observations to train on and 113 to test.

Exploratory Data Analysis

The dataset has about two-thirds benign and a third malignant observations. This is a fair mixture for the model to train on. The high amount of malignant cases likely means this test was performed once someone was suspected of having cancer, like after finding an anomaly in a mammogram.

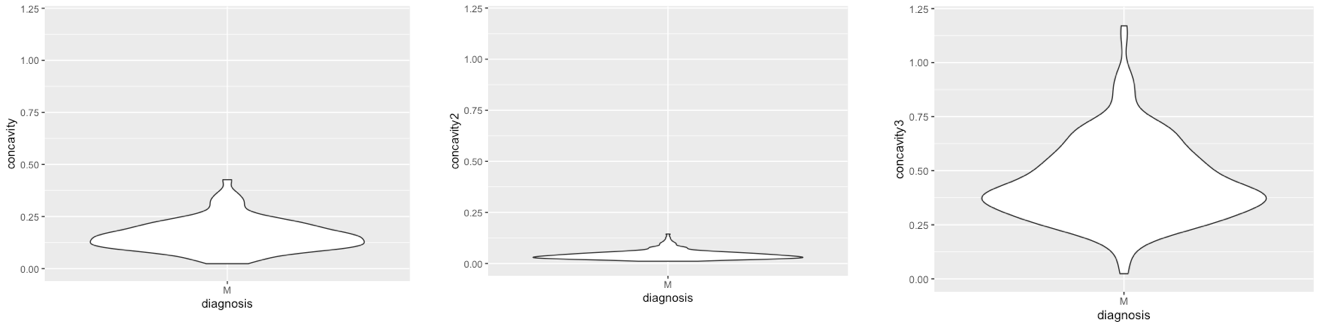


The texture value is likely to help the model distinguish between benign and malignant cells. The distributions below show the mode and the mean are different. Benign cells are more likely to have lower texture values. The texture value is the standard deviation of the gray-scale of the nucleus. The more color deviation a nucleus has the higher the texture value.



Compactness is another variable that can help discover malignant cells. Benign cells had no values over .23, but malignant cells had a significant amount above that. Malignant cells had a larger range and more positive kurtosis. Benign cells had a compactness close to zero while malignant had no such values below .05.



 The distributions of the values on the different nuclei can significantly vary as well. Below is a figure that shows the distribution of concavity of the three nuclei on malignant cells only. The second nuclei had a relatively small range and the third has lots of variation. There is likely lots of variation among the nuclei across the other variables, this is just one variable demonstrating it.

Model

The model used was from the ‘neuralnet’ package (version 1.44.2). It is a neural network which optimizes weights to neurons to have the best predictions on training data. Complex prediction environments need more neurons to capture the intricate relationships. The weights are adjusted by gradient descent until no adjustment can be made to improve on the model’s prediction cannot be improved on. The inputs and neurons that contributed to the correct predictions have their weights increased while incorrect ones have them decreased.

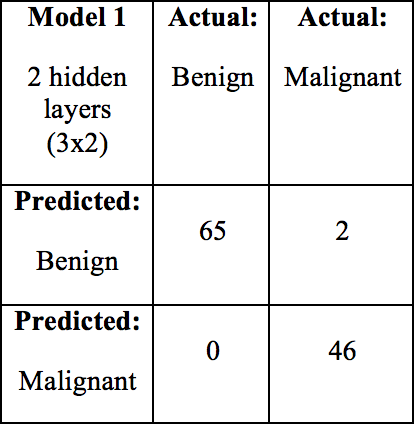
**Results and Model Evaluation**

The first model created had two hidden layers with three and two neurons respectively. The cross-entropy error function was used because the dependent variable had either a value of zero or one. The linear output parameter was set to false. On the training data, it correctly predicted 455 of the 456 observations. While this was a positive result, it is not all that useful to use. Below is a plot of the model.



Model Accuracy on Test Data

The true test of a model’s performance is on labeled data it has not seen before. This model had an accuracy of 99%, incorrectly predicting only one of the 113 test observations. The confusion matrix is below.



The two observations that the model got wrong was the 39th and 216th observation of the dataset. The model predicted the values to be .003 and .039 which was rounded to zero (meaning benign) but the actual value was a one because it was a malignant case. The model had significant confidence that its prediction was going to be correct because it was very close to zero.

New Model

Making the incorrect classification can have dire consequences. The model made two incorrect predictions that could have ultimately lead to no further investigation, resulting in a late stage cancer situation. Another model was built with more hidden layers in an attempt to get a model that doesn’t make such a mistake. This new neural network has three layers containing five, three and two neurons respectively.

The more complex model again incorrectly labeled two malignant observations as benign. One was the 39th observation again and the other was the 14th. The 216th observation was able to be predicted accurately potentially because the complex model better understood the relationship. The 14th observation may have been incorrectly predicted because of overfitting with more neurons.

Results

The nature of neural networks makes inspecting them very difficult. There is no good way to uncover how or why it made these predictors (like a logistic function where each piece of a model can be inspected). Overall the model was accurate, reporting 98% accuracy on unseen data.

The models had the same accuracy on unseen data, but the incorrect predictions where on different observations. The training and test data was the same on each model, so the differences seen can only be attributed to the different levels of complexity or the gradient descent for the models.

**Conclusion**

The models produced were 98% accurate on data that it had not seen before. This level of accuracy means it neural networks can be utilized as a tool in determining malignant cases. The both models produced false negatives which is cause for concern. This shows that models should not be the only approach to determining the presence of cancer.

All the variables that were provided were used in the model. This I mainly because the purpose was to achieve maximum accuracy. The concern is approaching this problem with a preconceived there is biased and may lead to information left behind from discarded variables. Conversely, overfitting becomes more of an issue though there is not strong evidence of it in these models. The model is capable of setting an input variables weight to zero, making the value of it have no influence on model’s outcome.

Machine learning methods should work on methods that can make accurate predictions using information from less intrusive test. The data used for this paper had a needle that has to be sanitized and trained medical professional to perform it. A model that can simply uses mammogram images will likely derive more value as they are a more common test.

References

American Society of Clinical Oncology. Breast Cancer: Statistics. January 2020. Retrieved from <https://www.cancer.net/cancer-types/breast-cancer/statistics>