Machine Learning in Healthcare

Study of machine learning algorithms and application on the Pima Indians Diabetes Dataset.

Alexander Roque Rodrigues

A dissertation presented for the degree of
Bachelor in Computer Science

Department of Computer Science
Smt. Parvatibai Chowgule College of Arts and Science
India
18 August 2019

Acknowledgements

Contents

Part I
Background

Data is everywhere. International initiatives coupled with global disruptive innovation are the leading causes for pushing datafincation forward.

Datafication refers to the modern-day trend of digitalizing (or datafying) every aspect of life.

This data creation is enabling the transformation of data into new and potentially valuable forms. Entire municipalities are being incentivized to become smarter. In the not too distant future, our towns and cities will collect thousands of variables in real time to optimize, maintain, and enhance the quality of life for entire populations. One would reasonably expect that as well as managing traffic, traffic lights may also collect other data such as air quality, visibility, and speed of traffic. As a result of big data from connected devices, embedded sensors, and the IoT, there is a global need for the analysis, interpretation, and visualization of data.

1 Diabetes Mellitus

More commonly referred to as "diabetes" – a chronic disease associated with abnormally high levels of the sugar glucose in the blood. Diabetes is due to one of two mechanisms:

Inadequate production of insulin (which is made by the pancreas and lowers blood glucose), or Inadequate sensitivity of cells to the action of insulin. The two main types of diabetes correspond to these two mechanisms and are called insulin dependent (type 1) and non-insulin dependent (type 2) diabetes. In type 1 diabetes there is no insulin or not enough of it. In type 2 diabetes, there is generally enough insulin but the cells upon which it should act are not normally sensitive to its action.

The signs and symptoms of both types of diabetes include increased urine output and decreased appetite as well as fatigue. Diabetes is diagnosed by blood glucose testing, the glucose tolerance test, and testing of the level of glycosylated hemoglobin (glycohemoglobin or hemoglobin A1C). The mode of treatment depends on the type of the diabetes.

The major complications of diabetes include dangerously elevated blood sugar, abnormally low blood sugar due to diabetes medications, and disease of the blood vessels which can damage the eyes, kidneys, nerves, and heart.

Part II
Review of Literature

1 Predicting Diabetes Mellitus With Machine Learning Techniques

Diabetes is segregated into 2 categories

Part III
Data Preprocessing

1 Feature Extraction

1.1 Filters

1.2 Wrappers

Part IV
Machine Learning Algorithms

1 Linear Regression

1.1 Introduction to Linear Regression

Linear regression is a forecasting technique that can be use to predict the future of a number series based on the historic data given. The perks of using a linear regression model are as follows:

- produces decent and easy to interpret results.
- is computationally inexpensive.
- conversion of algorithm into code does not take much effort or time.
- numeric values as well as nominal values support is offered.

However, a major drawback of linear regression is that it **poorly models nonlinear data**.
Considering a dataset that has values ranging from $X = \{x_1 + x_2 + x_3 + ... + x_n\}$ where all the entries of the dataset are real numbers. Each x_i is associated with a corresponding value of y_i from the dataset $Y = \{y_1 + y_2 + y_3 + ... + y_n\}$.

The most basic equation for linear regression can be expressed via this simple equation.

$$y = \beta_0 x + \beta_1 + \epsilon$$

So to minimize the error in the predictions, a way to calculate the error should be formulated. A loss function in machine learning is simply a measure of how different the predicted value is from the actual value. The Quadratic Loss Function to calculate the loss or error in our linear regression model. It can be defined as:

$$L(x) = \sum_{i=1}^{n} (y_i - p_i)^2$$

Therefore using the method of Least Squares, we can find the values of β_0 and β_1.

$$\beta_0 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The linear regression model with an error value close to 1.00 indicates a perfect model and those with values closer to 0.00 indicates a model that delivers poor performance.

2 Logistic Regression

Even if called regression, this is a classification method which is based on the probability for a sample to belong to a class. As our probabilities must be continuous in R and bounded between (0, 1), it's necessary to introduce a threshold function to filter the term z. The name logistic comes from the decision to use the sigmoid (or logistic) function:

3 K-Nearest Neighbours

The k-means algorithm is based on the (strong) initial condition to decide the number of clusters through the assignment of k initial centroids or means:
Then the distance between each sample and each centroid is computed and the sample is assigned to the cluster where the distance is minimum. This approach is often called minimizing the inertia of the clusters, which is defined as follows:
The process is iterativeonce all the samples have been processed, a new set of centroids K (1) is computed (now considering the actual elements belonging to the cluster), and all the distances are recomputed. The algorithm stops when the desired tolerance is reached, or in other words, when the centroids become stable and, therefore, the inertia is minimized. Of course, this approach is quite sensitive to the initial conditions, and some methods have been studied to improve the convergence speed. One of them is called k-means++ (Karteeka Pavan K., Allam Appa Rao, Dattatreya Rao A. V., and Sridhar G.R., Robust Seed Selection Algorithm for K-Means Type Algorithms, International Journal of Computer Science and Information Technology 3, no. 5, October 30, 2011), which selects the initial centroids so that they are statistically close to the final ones. The mathematical explanation is quite difficult; however, this method is the default choice for scikit-learn, and it's normally the best choice for any clustering problem solvable with this algorithm.

3.1 Finding Optimum Number of Clusters

One of the most common disadvantages of k-means is related to the choice of the optimal number of clusters. An excessively small value will determine large groupings that contain heterogeneous elements, while a large number leads to a scenario where it can be difficult to identify the differences among clusters. Therefore, we're going to discuss some methods that can be employed to determine the appropriate number of splits and to evaluate the corresponding performance.

4 Decision Tree

5 Random Forest

6 Gradient Boosting

7 Support Vector Machine

Support Vector Machines is an algorithm that is capable of handling linear as well as data that occurs non-linearly. For example, for a long time, SVMs were the best choice for MNIST dataset classification, thanks to the fact that they can capture very high non-linear dynamics using a mathematical trick, without complex modifications in the algorithm.

7.1 Linear Support Vector Machines

Let us consider a dataset of features we want to classify.

$$X = \{x_1, x_2, x_3, ..., x_n\}$$

For the target variable, we will consider the dataset Y, with target outcomes as $\{0, 1\}$ indicating a true or false condition.

$$Y = \{y_1, y_2, y_3, ..., y_n\}$$

8 Perceptron

The perceptron is the foundation of neural networks.

9 Multilayered Perceptron

Part V
Real World Application

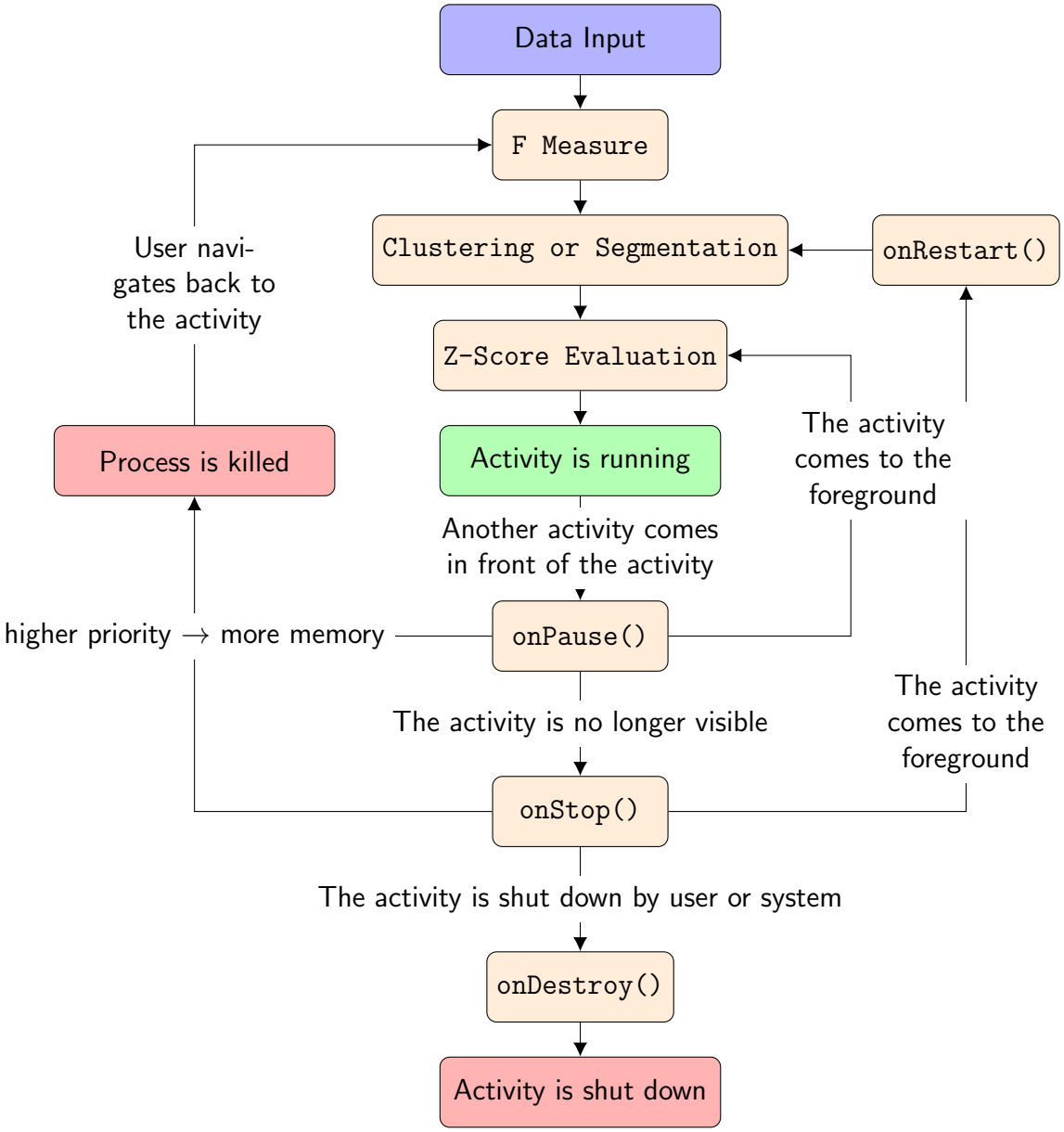
Part VI
Building Information Systems for Prediction

1 Introduction

In today's world we have many algorithms and multiple datasets along with sufficient data available for testing and training the algorithms.

2 Feature Selection

3 Models



4 Conclusion

Part VII
Conclusion

References

[1] Wei M, Gibbons LW, Mitchell TL *et al.* (1999) The Association between cardiorespiratory fitness and impaired fasting glucose and type 2 diabetes mellitus in men. *Ann Intern Med* **130**, 427-34.

[2] Jr., W. C. S. (2017, January 26). Definition of Diabetes mellitus. Retrieved November 17, 2019, from https://www.rxlist.com/script/main/art.asp?articlekey=2974.

[3] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y.,& Tang, H. (2018, November 6). Predicting Diabetes Mellitus With Machine Learning Techniques. Retrieved November 17, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/.