CS CAPSTONE TECHNOLOGY REVIEW

NOVEMBER 3,2019

REDUCING PATIENT DOSE FROM DIAGNOSTIC IMAGING USING MACHINE LEARNING

PREPARED FOR

OREGON STATE UNIVERSITY

STEVE REESE

PREPARED BY

GROUP 16

YIHONG LIU

Abstract

This report introduces the possible technology to implement the Linear Regression algorithms for our project, with different models, and the detailed solution about how to implement the algorithms step by step, and the visualization of data. Then summarizing the best suitable possible technology that can be used in the project.

CONTENTS

1 Introduction

The technology review examines three components of the ”Reducing Patient Dose from Diagnostic Imaging Using Machine Learning” project. For each component, three alternatives are explored for possible use in the project and each alternative will be evaluated based on their specifications.

2 Problem Description

Diagnostic imagery is a useful tool for viewing internal bodily structures. Radiation is required in order to create an image with high enough saturation that any objects of interest are clearly visible by humans. If the dosage is not enough, these objects won't be visible due to low saturation. However, the dosage also needs to be limited in order to avoid harmful side effects in the patient. If a doctor orders a diagnostic scan, they believe that the potential findings outweigh the risk imposed by additional radiation exposure.

Current diagnostic imagery tools emit radiation until an exposure threshold has been reached as detected by a radiation detector. This leads to needless exposure because a point is reached where additional radiation will not yield satisfactorily better imagery. The high likelihood of generating an image with enough quality comes at the expense of excess irradiation of the patient. Current tools are not sophisticated enough to know the optimal dosage with enough accuracy. Also, since the process happens so quickly, it wouldn't be feasible for a person to monitor the image quality and stop the machine manually. Finding a way to stop the imagery process earlier will in theory yield the same results with less radiation exposed to the patient.

3 Piece 1: Linear Regression Machine Learning Model

3.1 Overview of Criteria

The Linear Regression is one of the most important machine-learning algorithms in our project, regression analysis is a set of statistical processes fro estimating the relationships among variable, the focus is on the relationship between a dependent variable and one or more independent variables, in simple words, we want to predict y(the dependent variable, or target) based on a set of x's(dependent variables, or features), where y is continuous. with this algorithms, the system can train the data set with features to, and correct the machine learning model as the data goes wrong or inaccurately, then after optimizing the model, the machine learning model can automatically detect the peak point (target variable) of the specific images of radiations, which the system can know the right point to stop the unnecessary does of radiation, to make the patient suffer less pain.

3.2 Simple Linear Regression

With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and co-variance. All of the data must be available to traverse and calculate statistics.

3.3 Ordinary Least Squares

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations.

3.4 Gradient Descent

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

When using this method, the system must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.

it is useful when there is a very large data set either in the number of rows or the number of columns that may not fit into memory. [1]

4 Piece 2: The implementation of Linger Regression machine learning model

4.1 Overview of Criteria

The implementation of Linger Regression machine learning model describes the detailed (but still in a high level) design about how to achieve our goal with the algorithms.

4.2 Reading the data set file

Since the whole project is based on Python programming language, the Pandas is the one of the most suitable and necessary tool to read the data file. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. the process of reading data can be viewed as the way to collect data into our project, it can be achieved with something sudo code like this ” example = data.values[0:,1:]”.

4.3 Processing the data with python

Processing data is the way to correct and manipulate the data, one of the common utilization is feature extraction. For example, if clients only want to figure the correlation between size of house and the price of the house, then just extracting the ”size of house” only from bunch of variables, such as area, decoration, etc. except that, there are some method can be used to process data set, such as Normalization, one-hot encoding, regularization. The goal of normalization is to change the values of numeric columns in the dataset ti a common scale, without distorting differences in the ranges of values. The one-hot encoding representing each piece of data in a way that the computer can understand, which is a process by which categorical variables are converted into a from that could be accepted by computer.

Regularization is the process of adding information in order to solve an ill-popsed or to prevent over fitting. The major reason that Regularization is necessary is that the model will have a low accuracy it it is over fitting, because the model is trying too hard to capture the noised in the training dataset. This technique discourages learning a more complex or flexible model, so as to avoid the risk of over fitting. With the Scikit-learn, it provides the class linear Regression classifier, which can implement many processing function without writing any code extra. [2] [3]

4.4 evaluating the performance of the model

The Root mean squared error (RMSE) and coefficient of determination (R square score) will be the main method being used to determine the accuracy. RMSE is the square root of the average of the sum of the squares of residuals. R square score or the coefficient of determination explains how much the total variance of the dependent variable can be reduced by using the least square regression. Those value can tell whether the linger regression model achieves the accuracy requirements or not. [4]

5 Piece 3: visualization of data

5.1 WEKA

The Waikato Environment for Knowledge Analysis (WEKA) provides an implementation of the C4.5 algorithm. It also provides a rich Graphical User Interface (GUI) with tools for experimenting with the classifier parameters and output. it also contains tools for data pre-processing, classification, regression.

5.2 R

R is a programming language and software environment for statistical computing and graphic supported by R foundation for statistical computing, R is widely used among data miners and statisticians for developing statistical software and data analysis. The way R code implement in our project is for analyzing the data set, it has extensive visualization capabilities.

6 Recommendations

For the piece 1, those are some potential regression models, but previous two can not handle the dataset with too many dependent variables, so Gradient Descent will be the best choice for our project. For the piece 2, there is no recommendations, it's just all the detailed design about how to achieve implementing the linear regression model step by step. For piece 3, WEKA might be better than R in our project, although we might need to use both of R and WEKA, but WEKA can Handel most of algorithms that we need to implement, it has more selection, more packages, than R. [5]

References

[1] J. Brownlee, "Linear regression for machine learning." [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/

[2] P. Gupta, "Regularization in machine learning." [Online]. Available: https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

[3] A. AI-Masri, "How does linear regression actually work?" [Online]. Available: https://towardsdatascience.com/how-does-linear-regression-actually-work-3297021970dd

[4] A. Agarwal, “Linear regression using python.” [Online]. Available: https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2

[5] J. Brownlee, “A gentle introduction to scikit-learn: A python machine learning library.” [Online]. Available: https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/