Machine Learning Engineer Nanodegree

Capstone Proposal

Godaba Hari Chandana

February7th, 2019

Breast Cancer Wisconsin (Diagnostic) Classification

Domain Background:

In these days the Domain of Health Care is stepping to get engaged with Artificial Intelligence and Machine Learning. This project is based on the Domain of Health Care. This gives accurate results regarding the classification of Breast Cancer. It takes the features of cells in the body and predicts that the woman is affected with Breast Cancer or not. It's a good initiation to implement this project in HealthCare Centers. The aim of this application is to avoid the *Human Errors* in the Domain of **Health Care** .

The dataset is collected from UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

cite or link of an academic paper where machine learning was applied to this type of problem is :

https://www.sciencedirect.com/science/article/pii/S1877050916302575

Problem Statement:

Given a dataset which consists the features of cells in the body and an algorithm needs to be developed to classify the Breast Cancer and to determine if it is malignant or benign .This can then be used to automatically detect the women who are suffering from Breast Cancer.

Hence the goal is to classify a new data point to either Malignant or Benign on basis of the features of the given data point. The tasks involved in it:

1. Download the data from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
2. Cleaning the data and removing the unnecessary features which are not used to predict the target variable.
3. Visualising the data to know the characteristics of the features.
4. Train the data with Classification Models (without Feature Reduction) and then test to find which classifier performs good on the dataset by using appropriate methods then finally identifies the model which gives the best accuracy score.

5. Also Train the data with the Classification models (with feature reduction) and then test to find which the classifier performs good on the dataset and then finally identifies the model which gives the best accuracy score.

Datasets and Inputs:

The dataset which is used for Breast Cancer Wisconsin(Diagnostic) Classification is taken from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

This dataset consists of 569 instances with 32 attributes and it is a multivariate because it consists Categorical, real and integer valued attributes. For the best result I will split the data into training and testing sets. On a whole I will assign 70% of data to training set and 30% of data testing set.

Attributes are:

1. ID number

2. Diagnosis (M = malignant, B = benign)

3-32) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

Solution Statement :

Here, I am trying to predict whether the Breast Cancer is malignant or benign by using features of cells in the body. For this prediction I will use different classification models and in the result to find the model which gives the best accuracy.

In this project, I will implement classification models for Breast Cancer Wisconsin prediction, by using python and machine learning toolbox sklearn. Then I will find the accuracy scores given by all the classification models without feature reduction and then identify which model gives the best accuracy. Then I will execute the feature reduction process and then find the accuracy scores given by all classification models and finally identify which model gives the best accuracy. I explore the dataset by using read_csv and for visualization which helps me to better understand the solution, I used matplotlib.pyplot.

Benchmark Model:

Benchmark model is a model which we will take as reference and achieve the best result which is more than the benchmark model. In my project the Accuracy score will be compared between the different classification models(Random Forest Classifier , Logistic Regression, GaussianNB, KNeighborsClassifier, Decision Tree Classifier, SVM) and select the best one. I think Decision Tree model can be set as benchmark model and I'm sure that the final solution would outperform the Benchmark model.

Evaluation Metrics:

For the prediction of Breast Cancer Wisconsin I want to use accuracy score as an evaluation metric. It is the number of correct predictions made by the model over all kinds of predictions made. Firstly I will predict the accuracy score given by the classification models without feature reduction and then I will select a model which gives the best accuracy score. Then I will also predict the accuracy scores given by all classification models with feature reduction then I will select a model which gives the best accuracy.

Accuracy Score can be also calculated as:

Accuracy Score = TP+TN/TP+FP+FN+TN

True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

True negatives (TN): We predicted no, and they don't have the disease.

False positives (FP): We predicted yes, but they don't actually have the disease. (Also known as "Type 1error")

False negatives (FN): We predicted no, but they actually do have the disease.(Also known as "Type 2 error"

Project Design:

From the description and problem statement it can be inferred that Classification algorithms (Supervised or Unsupervised) can be employed for this problem.

Pre-processing:

Initially data exploration will be carried out to understand possible labels and also helps in preprocessing the data and can end up with better predictions.

In the process of preprocessing the first task is to read the data by using read_csv and perform visualizations on it to get some insights about the data .After that I will clean the data that is remove features which are not used for the prediction of target variable and also remove the unwanted data or replacing null values with some constant values or removing duplicates.

Training and Testing the data:

After the data exploration process, I will split the data into training and testing sets and then train the data by using different classification models like Random Forest Classifier, Decision Tree Classifier, SVC, Logistic Regression, KNeighborsClassifier, GaussianNB and then test all the models with testing data and then find the accuracy score given by the models and then finally identifies which model that gives the best accuracy score. Hence this model is declared as the best model to detect the Breast Cancer Wisconsin.