Capstone Proposal

Himadri Nandini Das Bebartta

Mar 10, 2019

Domain Background:

Machine Learning is used in a wide variety of fields today. In this project machine learning will be used to build a model that can predict whether a person suffers from breast cancer or not. Breast Cancer is a cancer that forms in the cells of the breasts. Breast Cancer is highly predominant in women in today's world. It is caused by uncontrolled growth of abnormal cells in the breast. It can start in the breast and can spread to other areas of the body in the course of time.

In the modern medical science there are plenty of newly devised methodologies and techniques for detection of breast cancer. Even though most of these techniques make use of highly advanced technologies such as medical image processing, but still most of the research work done till now detects the breast cancer at tumour stage and are not accurate to 100% and leads to false positive or false negative results which are highly dangerous. Implementation of different supervised learning such as Support Vector Machines, Ensemble methods, and Stochastic gradient descent classifier may lead to very high accuracy yielding true positive and true negative results.

Problem Statement:

A tumour can be in two stages and they are benign and malignant. Benign stage tumours are not dangerous to health and are non cancerous whereas malignant tumours has the potentiality of being dangerous and are cancerous. In this project, we will be using Breast Cancer Wisconsin (Diagnostic) Data Set to create a model that will be able to predict if a tumour is dangerous or not based on the characteristics that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

The physical characteristics of a malignant tumor differ from the benign tumor cells. Thus a measure on the different characteristics such as radius (mean of distances from center to points on the perimeter), texture_mean (standard deviation of gray-scale values), perimeter_mean, area_mean, smoothness_mean(mean of local variation in radius lengths), compactness, concavity, concave points, symmetry or fractal dimension helps us to understand that a new sample for classification belongs to which class of tumor i.e. Malignant or Benign.

The dataset is clearly a classification oriented problem as the column diagnosis consists of two values Malignant (M) and Benign (B). For obtaining a model we need to split the dataset into training set, validation set and testing set. A Testing set is required to predict how good the model performs on unseen data.

The datasets and inputs:

To create a model to predict whether a tumor cell is benign or malignant we will use a labelled dataset which stems from kaggle and is available at https://www.kaggle.com/uciml/breast-cancer-wisconsin-data .

This was created in 1995 by:

Dr. William H. Wolberg (General Surgery Dept., University of Wisconsin Clinical Sciences Center),
W. Nick Street (Computer Sciences Dept., University of Wisconsin)
and Olvi L. Mangasarian (Computer Sciences Dept., University of Wisconsin).

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The dataset have the following structure:

- Number of instances: 569
- Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)
- Diagnosis (M = malignant, B = benign)
- Missing attribute values: none
- Class distribution: 357 benign, 212 malignant
- All feature values are recoded with four significant digits.
- 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.

The dataset is appropriate for the problem at hand because it contains all the required information to train a classification model that determines whether a cell is benign or malignant.

Solution Statement

This is a supervised learning problem because the dataset is labelled. After performing some preprocessing steps, we have to develop an algorithm that could be used for breast cancer detection. At the end of the project, the code will accept an image of cell and predict whether it is in benign or malignant stage. This implies the use of a binary classification model. For each row in the labelled dataset it will be possible to determine whether the model predicted the correct class.

The solution will be clearly quantifiable and measurable. For example by dividing the number of correctly classified entries by the total number of entries in the dataset one can calculate the accuracy of the model.

Benchmark Model

Logistic Regression will be used as a benchmark model for the dataset. Along with logistic regression we will also be using SVM, and different ensemble methods to find out that which model outperforms in identifying, if a tumour falls under Benign or Malignant category.

The judgement on accuracy can be performed using F1 score or area under the ROC curve (auc)metric. These metrics will be used to used to compare the results as F1 Score, ROC can be used to quantify the performance of the model because they can be used for binary classification problems

Metric

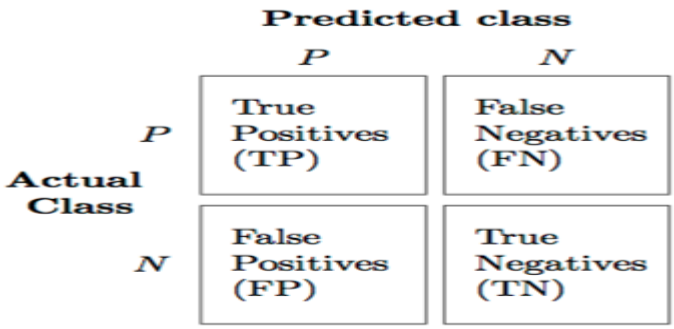
In tumor cells classification is important to avoid false negatives because if a malignant tumor is predict as benign the patient will not receive treatment. That's why F1 is a ideal metric to score our model.

recall = True positive / (True positive + False negative)

precision = True positive / (True positive + False positive)

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

F1 score is the harmonic average of the precision and recall. This score can range from 0 to 1, with 1 being the best possible F1 score i.e. perfect precision and recall. And 0 is the worst F1 score.



Using F1 Score formula.

In statistical analysis of binary classification, the F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Project Design

First of all the data need to be explored. Libraries and dataset is to be loaded. Then data pre processing/cleaning is required where if any outliers found should be removed. Missing data will be filled in or the whole row will be dropped. Identification of feature and target columns is required.

The dataset will be split into training, validation and testing set. Testing set is safely kept to find out how the model performs for unseen data. Many different models will be trained using standard parameters. The benchmark model for determining the breast cancer to be in Malignant or Benign stage is Logistic Regression. Several other classifiers such as Support Vector Machines, Ensemble methods such as AdaBoost, Random Forest, Gradient Boosting will also be implemented along with logistic regression.

Thus different algorithms will be evaluated by building models and selecting the best model. The model will be quantified based on F1 score. Other metrics which are applicable for classification problems such as ROC Curve, or F score might also be taken into consideration

References

1. https://www.researchgate.net/publication/310796041_Predicting_Breast_Cancer_Recurrence_using_effective_Classification_and_Feature_Selection_technique
2. https://github.com/CesarTrevisan/Breast-Cancer-Detection
3. https://github.com/ck24/capstone-proposal/blob/master/proposal.pdf
4. https://www.kaggle.com/cihanyatbaz/machine-learning-with-breast-cancer-dataset