Machine Learning Engineer Nanodegree

Using Supervised Learning to predict whether the breast cancer is benign or malignant.

Poli Naidu Sigilipalli

Jan 5th 2019

Proposal

Domain Background

Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the US, and it is the second leading cause of cancer death among women. It occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a tumor. A tumor does not mean cancer – tumor can be benign (non cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed.

Similar topic links:

https://www.sciencedirect.com/science/article/pii/S2001037014000464

https://www.sciencedirect.com/science/article/pii/S093336571730009X

https://www.sciencedirect.com/science/article/pii/S0933365710000679

Problem Statement

Given breast cancer results from breast fine needle aspiration (FNA) test (is a quick and simple procedure to perform, which removes some fluid or cells from a breast lesion or cyst (a lump, sore or swelling) with a fine needle similar to a blood sample needle). Since this build a model that can classify a breast cancer tumor using two training classification:

1= Malignant (Cancerous) - Present
0= Benign (Not Cancerous) -Absent

Since the labels in the data are discrete, the predication falls into two categories, (i.e. Malignant or benign). In machine learning this is a classification problem. Thus, the goal is to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period. To achieve this we have used machine learning classification methods to fit a function that can predict the discrete class of new input.

Features and Description:

- Id - ID number
- Diagnosis - The diagnosis of breast tissues (M = malignant, B = benign)
- radius_mean - mean of distances from center to points on the perimeter
- texture_mean-standard deviation of gray-scale values
- perimeter_mean-mean size of the core tumor
- area_mean
- smoothness_mean - mean of local variation in radius lengths
- compactness_mean - mean of perimeter^2 / area - 1.0

- concavity_mean - mean of severity of concave portions of the contour
- concave points_mean - mean for number of concave portions of the contour
- symmetry_mean
- fractal_dimension_mean - mean for "coastline approximation" - 1
- radius_se - standard error for the mean of distances from center to points on the perimeter
- texture_se - standard error for standard deviation of gray-scale values
- perimeter_se
- area_se
- smoothness_se - standard error for local variation in radius lengths
- compactness_se - standard error for perimeter^2 / area - 1.0
- concavity_se - standard error for severity of concave portions of the contour
- concave points_se - standard error for number of concave portions of the contour
- symmetry_se
- fractal_dimension_se - standard error for "coastline approximation" - 1
- radius_worst - "worst" or largest mean value for mean of distances from center to points on the perimeter
- texture_worst - "worst" or largest mean value for standard deviation of gray-scale values
- perimeter_worst
- area_worst
- smoothness_worst - "worst" or largest mean value for local variation in radius lengths
- compactness_worst - "worst" or largest mean value for perimeter^2 / area - 1.0
- concavity_worst - "worst" or largest mean value for severity of concave portions of the contour
- concave points_worst - "worst" or largest mean value for number of concave portions of the contour
- symmetry_worst
- fractal_dimension_worst - "worst" or largest mean value for "coastline approximation" – 1

Data Sets and inputs

I've taken this dataset from UCI Machine Learning Repository. It contains 569 rows and 32 columns. In this dataset , out of 569 records 357 records are under benign and 212 are under malignant. It is some what evenly distributed and balanced set. But it is always better to use other all performance metrics than accuracy_score to avoid getting false negatives i.e people who are malignant but predicted as benign. We cannot let effected people go to their homes without being treated. In this case, false negatives are to be taken more care than false positives.

Solution Statement

This dataset can be viewed as classification problem. I will use classification supervised models for this problem. I will use different classification models and calculate accuracy and F1 scores for every model and select the best one out of them. I will use methods like GridSearch to tune the parameters for getting better results from a model. The different classification algorithms are SVM, RandomForestClassifiers, SGDC will be used. After comparison, I will select the best model.

Benchmark Model

I will use SVM as my benchmark model. If SVM gives the best score in my chosen metrics than other models, I will consider this as my model.

Evaluation Metrics

As I have mentioned above it is really important to predict effected people accurately. Suppose after dividing our total set into training set and testing set. What if we are left with the training set in which we have 80% of set are benign and 20% of set are malignant. In this

case, our model tries to predict a record as benign always which gives 80% accuracy score. That's a good score. But it's actually a terrible model. We are actually predicting an effected person as benign which is a terrible mistake. So, its not a good practice to use accuracy score as metric for this. Instead we can go with **confusion matrix** and **f_score.**

I will use F1 score and confusion as evaluation metrics.

F1 Score = 2 * (precision * recall) / (precision + recall)

Precision = (TruePositives)/(TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegative)

Project Design

Firstly, I will load csv file to a dataframe using panda library. After that, its data exploration stage. In this stage I'll get familiar with the data through visualization techniques using python libraries (Pandas,matplotlib). This is really important because it will provide useful information for data pre-processing. Next stage is pre-processing our data. In this stage, we will use feature selection to reduce high-dimension. Our main goal is to prepare the data and make it suitable for performing predictions. Next I'll construct a predictive model using SVM machine learning algorithm to predict the diagnosis of a breast tumor. The diagnosis of a breast tumor is a binary variable (benign or malignant). I'll also evaluate the model using fscore, confusion matrix the receiver operating curves (ROC), which are essential in assessing and interpreting the fitted model.

References

- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

- https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance