Machine Learning Engineering Nanodegree

Capstone Proposal

Hsiao-Tien Fan

June/25/2017

Domain background

Breast cancer is the second most common type of cancer behind skin cancer in the United States, with an estimated number of a quarter of a million cases expected to be diagnosed in 2017. [1] The diagnosis for breast cancer is usually conducted with a variety of tests such as a physical examination of the breast, mammograms, ultrasounds or biopsy. The selection of which of these tests is to be conducted is at the discretion of the doctor depending on the situation of a particular case. [2] However, it has been shown in previous studies that by analyzing the cytological information obtained from images of biopsied cells using Fine Needle Aspiration (FNA), it is possible to determine whether the sample is cancerous [3][4]. This project will attempt to apply machine learning algorithms datasets obtained from the FNA method and attempt to classify the tumor.

Problem Statement

The goal of this project is to create a python based model that is able to assess whether the biopsy from a subject's tumor is benign or malignant. The python script will:

- Load the dataset and preprocess the data for training.
- Train a classifier to determine if a subject is benign or malignant.

Datasets and Inputs

The dataset used for this project is available from the UCI Machine Learning Repository or from Kaggle's database section [6][7]. There are 569 samples in total, with 357 benign and 212 malignant. Each sample has the following features where each is in relation to the 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)

Solution Statement

For this project, as the number of features is relatively low, it may be a good idea to start with a decision tree and evaluate the performance from there. The decision tree will be trained with a subset of the dataset to predict whether the sample is benign or malignant, with the remaining subset used for testing. The quality of the solution will be measured by the performance of the model's prediction on the test set. The evaluation metrics for determining the performance will be described below.

Benchmark Model

A simple benchmark model to use would be the naïve predictor. For this project to be meaningful, it needs to perform better the probability of getting the correct result from purely guessing. The proposed benchmark is the probability of being correct when always assigning the prediction to be malignant. This was chosen, as with cancer prediction, it is better to have false positives rather than false negatives.

Evaluation Metrics

Accuracy - This metric tells us the amount of correct predictions made by the classifier by looking at the true positives and true negatives.

$$accuracy = \frac{No.of\ true\ positives + No.of\ true\ negatives}{Dataset\ size}$$

False negative rate – This metric is important for this particular application as for cancer detection, the worst case scenario is to not recognize the presence of cancer and declare the patient healthy. It is important to tune the model so that the false negative rate is as low as possible.

$$False\ negative\ \% = \frac{No.of\ negatices + No.of\ false\ negatives}{No.\ of\ negatives}$$

Project Design

For this project, the first aspect that needs to be addressed is the nature of the data. Data exploration needs to be performed and possibly visualized in the form of graphs to observe any skews in the data. Proper normalization techniques need to be applied to preprocess the data before it can be used for training.

Following the preprocessing, the data will be applied to a classifier for training. The proposed method of choice is the decision tree, which is easy to train and performs well on classification problems. 80% of the total dataset will be used for the training process while the remain 20% will be used for testing. The testing set will be further split into 5 groups for cross validation to prevent overfitting, which is a common issue with decision trees.

The evaluation of the model will be based on the accuracy and the percentage of the false negatives with respect to the total number of predicted negatives. Obviously the accuracy of the system is important, as it determines whether the model is useful in achieving the intended application. It is also important to observe the false negative percentage, as it is important for the system to minimize the number of malignant diagnosed as benign.

References

[1]http://www.breastcancer.org/symptoms/understand_bc/statistics

[2]http://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/diagnosis/dxc-20207942

[3]https://en.wikipedia.org/wiki/Fine-needle_aspiration

[4]Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, *43*(4), 570-577.

[5]W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, U.S.A., 87:9193-9196, 1990.

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

[7] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data