Machine Learning Engineer Nanodegree

Capstone Proposal

CL. Chen
March 15st, 2018

Proposal

Breast Cancer Prediction

(approx. 2-3 pages)

Domain Background

(approx. 1-2 paragraphs)

Breast cancer is the most prevalent diseases and common cancer of the women in the worldwide. Recently, the death rate of breast cancer has gradually increased. Approximately 60% of the patients with breast cancers are able to survive for ten more years with early diagnosis coupled with appropriate treatment. It is important to accurately diagnose early to increase the survival rate. With technology continuously progress, data mining and classification method can be used to process a huge amount information of hospital patients. Now, we can use machine learning method to extract the features of cancer cell nuclei image and classify them. It would be helpful to determine whether a given sample appears to be benign ("B") or malignant ("M").

The related academic paper: Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis by Hiba Asri et al. / Procedia Computer Science 83 (2016) 1064 – 1069

link https://doi.org/10.1016/j.procs.2016.04.224

Problem Statement

(approx. 1 paragraph)

In this problem we have to predict the Stage of Breast Cancer M (Malignant) and B (Bengin). The project focus on finding classifier that can accurately determine malignancy of sample. The Breast Cancer (Wisconsin) Diagnosis dataset is used for training. 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.

Datasets and Inputs

(approx. 2-3 paragraphs)

This database is also available through the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/. Also can be found on UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

The Breast Cancer (Wisconsin) Diagnosis dataset contains the diagnosis and a set of 30 features describing the characteristics of the cell nuclei that present in the digitized image of a of a fine needle aspirate (FNA) of a breast mass. There are ten features with real value which are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter);

2. texture (standard deviation of gray-scale values);

3. perimeter;

4. area;

5. smoothness (local variation in radius lengths);

6. compactness (perimeter^2 / area - 1.0);

7. concavity (severity of concave portions of the contour);

8. concave points (number of concave portions of the contour);

9. symmetry;

10. 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

(approx. 1 paragraph)

All data points are labeled in the dataset. We can use supervised learning to solve this problem. In this case, all samples will be divided into two groups (Malignant and Bengin) which have the binary result. Some binary calssifier will be able to used like LR (logistic regression), SVM, Random Forest.... All dataset will be separated into 70% training and 30% test set randomly then compute their accuracy and F1 score. Feature selection will be performed to reduce the numbers of dimensions and to increase its accuracy.

Benchmark Model

(approximately 1-2 paragraphs)

This dataset was analyzed in the public Kaggle. I choose Buddhini W's 'Breast cancer prediction' as my benchmark model. In her analysis, the best model to be used for diagnosing breast cancer as found is the Random Forest model with the top 5 predictors, 'concave points_mean','area_mean','radius_mean','perimeter_mean','concavity_mean'. It gives a prediction accuracy of ~95% and a cross-validation score ~ 93% for the test data set.

Evaluation Metrics

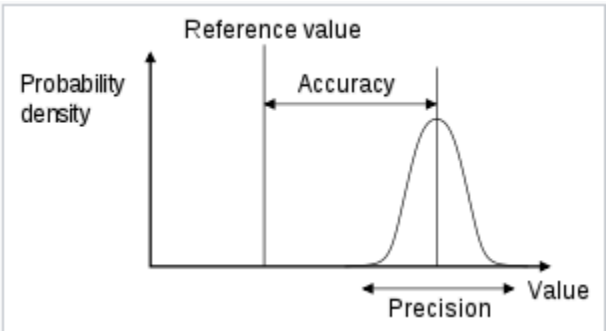
(approx. 1-2 paragraphs)

Accuracy and F1 score would be used to compare the result with the benchmark model.

Accuracy: the ratio of how many samples are correctly classified to all samples.

F1 score: the harmonic mean of precision and sensitivity

$$Accuracy = \frac{True\ Positive + True\ Negative}{All\ samples}$$



$$F = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Project Design

(approx. 1 page)

This project will be separated into several steps. In sequence they are data exploration, feature visualization, feature selection,data preprocessing and model classification. For data exploreation, I will have a look at the data. Explore the data type and check if ther is any data column unnecessary then drop them and realize the counts distribution of the diagnosis in this data set. For feature visualization, I will observe the data and analyze the feature correlation through the correlation plot like heatmap,scatter plot because the correlated features may lead to model failed. For feature selection, I may use random forest model to select the top5 important features to improve the model efficiency. For data preprocessing, I will scale the data before using SVM model. For model classification, all dataset will separated into train and test randomly. Perform k-fold cross-validation with different classification model, such as Logistic Regression, Random Forest, Gaussian Naive Bayes, Adaboost and SVM (Support Vector Machine) to find the best model based on accuracy and F1 score.

Reference

Kaggle, Breast cancer prediction, Buddhini Waidyawansa (12-03-2016)

https://www.kaggle.com/buddhiniw/breast-cancer-prediction