Machine Learning Engineer Nanodegree

Capstone Proposal

Jeyvison Nascimento

May 06, 2017

Proposal

Domain Background

Breast cancer is the most dangerous type of cancer,besides lung cancer,a woman may develop. It's death rate is high and it is expected that 12% of the american women will develop it during their lifetime.

Problem Statement

The goal is to create a program to help the doctor choose a protocol to deal with breast cancer when it's malignant or benign.Many informations will be considered like radius, perimeter and smoothness. With those information the cancer will be classified as malignant or benign and the doctor can decide which steps to take next

The program will:

- Train the model based on a dataset
- Identify if the cancer is malignant or benign
- Inform the doctor about the result

Datasets and Inputs

The dataset that will be used has 30 features and a binary target feature that tells if the cancer was malignant or benign. The input of the program will be a csv containing values for classification.

The dataset has 569 instances and is unbalanced so it will need to be stratified.For training purposes, the training will be splitted on 70% for training and 30% for testing using K-fold cross validation.

Here follows the list of the 30 features:

1. radius_mean
2. texture_mean
3. perimeter_mean
4. area_mean
5. smoothness_mean
6. compactness_mean
7. concavity_mean
8. concave points_mean
9. symmetry_mean
10. fractal_dimension_mean
11. radius_se
12. texture_se
13. perimeter_se
14. area_se
15. smoothness_se
16. compactness_se
17. concavity_se
18. concave points_se
19. symmetry_se
20. fractal_dimension_se
21. radius_worst
22. texture_worst
23. perimeter_worst
24. area_worst
25. smoothness_worst
26. compactness_worst
27. concavity_worst
28. concave points_worst
29. symmetry_worst
30. fractal_dimension_worst

Solution Statement

As we have many real-valued features in the dataset we will need to scale these values.SVC(SVM) will be used to perform the classification because we have a small dataset. Naive Bayes could be used for this one too but it does not provide a reliable probability prediction which can be useful for choice making.

Benchmark Model

For benchmarking, a Scikit learn Dummy classifier will be used to compare the results with the SVC classifier.

Evaluation Metrics

We have two metrics we may use to evaluate the model, Accuracy andw F1 score. Both will be used to measure the model's performance.

Project Design

The steps needed to develop the project are the following:

- Develop a simple python script
- Make the script load the dataset file(CSV)
- The dataset is given to the classifier for training
- The script loads a CSV evaluation file and uses its content for prediction
- The result showing the type of the breast cancer and the prediction probability will be shown in the screen.