Machine Learning Nanodegree
Capstone Proposal

Version: v1.0
Authors: edgar.martinez.rico@gmail.com (Edgar Martinez Rico)
Status: In Progress
Last Updated: 2017/05/15

Proposal

Classification of breast cancer among women.

Domain Background

Cancer is a disease in which abnormal cells grow invading healthy cells in the body[1]. Cells constitutes the basic building blocks to build tissue and organs in the body. The process of cell growth could at times go wrong and lead into the development of new cells when the body does not need them. Or when old or damaged cells do not properly die. This will result in the accumulation of cells that forms a mass of tissue that is referred as tumor, lump or growth. Breast cancer begins with the formation of cancer cells in the breast that invade the nearby tissues or spread (metastasize) to other areas of the body. These cells can spread from the original malignant tumor if they entered into the blood vessels or lymph vessels.

The recognition of breast cancer as early as possible provides the better chance of treatment. As a result, knowing your breast is key, it is important to understand how your breast look and feel. Additionally regular mammograms and screening tests help to identify breast cancer in its early stage even before symptoms appear. The most common symptom is the development of a new lump or mass. This mass could be painless, tender, soft, rounded or have irregular edges which are more likely to be cancer.

Men breast cancer is rare, less than one percent of breast cancer are develop in men. According to the American Cancer Society breast cancer is the second leading cause of cancer death in women in the United States. The American Cancer Society estimates that 1 in 37 women will die from breast cancer[2]. On average, every 2 minutes a woman is diagnosed with breast cancer and one woman will die of breast cancer every 13 minutes. And according to the World Health Organization breast cancer is the most common cancer in women in developed and developing countries[3].

Problem Statement

The diagnosis of breast cancer is a critical step. In many cases this can be diagnose with the use of a diagnostic mammogram, breast ultrasound or breast MRI. However, sometimes a biopsy might be required. There are two types of biopsies to diagnose breast cancer. Fine Needle biopsy removes cells from a suspicious tumor in the breast. Surgical biopsy provides the complete information in regards a lump and is the most accurate way to diagnose breast cancer.

---

[1] "What Is Cancer?" The National Breast Cancer Foundation.
[2] "How Common is Breast Cancer?" American Cancer Society.
[3] "Breast Cancer: Prevention and Control." World Health Organization.

However, the diagnosis with the fine needle biopsy has mixed success[4]. Different features such as training, experience and technical expertise of the physician performing the diagnosis[5]. Therefore, machine learning can be used for the diagnosis to determine if a lump is malignant or benign. This can be achieve with the use of a supervised learning technique that allow the classification of a lump in either benign or malign. The supervised learning model can be built taking in consideration different features[4] of a lump such as radius, perimeter, area, compactness and symmetry to name a few.

Datasets and Inputs

The data set for this project will be downloaded from Kaggle[6]. This dataset is divided into two main classes of tumors malign or benign. Each tumor is described using several features computed from digital images taken from a fine needle aspirate of a breast lump. These describes the characteristics of a cell nuclei in a three dimensional space[7]. The characteristics include radius, texture, perimeter, smoothness, compactness and concavity to name a few. The data consist of 357 benign tumors and 212 malignant tumors. The data consist of 30 features in which all them are numeric with the exception of the *diagnosis* feature. The feature called "id" does not provide value for the model since it represents a unique identifier. As a result, this feature will be dropped from the model. The target feature *diagnosis* has two values *M* for malign and *B* for benign. Such feature will be extracted and the remaining 28 features will be used for training and testing the model.

Solution Statement

A supervised learning approach will be used to model the characteristics of a tumor. A classification model will be built to be able to determine if the characteristics provided of a tumor belongs to either benign or malign lump. This model should be able to generalize well enough for unseen data as a result, it should predict the corresponding output. Different supervised learning algorithms will be evaluated to determine the best model. Among the algorithms that will be evaluated are Gaussian Naive Bayes, Support Vector Machine, Decision Trees to name a few[8].

---

[4] Frable, William J. *Thin-needle Aspiration Biopsy*. Philadelphia: Saunders, 1983.
[5] "Cytopathology Palpation Guided Fine Needle Aspiration Analysis of FNA Procedure and Diagnosis" Pathology Outlines.
[6] *"Breast Cancer Wisconsin (Diagnostic) Data Set" Kaggle.*
[7] Bennett, Kristin, and O. L. Mangasarian. "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets."
[8] *Choosing the Right Estimator — Scikit-learn 0.18.1 Documentation*.

Benchmark Model

Sklearn provides a Dummy Classifier which can be used as a baseline. This classifier performs predictions using rules. This dummy classifier will be used to compare the results with the real classifier model. The strategy that will be used for the dummy classifier will be *stratified.* This strategy generates predictions by respecting the training set's class distribution. Another approach is to compare prior results published on kaggle. However, some of the results are not published completely.

Evaluation Metrics

To evaluate the performance of each algorithm four measures will be used to properly calculate the F1 score:

		Predicted Class	
Actual Class		*Malign*	*Benign*
	Malign	TP	FN
	Benign	FP	TN

The measure above can be explained as follow:

- TP: A True Positive is when a malign tumor is predicted and is an actual malign tumor.
- FP: A False Positive is when a benign tumor is predicted as malign tumor.
- FN: A False Negative is when a malign tumor is predicted as benign tumor.
- TN: A True Negative is when a benign is predicted and the diagnosis is benign tumor.

Project Design

1. Obtain the information from kaggle.
2. Analyze the data using statistics to understand the data. Determine if there are outliers on the data. Remove any possible outliers from the data.
3. Split the data (features and its correspondent labels) into training data and testing data.
4. Use sklearn and explore different algorithms such as Gaussian Naive Bayes, Support Vector Machine, Decision Trees, Random Forest to name a few.
5. Train each of the algorithms from the prior step and test each model to determine which one provides better results.
6. Once an algorithm has been identified to predict better it will be further improved using Grid Search to guarantee the best results.

References

- Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle. Web. 15 May 2017.
- Bennett, Kristin, and O. L. Mangasarian. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. Optimization Methods and Software 1.1 (1992): 23-34. Print.
- Breast Cancer: Prevention and Control. World Health Organization. World Health Organization. Web. 14 May 2017.
- Choosing the Right Estimator. Choosing the Right Estimator — Scikit-learn 0.18.1 Documentation. Web. 21 May 2017.
- Frable, William J. Thin-needle Aspiration Biopsy. Philadelphia: Saunders, 1983. Print.
- General Cytopathology - Analysis of FNA Procedure and Diagnosis. Web. 15 May 2017.
- How Common Is Breast Cancer? American Cancer Society. Web. 14 May 2017.
- Nbcf. What Is Cancer? The National Breast Cancer Foundation. www.nationalbreastcancer.org. Web. 14 May 2017.