Breast Cancer Prediction

Dane Acena

Breast cancer is one of the most common cancer diagnosed in women in the United States[1]. Diagnosis of breast cancer is performed when an abnormal lump is found, either from self-examination or x-ray. After the detection of suspicious lump, the doctor will conduct a diagnosis to determine whether it is cancerous[2].

I want to train an algorithm to predict whether a certain mass detected is either malignant or benign using several features of the mass.

For this project, I will need a dataset that contains information regarding malignant and benign (class) breast cancer masses. Each of the instances in the data has a class assigned with features that describes the certain mass.

The dataset that will be used for this project will be gathered from Breast Cancer Wisconsin Data Set[3] available from UCI Machine Learning Repository[4]. This data set has 699 instances, with 10 attributes plus the class attribute.

1. Sample code number	Id number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class	2 for benign, 4 for malignant

To train this algorithm to predict breast cancer I will use a logistic regression algorithm.

To evaluate the results, I will split the dataset into 80% training set and 20% test set. I will also compare the results of my logistic algorithm with the logistic regression from sklearn.

[1] https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[2] https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset

[3] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[4] William 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., Volume 87, December 1990, pp 9193-9196.