Classification of Cancer Cells Using Machine Learning Model (SVM)

Arunima Varma

September 19, 2020

1. Introduction

1.1 Background

Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. These cells usually form tumours that can be seen via X-ray or felt as lumps in the breast area.

Early diagnosis significantly increases the chances of survival. The key challenges upon detection is how to classify tumours into malignant or benign. A tumour is considered malignant if the cells can grow into surrounding tissues or spread to distant areas of the body. A benign tumour does not invade nearby tissue nor spread to other parts of the body the way cancerous tumours can.

Machine Learning is a sub-field of Artificial Intelligence that gives systems the ability to learn themselves without being explicitly programmed to do so. Machine Learning can be used in solving many real-world problems. ML techniques can dramatically improve the level of diagnosis in breast cancer. Research shows that experienced physicians can detect cancer by 79% accuracy, while a 91% accuracy can be achieved using ML techniques.

1.2 Problem

In this study, my task is to classify tumours into malignant (cancerous) or benign (non-cancerous) using features obtained from several cell images.

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.

2. Data Acquisition and Cleaning

2.1 Data Sources

Scikit-learn comes with a few small standard datasets that do not require downloading any file from any external website. The dataset that I will be using for our machine learning problem is the Breast cancer Wisconsin (diagnostic) dataset. The dataset includes several data about the breast cancer tumours along with the classification labels, viz., malignant or benign.

2.2 Data Cleaning

Attribute Information:

1. ID number
2. Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

1. Radius (mean of distances from centre to points on the perimeter)
2. Texture (standard deviation of grey-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)

2.3 Feature Selection

The important attributes that we must consider from that dataset are ‘target-names'(the meaning of the labels), ‘target'(the classification labels), ‘feature names'(the meaning of the features) and ‘data'(the data to learn).

To get a better understanding of what the dataset contains and how we can use the data to train our model, let us first organize the data and then see what it contains.

So, we see that each dataset of a tumour is labelled as either ‘malignant’ or ‘benign’. From here, , each label is linked to binary values of 0 and 1, where 0 represents malignant tumors and 1 represents benign tumours.
Here, there are 30 features or attributes that each dataset of the tumor has. We will be using the numerical values of these features in training our model and make the correct prediction, whether or not a tumor is malignant or benign, based on these features.