Support Vector Machine (SVM)

Saurabh Burewar (B18CSE050)

Dataset

Link to the dataset: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

The dataset describes characteristics of the cell nuclei present in the digitized image of a fine needle aspirate of a breast mass. It has ID for every entry, a list of attributes to describe the characteristics and the final diagnosis.

These ten real-valued features are computed for each cell nucleus:

Features	Definition	Key
Radius	Mean of distances from center to points on perimeter	
Texture	Std deviation of gray-scale values	
Perimeter	Perimeter of the mass	
Area	Area of the mass	
Smoothness	Local variation in radius lengths	
Compactness	Perimeter^2/area - 1.0	
Concavity	Severity of the concave portions	
Concave points	Number of concave portions	
Symmetry	Symmetry exists or not	
Fractal dimension	Coastline approximation - 1	
Diagnosis (Target)	Malignant or Benign	M = Malignant, B = Benign

		M = 1, B = -1

🟨 Features used in the training of the models

🟥 Target variable

Dataset Processing

- The last column in the dataset is an unnamed attribute with null values, so we are dropping that column. Also, the “ID” is not needed, so we are dropping it as well.
- For numerical data, the “Diagnosis” feature is mapped as { M: 1, B: -1 }.
- The features and the output are separated and the features are normalized. Now, we use this normalized version as the data.
- The data is split into a 70:30 ratio for training and testing.

Features used

All the features present in the data are used except the ID number and the last unnamed feature.

Classification

The task here is to classify the data, that is find a hyperplane that can separate the malignant and the benign samples. So, the target variable is “Diagnosis” which shows 1 for Malignant and -1 for Benign. Therefore, it is a case of binary classification.

Performance

Hard SVM by solving QP

Accuracy of the model: 40.35%

Soft SVM by solving QP

Accuracy of the model: 45.35%

Soft SVM with SGD

Accuracy of the model: 59.64%