Breast Cancer Prediction - ML Techniques Explained

Breast Cancer Prediction - Machine Learning Components Explained

1. Logistic Regression:

Logistic Regression is a linear classification model used for binary classification tasks. It works by computing a weighted sum of input features, then applying the sigmoid function to map this sum to a probability between 0 and 1. The model predicts class 1 if the output is greater than a threshold (commonly 0.5), otherwise class 0.

2. Sigmoid Function:

The sigmoid function is defined as:

sigmoid(z) = 1 / (1 + exp(-z))

It smoothly maps any real-valued number to the range [0, 1], which makes it ideal for expressing probabilities in logistic regression.

3. LabelEncoder:

In machine learning, categorical variables need to be converted into numerical format. The LabelEncoder is used to convert labels such as 'M' (Malignant) and 'B' (Benign) into numeric values (1 and 0 respectively), which the model can process.

4. fit_transform vs transform:

- fit_transform(): This method first learns the parameters from the data (e.g., mean and standard deviation for scaling) and then applies the transformation.
- transform(): This method applies the learned transformation to a different dataset (e.g., test set), ensuring consistency between train and test data.

5. StandardScaler:

StandardScaler is used to scale numerical features so that they have a mean of 0 and a standard deviation of 1. This is essential for models like logistic regression that are sensitive to the scale of features.

Example:

Original values: [10, 20, 30]

Standardized: [-1.22, 0, 1.22]

6. Why Standardize Features?

Unscaled features with large ranges can dominate features with small ranges. Standardizing puts all features on the same scale, improving the model's ability to converge and ensuring fair influence from each feature.

7. accuracy_score:

This metric measures the percentage of correct predictions made by the model. While simple, it can be misleading in imbalanced datasets.

8. classification_report:

This function provides a summary of precision, recall, F1-score, and support for each class. It gives a more complete view of performance, especially in medical datasets where false negatives and false positives carry different costs.

9. Confusion Matrix:

A confusion matrix shows how predictions compare to actual labels in a 2x2 format:

Predicted

```
| 0 | 1
----------------
True 0 | TN  | FP
True 1 | FN  | TP
```

Where:

- TN = True Negative, FP = False Positive
- FN = False Negative, TP = True Positive

10. Precision and Recall:

- Precision = TP / (TP + FP): How many predicted positives were truly positive?
- Recall = TP / (TP + FN): How many actual positives were correctly identified?

11. ROC-AUC:

ROC (Receiver Operating Characteristic) curve plots the true positive rate against the false positive rate across thresholds. AUC (Area Under Curve) reflects the model's overall ability to distinguish between classes. AUC = 1.0 indicates perfect classification.

12. Threshold Tuning:

The default threshold for classifying probabilities is 0.5. However, this may not always be optimal. By adjusting the threshold, we can balance precision and recall based on the application's needs (e.g., high recall in cancer detection to avoid missing positives).

13. Threshold Tuning Results:

The threshold was varied from 0 to 1 and metrics like precision, recall, and F1-score were evaluated. The best F1-score occurred at threshold ~ 0.38. This resulted in:

- Precision: 0.9767

- Recall: 0.9767

- F1 Score: 0.9767

- ROC-AUC: 0.997

14. Summary:

This project demonstrates the importance of not only selecting the right model (logistic regression) but also applying correct preprocessing (scaling, encoding), choosing suitable metrics, and fine-tuning classification thresholds to meet domain-specific requirements.