Feature Construction

Mohammad Reza Ahmadizadeh
Second Assignment

October 25, 2025

1 What Will I Do?

For this week's assignment we have **Feature Construction**.

Creating new features based on the existing features for improving feature space and to see if the newly added feature have what impact on our result.

2 What Is My Approach?

For this I have to options:

Option 1: Add the New Features

I saw that in In most feature constructions, They keep all the original features and simply add the new ones as extra columns.

- The goal of feature construction is to enrich your dataset, not replace it
- You'll then evaluate whether these additional features improve model performance.

Option 2: Replacing the old features

If i do this I will:

- The new features summarize or combine correlated ones (to reduce dimensionality).
- I plan to focus on interpretability rather than raw accuracy.

With this case being said I will choose to add the new freature to already existing ones because the newly constructed features were added to the original dataset to enrich the feature space.

3 Feature Construction

Based on the article that I read, It was said the size related feature such as; *area*, *radius* and *parimeter* have strong influence on breast cancer classification.
And with this being said several new features were constructed.

1. **Area-to-Perimeter Mean Ratio (area_to_perimeter_mean)**

   Formula:

   $$area\_to\_perimeter\_mean = \frac{area\_mean}{parimeter\_mean}$$

   Meaning:

   This ratio describes the compactness of a cell nucleus. A higher value suggests a larger area relative to its perimeter, which may indicate a more irregular or enlarged cell — often associated with malignant tumors.

2. **Radius Range (radius_range)**

   Formula:

   $$radius\_range = radius\_worst - radius\_mean$$

   Meaning:

   This feature measures the variation between the average and the worst radius measurement for each sample. Larger differences may represent higher irregularity in cell size, which can be a sign of malignancy.

3. **Area Range (area_range)**

   Formula:

   $$area\_range = area\_worst - area\_mean$$

   Meaning:

   Captures how much the cell area changes between mean and worst observations. This reflects heterogeneity — a typical characteristic of malignant cells.

4. **Concavity-to-Compactness Ratio (concavity_to_compactness_mean)**

   Formula:

   $$concavity\_to\_compactness\_mean = \frac{concavity\_mean}{compactness\_mean}$$

   Meaning:

   Shows how much concavity (degree of inward curvature of the cell boundary) exists relative to compactness. A higher ratio indicates more concave, less compact cell structures, possibly indicating malignancy.

5. **Severity Index (severity_index)**

   Formula:

   $$severity\_index = area\_worst + concavepoints\_worst$$

   Meaning:

   Combines two of the most influential predictors identified by Hoque et al. (2024) — "area_worst" and "concave_points_worst" — into a single index. It represents an overall measure of cell abnormality severity.

6. **Shape Complexity Mean (shape_complexity_mean)**

   Formula:

   $$shape\_complexity\_mean = mean(smoothness\_mean, compactness\_mean, concavity\_mean, concavepoints\_mean)$$

   Meaning:

   Summarizes shape-related characteristics of each sample into one composite measure. A higher value indicates more structural irregularities in the cell nucleus.

Each of these constructed features is designed to enhance the separability between **benign** and **malignant** cases by combining or comparing existing measurements in meaningful ways.

The features were created in an `.ipynb` file using the `pandas` liberary and `df` (DataFrame) method. The result is saved in a new csv file named `feature_constructed_dataset.csv`.