Cancer Detection -- Wisconsin Diagnostic Breast Cancer (WDBC) Malignancy Detection

Situation: The application of machine learning based medical diagnosis based on radiology data is a very hot topic at the intersection of AI and medicine. We will use a classic dataset that contains breast cancer diagnosis performed on 569 patients to see if we can train an algorithm to detect the presence or absence of breast cancer in a new patient. As you will see in the dataset below there is an attribute (column) called 'Class' that takes on one of two values, 'Benign' (357 cases), and 'Malignant' (212 cases). This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The 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.

Complication: There are many regions of the world that do not have access to trained radiologists. Can digitized images of a fine needle aspirate (FNA) of a breast mass combined with machine learning be used to detect breast cancer? There are 10 FNA real-valued features that are computed for each cell nucleus. These are:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

For each of these 10 features we also compute the mean, standard error, and "worst" or largest (mean of the three largest values) of these features, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

Are these features useful in *accurately* classifying a new cell into 'Benign' or 'Malignant?'

Key question: Can we use AI instead of a trained radiologist to detect breast cancer given the FNA of the patient's breast mass?

Solution approach: We are going to use this dataset to dive into the fundamentals of a particular class of very popular prediction problems, namely **binary classification**. Our goal is to learn to build and evaluate the performance of binary classification models from a predictive angle. We will start with basic models such as k-nearest neighbors, decision trees and logistic regression. The first two are classic ML models, whereas logistic regression is a classical statistical model.

Deck: Predictive modeling basics

Dataset[1]: wdbc.data **Data Dictionary**: see folder

R Code: see folder (R file)

Discussion questions:

1. There are so many machine learning models for binary classification. Which of these should we start with?
2. Should we care about model explainability for this problem? In general? How should we think about trading off model accuracy with explainability?
3. How should we choose the best value of *k* in k-nearest neighbors?
4. What is a general data splitting strategy for prediction problems?
5. Why do we need cross-validation when we have sample splitting between train and test?
6. What/which data should we use at model deployment stage?

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29