Itzel Salgado

Homework 2

Introduction: With the exception of skin cancer, breast cancer is the most common cancer which women are diagnosed with. 1 in 8 women will be diagnosed with breast cancer in their lives; therefore, it is important to determine new ways to diagnose this disease since it is something which is quite common. In order to detect breast cancer, tests such as mammograms, ultrasounds, and other forms of imaging are performed. At times, these images are enough to determine whether tumors are malignant or benign; however, there are times when biopsies are needed, which require surgery and may also leave scarring which is inconvenient when tumors are benign. A solution to this could be a mechanistic data science approach where imaging can be used to determine a pattern between benign and malignant tumors based on the imaging performed and past knowledge of these tumors.

Objective: For this project, a data driven model will be used to determine whether tumors from breast cancer are benign or malignant based on the features which are given in an existing database. The goal of this project is find the correlation between the given features and malignant or benign tumors to improve the detection of breast cancer.

Multimodal data generation and collection: An existing database will be used to develop this model. The database was obtained by the University of Wisconsin Hospitals in Madison from Dr. William H. Wolberg and can be accessed at the following link <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>. This database has the normalized values for the clump thickness, uniformity of the cell size, uniformity of the cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and whether the cell is malignant or benign. There are a total of 699 data points in this database.

Featuring engineering: From the collected data, the features will be used to determine which inputs provide the best correlation to determining whether the given tumors are malignant or benign.

Dimension reduction, regression, and classification: Since there are a large amount of features in this model, the model is looking at a variety of features, there are 2 main groups these features can be divided into: measurements (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, and single epithelial cell size) and the cell properties (bare nuclei, bland chromatin, normal nucleoli, and mitosis). For this model, we are planning on using regression to determine the relationships between the malignant/benign tumors and these features.

System and design: The aim of this model would be to determine whether tumors from breasts are malignant or benign based on the given features and to determine if there are certain features which lead to a greater correlation when compared to others.

Summary: A model will be created to determine if a tumor is malignant or benign based on the database provided by Dr. William H. Wolberg. A mechanistic data science solution will be used to classify the breast tumors and improve the current detection methods of breast cancer.

Itzel Salgado

References:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

[1] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[2] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

[3] O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

[4] K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).