Project 25% Status Report: Breast Cancer Wisconsin

Ziyuan Shen

March 19, 2019

1 Problem Description

Over the past few decades, machine learning techniques have been increasingly applied to disease studies [1]. The advanced image technology provides increasing clinical image information. The development of single-cell technology greatly facilitates the availability of multi-omics data [2, 3]. Accordingly, delicately designed data mining and machine learning techniques are desired to explore the biological conclusions underlying the large datasets. Relevant studies can aid disease diagnosis and precision medicine, as well as disease early detection. Cancer classification is one of the top promising topics [4]. By far, various biological datasets [5, 6] have been used for cancer classification studies. Improving the classification performance is undoubtedly meaningful as human beings' overall health conditions will be improved by these efforts.

Cancer classification historically requires prior biological knowledge. However, most machine learning engineers do not have a biological background. Therefore, the general classification performance is worth improving. Prediction models with high accuracy are aimed to aid oncologists with diagnosis and prognosis.

The goal is to create a good predictor with samples from known classes, and possibly identify hidden cancer subtypes, without any biological priors. Most existing works develop models based on specific cancer types [5, 7, 8]. Thus, a more general and systematic procedure remains open-ended. Although biological prior is not require for performing classification, different estimators will yield different performances when tested on a certain dataset. The task is to explore the data structure, choose appropriate algorithms, compare different algorithms and analyze the result.

2 Data Description

2.1 Background

Breast Cancer Wisconsin is a classic caner dataset for classification and has been explored by many machine learning researchers for testing algorithms [8–13]. The dataset is composed of digital image information of breast cancer cell nuclei. The cells are labeled as malignant and benign.Therefore, cancer subtypes have the potential to be classified with only cells data, without any prior biological knowledge. The dataset contains two classes, thus should serve as a good example for conducting classification algorithms. According to the original paper [9] generating this dataset, discrimination between the two classes is hard by one single characteristic. By integrating all features, the two masses are classifiable.

2.2 Summary Statistics

The data contains 569 samples in total with two classes: malignant cells and benign cells. The original image provides 10 characteristics for each sample: radius, texture, perimeter, area, smoothness,

compactness, concavity, concave points, symmetry and fracture. The mean value, standard error, and mean of the three largest values are calculated from the original data, respectively, for each data sample. Therefore, 30 features are generated for each sample. Statistics of the dataset are summarized in Table 1. Statistics of the first ten features are summarized in Table 2.

Table 1: Summary Statistics of Breast Cancer Datasets.

Classes / Statistics	Malignant	Benign
#Characteristic	10	
#Features	30	
#Total Samples	569	
#Samples	212	357
#Proportion	37.3%	62.7%

Table 2: Summary Statistics of The First Ten Features.

Statistics	radius mean	texture mean	perimeter mean	area mean	smoothness mean	compactness mean	concavity mean	concave points mean	symmetry mean	fractal mean
Mean	14.12	19.28	91.97	654.89	0.096	0.104	0.089	0.049	0.18	0.063
Median	13.37	18.84	86.24	551.10	0.096	0.093	0.062	0.034	0.18	0.062
Std	3.52	4.30	24.3	351.9	0.014	0.053	0.080	0.039	0.027	0.007

To explore the correlation between these samples, the correlation heat map of the first 10 features, which are the mean values of the 10 characteristics, is plotted in Figure 1. By observing the heat map, radius, perimeter and area are highly correlated. We can also draw this conclusion from the definition of these three characteristics. Therefore, we can choose any of these three features when performing classification. Similarly, compactness mean, concavity mean and concave points mean are highly related.

2.3 Data Visualization

2.3.1 Pairwise Scatter Plot

By dropping features with high correlation, six features are extracted for pairwise scatter plot as depicted in Figure 2. From the graph, statistics of the two classes overlap each other but should be separable if treated with appropriate classifier.

2.3.2 t-SNE

According to Table 2, different features of this dataset have entirely different magnitudes. Therefore, some preprocessing might be necessary before visualizing the data. In order to conduct data visualization, the data is firstly reprocessed with z-scoring so that the processed data has zero mean and standard error. T-SNE [14] visualization is shown in Figure 3. A 3D plot for three components is provided as well as 2D plots for every two components.

2.3.3 Principle Component Analysis

As the number of features of this dataset is high, Principle Component Analysis (PCA) [15] is appropriate for visualizing the data. According to PCA, the first three components only contains 72% of the variance, which is large enough to represent the main structure of the data. PCA data visualization is shown in Figure 4.

2.4 Source

The source data can be downloaded from UCI machine learning repository [16] or kaggle website [17].

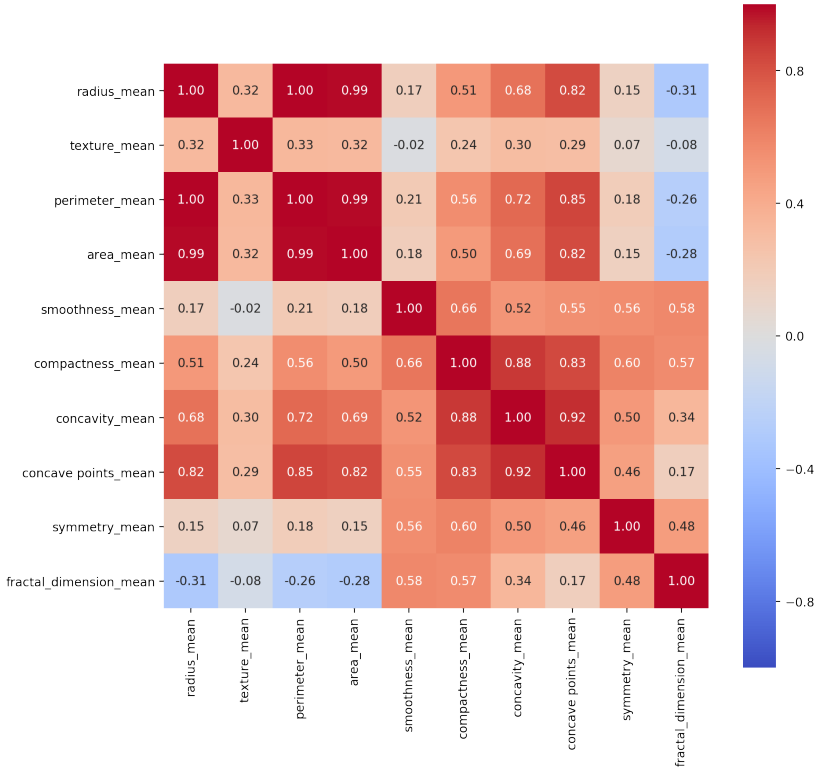

Figure 1: Correlation heatmap.

3 Anticipated Technical Approach/Method

3.1 Preprocessing

As the data features have entirely different scales, preprocessing is important for obtaining accurate prediction performance.
Normalization: z-Scoring may be applied to scale data to zero mean and unit variance.

3.2 Dimension Reduction

As the dataset has high dimensions, an approach for dimension reduction will be important to data visualization and the prediction performance.
Principle Component Analysis: Principle Component Analysis (PCA) [15] will be a useful tool for dimension reduction.
Feature Selection: Apart from PCA, feature selection may also be a good way to reduce dimensions.
Trade-offs: Some anticipated trade-offs are summarized in Table 3.

Table 3: Trade-off Analysis of Dimension Reduction.

Trade-offs	Fewer Features	More Features
Interpretability	High	Low
Overfitting	Less likely	Likely
Prediction Accuracy	Low	High

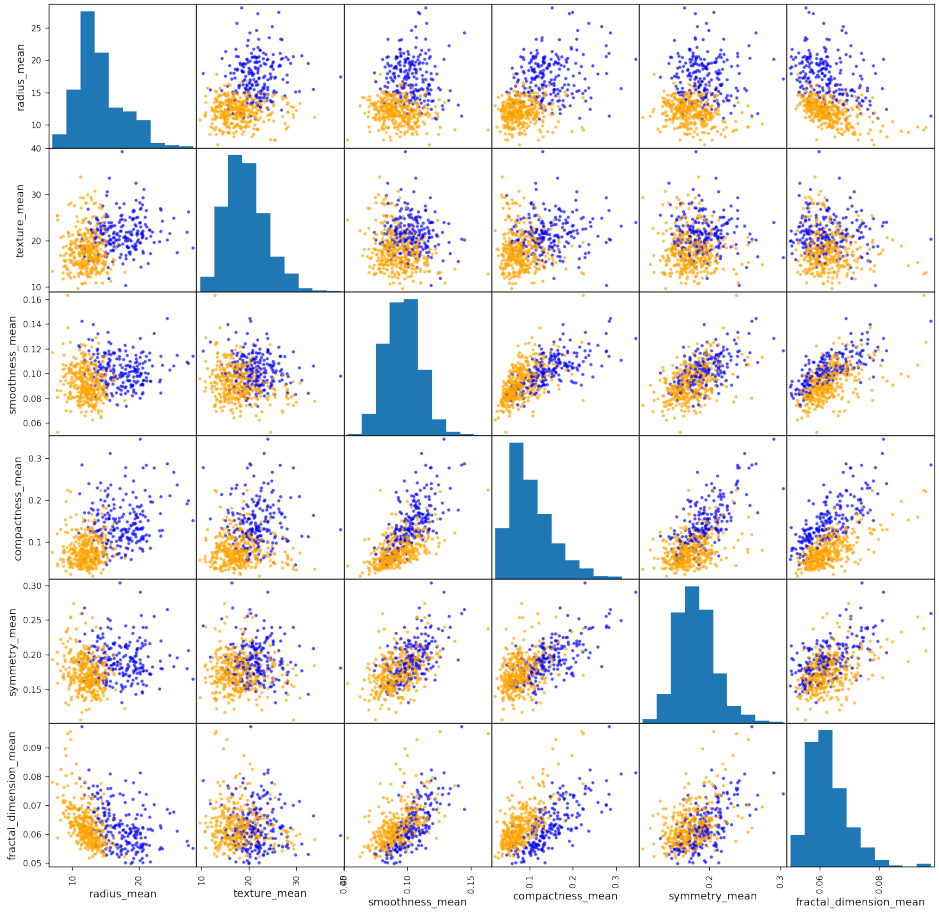

Figure 2: Pairwise scatter plot of six features.

3.3 Classifying Algorithms

Defined as a classification problem with two classes, the dataset can be tested by the following algorithms:
1) **Naive Bayes** [18]
2) **Logistic Regression**
Logistic regression [19] aims to model the posterior probabilities of all classes via linear functions of all features.
3) **Support Vector Machine**
A support vector classifier [20] generates a hyperplane for separating two classes.
4) **K-Nearest Neighbors**
A K-Nearest Neighbors classifier [21] adopts information around each query point for classification.
5) **Decision Tree Classifier** [22]
6) **Random Forest Classifier** [23]

3.4 Evaluation

As the number of samples is limited in the given dataset, cross-validation [24] will be necessary to evaluate performance of the predictor. When evaluating a predictor, accuracy is certainly of high priority. Apart from accuracy, computational complexity, model interpretability and potential of overfitting should also be taken into consideration.

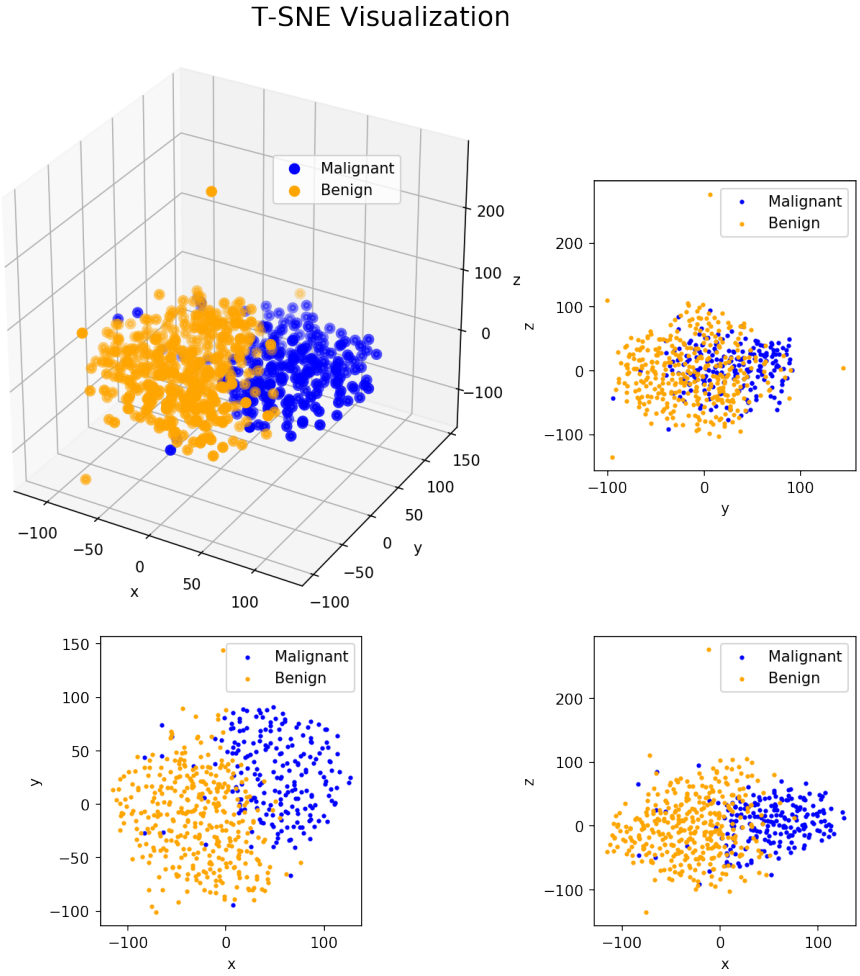

Figure 3: T-SNE visualization.

References

[1] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.

[2] M. Ibrahim and K. Chakrabarty, "Cyber-physical digital-microfluidic biochips: Bridging the gap between microfluidics and microbiology," *Proceedings of the IEEE*, 2017.

[3] M. Ibrahim, K. Chakrabarty, and U. Schlichtmann, "Synthesis of a cyberphysical hybrid microfluidic platform for single-cell analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[4] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[6] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, 2003.

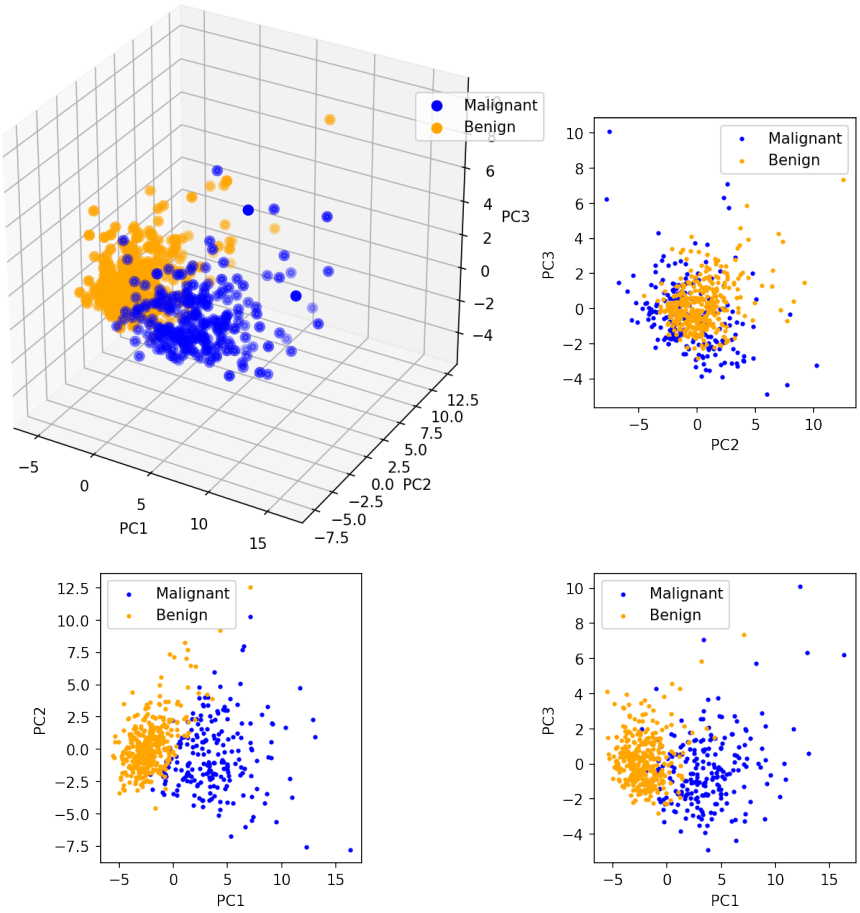

Figure 4: First three principle components visualization.

[7] S. Pineda, F. X. Real, M. Kogevinas, A. Carrato, S. J. Chanock, N. Malats, and K. Van Steen, "Integration analysis of three omics data using penalized regression methods: an application to bladder cancer," *PLoS genetics*, vol. 11, no. 12, p. e1005689, 2015.

[8] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*. ACM, 2018, pp. 5–9.

[9] W. 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*, vol. 87, no. 23, pp. 9193–9196, 1990.

[10] W. H. Wolberg, O. L. Mangasarian, and R. Setiono, "Pattern recognition via linear programming: Theory and application to medical diagnosis," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1989.

[11] K. P. Bennett, "Decision tree construction via linear programming," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1992.

[12] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, vol. 1, no. 1, pp. 23–34, 1992.

[13] K. Liu, G. Kang, N. Zhang, and B. Hou, "Breast cancer classification based on fully-connected layer first convolutional neural networks," *IEEE Access*, vol. 6, pp. 23 722–23 732, 2018.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Principal component analysis," 2017.

[16] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[17] "Breast Cancer Wisconsin (Diagnostic) Data Set: Predict whether the cancer is benign or malignant," https://www.kaggle.com/uciml/breast-cancer-wisconsin-data, 2019.

[18] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[19] H. Trevor, T. Robert, and F. JH, "The elements of statistical learning: data mining, inference, and prediction," pp. 119–127, 2009.

[20] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 417–438, 2009.

[21] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 463–471, 2009.

[22] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 305–313, 2009.

[23] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 587–601, 2009.

[24] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 241–247, 2009.