* This is a poster. More details please refer to **Full Report**

Machine Learning for Breast Cancer Wisconsin(Diagnostic)

Yuqin Shen

Abstract

The application of machine learning increases visual interpretation correctness and reduces manual work on early detection of breast cancers. The project applied six typical machine learning algorithms on the data set of Breast Cancer Wisconsin (Diagnostic) to predict whether the cancer is benign or malignant. The dataset consists of **569** instances (357 benign – 212 malignant), and each instance has **32** attributes. The project did data preprocessing to clean noises or outliers, then used PCA and t-SNE to visualize clustering performance before and after data normalization. Methods like KNN, Logistic Regression, Bayes Classification, SVM, Cross Validation and Feature Selection are used to train models. Model performance is measured on accuracy, AUC, precision and recall.

Objectives

- Learn the standard process to apply machine learning on dataset
- Explore data preprocessing methods and dive into their advantages
- Implement different machine learning models on data set and get to know each model characteristics by comparing their decision surfaces
- Master the ways to evaluate model performance, especially for medical data set

Data Preprocessing

Clear Outlier

Outlier is a value that lies in a data series on its extremes, which is either very small or large and thus can affect the overall observation made from the data series. Our data set is noise-free, thus there is no need to clean outliers as every data is a legitimate value. ROC curve of original data set has better performance.

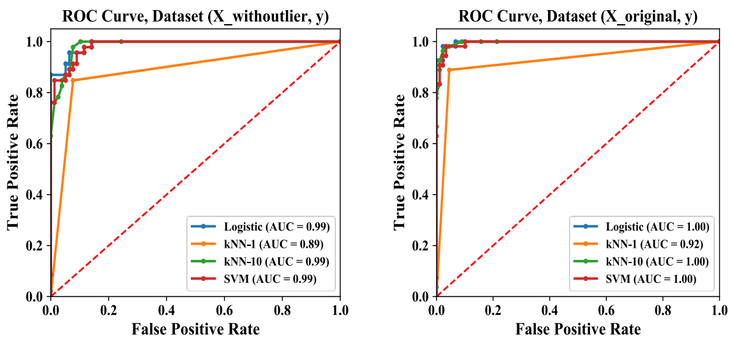

Figure 1: ROC Curve of data set with outliers vs original

Data Normalization

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

Data Preprocessing (continued)

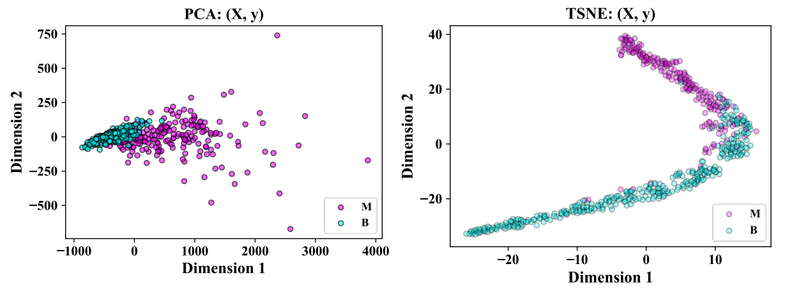

Figure 2: PCA & t-SNE scatter Plot before normalization

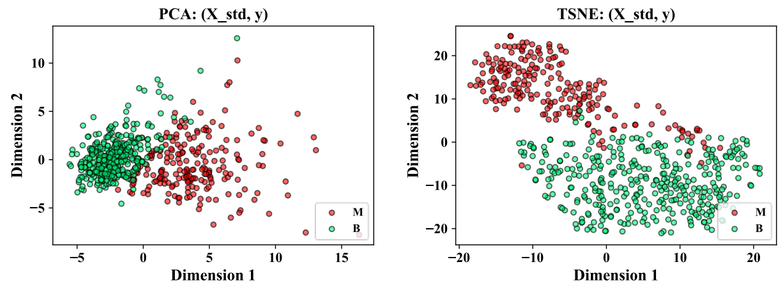

Figure 3: PCA & t-SNE scatter Plot after normalization

Model Training & Classification

Decision Statistic Surface

Decision boundaries or decision surfaces are the surfaces that separated the decision regions. It is a good way of classification visualization to explore the data and compare different classifiers. We explored KNN, Bayers, Logistic Regression and SVM characteristics by visualized their decision statistic surface. The color bar indicates the prediction of class 1 (Malignant) or class 0 (benign).

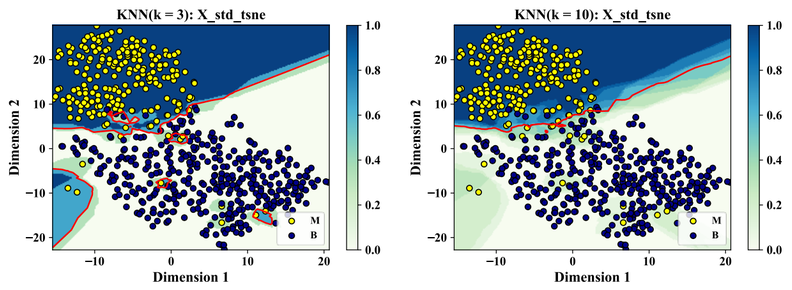

Figure 4: Decision statistic surfaces for KNN (k = 1) vs KNN (k = 10)

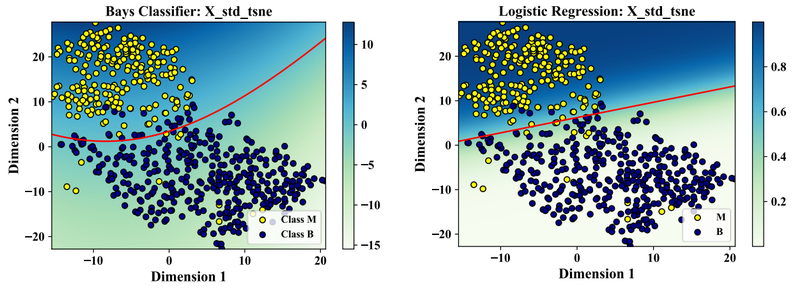

Figure 5: Decision statistic surfaces for Bayes Classifier vs Logistic

Model Training & Classification

Cross Validation

Cross-Validation is the most common technique that addresses some of the downsides of the training/test splits as it splits the data many times and averages the error over the splits.

The most common form of cross-validation is known as **k-fold** validation.

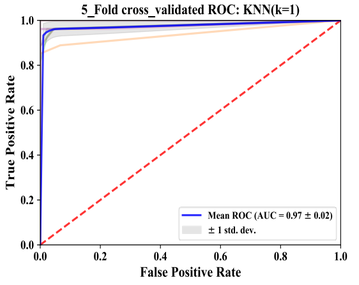

Figure 6: ROC Curve of 5-Fold on KNN (k=1)

Feature Selection

Feature selection is the process of removing redundant or irrelevant features from the original data set. This reduces the dimensionality of the data and enables learning algorithms to operate faster and more effectively.

The heatmap of dataset or feature selection algorithms could be used to select features for model training.

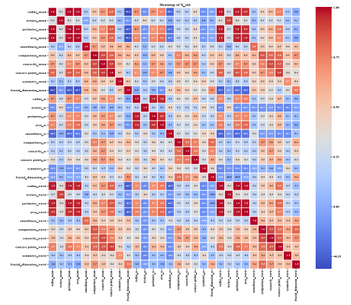

Figure 7: Heatmap of Full Features (full image)

Conclusion

Overall the performances of the models are very high because the data is typical used for machine learning, is neat, mid-size and binary classes are relatively balanced. Data preprocessing and data normalization, feature selection tends to refine classifiers. Logistic Regression, SVM, KNN-10 did around 5% better performance than KNN-1 due to overfitting problem.

References

1. Borges, L.R., 1989. Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. Group, 1(369).
2. Bhattacherjee, Aindrila, et al. Classification approach for breast cancer detection using back propagation neural network: a study. Biomedical image analysis and mining techniques for improved health outcomes. IGI Global, 2016. 210-221.

* More please refer to full report

Acknowledgements: This work was a course research project on 681 Pattern Classification and Recognition, taught by Ph.D. Stacy L. Tantum. Data set used is public in Kaggle.

Contact: ys238@duke.edu