Applying Different Machine Learning Models to Predict Breast Cancer

Lydia Xu (leediaxu@gmail.com) , Vera Xu (veraxrl@stanford.edu)

Stanford University

Abstract

Breast cancer is by far the most commonly diagnosed cancer among women worldwide. In this project, we aim to apply and compare various machine learning models (Logistic regression, NB, SVM and Random Forest) for breast cancer diagnosis based on cytopathology data and analyze the reasons behind their performances. The outcome will be an application that utilizes the most suitable machine learning principle to streamline the diagnosis process with the most accurate results.

Data & Features

Data: obtained from UCI database and collected from Wisconsin hospital. Each row contains 30 different features and the diagnosis of breast cancer (0 for benign and 1 for malignant). The 30 features represent the mean, standard deviation and the worst of 10 different cytopathology measurements, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. We have 569 entries in total.

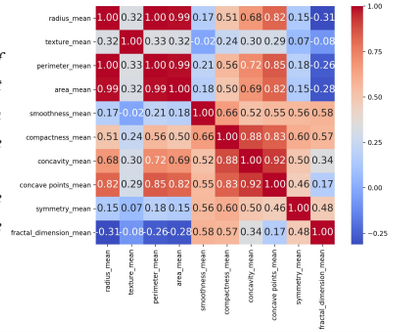
Feature Selection: 3 methods to avoid overfitting

1. **PCA (principle component analysis)**
Reduce dimensionality of the dataset into linearly uncorrelated principle components by multiplying a PCA transformation matrix.

2. RFE (recursive feature elimination):
Start with the initial set of features, and recursively remove one feature that is the least important until the desired number of features is reached.

3. Correlation Heat Map
Generate heatmap based on correlation of the "mean" features. It's noted that radius, perimeter and area are closely related and can thus should be grouped together. Same for concave_points and concavity.

Figure 1. Heatmap Analysis



Note[1]: Figure 1, Heatmap

Note[2]: Figure 2 shows a comparison of the number of features selected with respect to Logistic Regression, SVM and Random Forest. It shows that 5 features render the best accuracy.

Note[3]: PCA doesn't seem to improve performance of our models, due to the limited correlation among features.

Models

Four classification models were applied to the feature set:

Logistic Regression:

$$h_\theta(x) = \frac{1}{1 + exp(-\theta^T x)}$$

Support Vector Machine:

hinge loss function

$$\varphi_{hinge}(z) = [1 - z]_+ = max\{1 - z, 0\}$$

Naive Bayes:

$$\phi_{j|y=k} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = k\}}{\sum_{i=1}^{m} 1\{y^{(i)} = k\}} \quad where\ k = 0, 1$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

Random Forest:

Random Forest builds different decision trees based on various sub-examples of the dataset, and averages the decision to reduce variance and overfitting. Gini impurity is used to evaluate quality of a split:

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

Results & Discussion

The data is divided into 70% training (398) and 30% testing set (171). Table 1 shows the train and test errors for different models. Generally, models perform better on training set with the exception of Naive Bayes. All four models achieve good performance with test errors smaller than 0.1. Random Forest has the best performance for both training and test set, while SVM is a comparatively bad model. The good performance of Naive Bayes is unexpected because NB assumes features are independent, however, many features we used are strongly correlated.

Table 1. Train error and test error of difference models

	Train Error (%)	Test Error (%)
Logistic Reg	3.29	5.32
SVM	8.77	9.32
Random Forest	0.25	3.09
Naïve Bayes	5.53	3.51

Figure 2 shows that if we perform RFE feature selection, the test accuracy tends to improve when first reducing the feature set size, but it decreases when the feature set is too small, and the threshold is different for different models. The improvement is huge for SVM and random forest, suggesting that these models are overfitted when there are too many features.

Figure 3 visualizes commonly-used evaluation metrics (accuracy, recall, specificity and F1 score) to evaluate performance of all four models w/ and w/o feature selection. F1 score, the harmonic mean of precision and recall, reflects the weighted average of the two important metrics. With feature selection, Logistic Regression and Random Forest models render no False Positive and reach 100% specificity

Figure 2. Different models with RFE feature selections

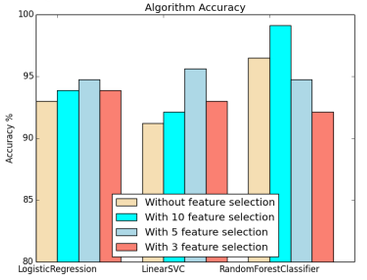
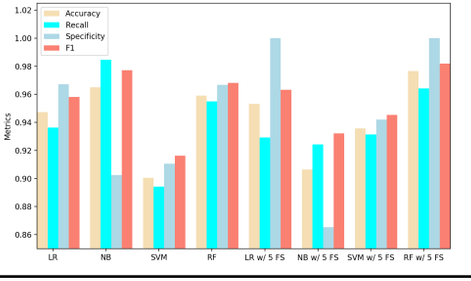


Figure 3. Evaluation metrics of different models w/ & w/o selection



Future Work

1. Further investigate the reason behind the unexpected high accuracy of Naive Bayes with more dataset and scientific background.
2. The current dataset was collected in the early 90s. If given more time, we will perform the same experiments on a more recent dataset as the new features collected maybe more accurate.

References

1. "Decision Tree Learning," Wikipedia, 03-Dec-2017. [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity. [Accessed: 11-Dec-2017].
2. "Random Forest - Fun and Easy Machine Learning," YouTube, 12-Jul-2017. [Online]. Available: https://www.youtube.com/watch?v=D_2LkhMJcfY. [Accessed: 11-Dec-2017].
3. "Breast Cancer Wisconsin (Original) Data Set," UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast Cancer Wisconsin %28Original%29. [Accessed: 11-Dec-2017].