Eric Yang

Breast Cancer Classification with SVM and Decision Tables

Goal and Dataset
The goal of this study is to provide diagnosis support for breast cancer malignancy using breast mass images. A successful model can aid oncologists in cancer interpretation, thereby reducing error and leading to more appropriate treatments. The dataset used in this study was created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian of the University of Wisconsin. 569 images of breast masses were obtained from fine needle aspirates. The diagnosis of each breast mass was also given, classified as either benign or malignant. For each image, cell nuclei were manually traced, and 10 features were extracted. These include radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1). In addition, the mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.
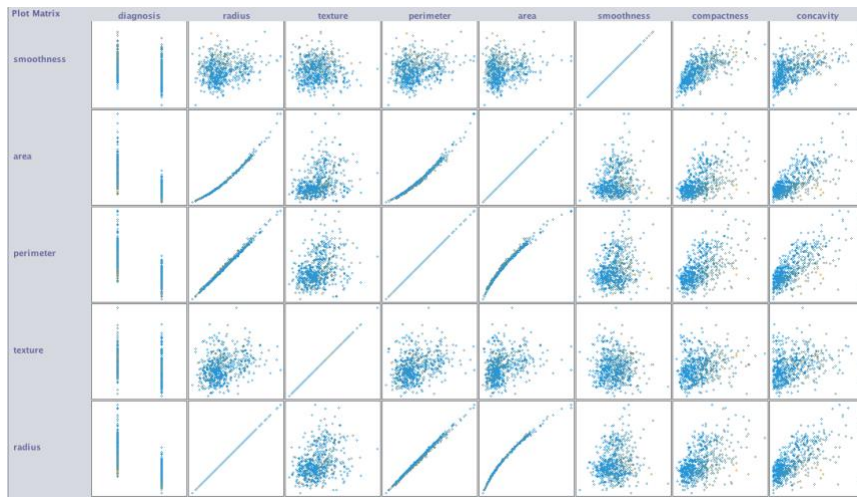
Figure 1. Subset of features plotted against one another.

Many features are linearly correlated via visual inspection (Figure 1). Specifically, the radius feature is strongly positively correlated with area, perimeter, worst area, worst radius, and worst perimeter. This makes sense mathematically as area and perimeter are geometrically related to radius. Below, I will classify cancer malignancy with all the features included and compare its performance with one where the five features correlated with radius are removed.

Classification with SVM and Decision Tables
The two classification methods used in this study were SVM and decision tables. Briefly, SVM finds a hyperplane that maximizes the margins to separate datapoints into two groups. With support of kernels, SVM is a computationally efficient binary classifier that is effective in high dimensional spaces. However, it is not the most interpretable model and it does not provide probability estimates on points belonging to certain groups.
The decision table classifies the dataset by matching feature values of testing data to a set of rules generated during training. To illustrate this, the rules used in one of the classifiers is shown

in a decision table (Figure 2). For example, an unseen observation with worst radius of 10, worst smoothness of 1 and worst concavity of 1 would be classified as benign based on the first rule, regardless of its values in the remaining 27 features. If an observation does not match any of the rules, then the majority of the classification in the decision table is used. To generate the decision table, the algorithm looks at a subset of features at a time and finds rules via induction that maximize prediction accuracy. Detailed descriptions of the specific heuristic search for rules within a subset can be found in "The power of decision tables" (Kohavi, 1995). The subset that yields the highest accuracy is kept. Some strengths of decision tables include its interpretability and fast classification. However, it does not do well with continuous features and can be sensitive to training data similar to decision trees. The assumption that decision trees make is the data type uniformity of each feature column. The train-test split for both SVM and decision table is 0.66-0.34.

```
Rules:
==================================================================
worst radius      worst smoothness  worst concavity  diagnosis
==================================================================
'(-inf-14.905]'   '(0.13615-inf)'   '(0.3663-inf)'     B
'(16.795-18.225]' '(0.13615-inf)'   '(0.3663-inf)'     M
'(14.905-16.795]' '(0.13615-inf)'   '(0.3663-inf)'     M
'(18.225-inf)'    '(0.13615-inf)'   '(0.3663-inf)'     M
'(-inf-14.905]'   '(-inf-0.13615]'  '(0.3663-inf)'     B
'(14.905-16.795]' '(-inf-0.13615]'  '(0.3663-inf)'     B
'(18.225-inf)'    '(-inf-0.13615]'  '(0.3663-inf)'     M
'(-inf-14.905]'   '(0.13615-inf)'   '(0.2164-0.3663]'  B
'(14.905-16.795]' '(0.13615-inf)'   '(0.2164-0.3663]'  M
'(18.225-inf)'    '(0.13615-inf)'   '(0.2164-0.3663]'  M
```
Figure 2. Subset of decision table rules that best classifies training data.

Results and Discussion

Decision table with all features:
```
 a   b   <-- classified as
70   4 |   a = M
 6 113 |   b = B
```

SVM with all features:
```
 a   b   <-- classified as
69   5 |   a = M
 0 119 |   b = B
```

Decision table with all features minus 5:
```
 a   b   <-- classified as
70   4 |   a = M
 6 113 |   b = B
```

SVM with all features minus 5:
```
 a   b   <-- classified as
68   6 |   a = M
 1 118 |   b = B
```

Figure 3. (Top) Confusion matrix of cancer malignancy classification with decision table (left) and SVM (right) using all features. (Bottom) Confusion matrix of cancer malignancy classification with decision table (left) and SVM (right) using all features minus area, perimeter, worst area, worst radius, and worst perimeter.

Given unseen test data, the decision tables classify tumor malignancy with accuracies of 94.8% for each of the conditions while SVM classifies better using all features compared to the condition with reduced features (97.4% vs 96.4%). It seems that the decision table performed worse as rule generation was more difficult for continuous features. In addition, the decision table only used three features for classification, potentially leaving out valuable information contained in other features. It is also interesting that addressing multiple collinearity in both approaches did not improve performance. This makes sense as rule-based algorithms like decision tables are indifferent to collinearity and SVM would generate the same hyperplanes with features containing essentially identical values. Regression-based approaches would be more prone to the multiple collinearity problem. In summary, SVM is the more reasonable approach here given continuous features and its comprehensive utilization of features.

References
1. Dua, D. and Graff, C. (2019) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. W.N. Street, W.H. Wolberg and O.L. Mangasarian. (1993) Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA.
3. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016) The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
4. Kohavi R. (1995) The power of decision tables. Proceedings of the European Conference on Machine Learning. Lecture Notes in Artificial Intelligence pp174-189