Lab 9: Classification

You are going to use classification methods to predict whether breast tissue samples are malignant or benign. You will set up an experiment, train a baseline model, perform feature selection, statistically compare candidate models to see if one approach outperforms another, and then do a final training and characterization of the logistic regression model using the appropriate features.

The data set is available here. Use the wdbc.data and wdbc.names files. Alternatively, you may import the data by following the "Import in Python" directions in the link above.

1 Load the data

1. Load the data into a DataFrame and standardize your features. You may do this 'by hand' or using StandardScaler from sklearn.preprocessing.
2. Characterize the dependent variable – plot the a bar chart of the diagnosis values.
3. Include a short (∼1 paragraph) description of the data set in a markdown cell in your notebook. What do the 30 features represent? What is the target variable?

2 Devise your experimental approach

1. Perform a (stratified!) train/test split using a 70/30 ratio.
2. In section three you will be using K-Fold Cross Validation when evaluating the models you create. Answer the following question in a markdown cell in your notebook:
   (a) If k=10, how many rows will each fold have (on average)?

3 Modeling

1. Use 10-fold cross validation (see sklearn.model_selection.cross_val_score) to create and assess three models:
   (a) A logistic regression baseline model that uses all standardized features.
   (b) A logistic regression model built using a "greedy" approach. Build a model for each variable individually. Sort the variables from most accurate to least accurate. Starting with the best performant single variable model, add each new variable one at a time to the model. If variable improves the model accuracy over the last model, keep that variable in the model. By "improves the accuracy" with k-fold CV, you should only include the feature if the average accuracy across all folds is increased. If the variable does not improve the accuracy, skip that variable. At the end, you should have a single model with multiple variables.

(c) A Random Forest classifier that uses all standardized features.

2. Run a statistical comparison between the accuracy of the three models. Identify if there is evidence to support one model being "best". Report the p-values and plot the accuracy and standard deviation of the two models. This should be done using the results of the k-fold training procedure.

3. Retrain your best performing model using all of the training data, then evaluate that model's performance on the test set (clearly display this accuracy in your notebook) and plot the ROC curve. The AUC should be reported in the figure legend.

4. Answer the following question in your writeup:

   (a) Was there statistically significant evidence that any of your models was "best"?

   (b) Did the features chosen by the greedy LR procedure align with the most important features used by the Random Forest? *Hint: If the random forest model was not your best model, retrain a single random forest on all of your training data and examine the feature_importances_ of your random forest.* Given the performance of the two models, how well would you say the greedy procedure works for feature selection?

4 Submission Instructions

- Convert your notebook to HTML and submit it to Canvas.