**Classification Methods**

model categorial outcome

classification rule = mathematical function that predicts the outcome of a new sample unit when the values of the covariates are known

split the data into a training and a test set

use the training dataset to build the classification rule

use the test dataset to evaluate how the classification rule labels new cases with known labels using ROC analysis

**Evaluating Classification Rule**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Misclassification Error Table | | Predicted Class | | Total True/False |
| **0** | **1** |
| True Class | **0** | true negative | false positive |  |
| **1** | false negative | true positive |  |
| Predicted True/False | |  |  |  |

accuracy = rate of correctly classified labels in the test set

false positive = predicted value is 1 and true value is 0

false negative = predicted value is 0 and true value is 1

sensitivity = true positive proportion

specificity = true negative proportion

positive predictive value (PPV) = proportion of true positives in predicted positives

negative predictive value (NPV) = proportion of true negatives in predicted negatives

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Misclassification Error Table | | Predicted Class | | Total True/False |
| **0** | **1** |
| True Class | **0** | 0  true negative | false positive |  |
| **1** | 0  false negative | true positive |  |
| Predicted True/False | |  |  |  |

test with maximum sensitivity will always predict 1 so specificity is 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Misclassification Error Table | | Predicted Class | | Total True/False |
| **0** | **1** |
| True Class | **0** | true negative | 0  false positive |  |
| **1** | false negative | 0  true positive |  |
| Predicted True/False | |  |  |  |

test with maximum specificity will always predict 0 so sensitivity is 0

cannot maximize sensitivity and specificity simultaneously in one test

**ROC Analysis**

ROC curve simultaneously displays the two types of errors for all possible thresholds

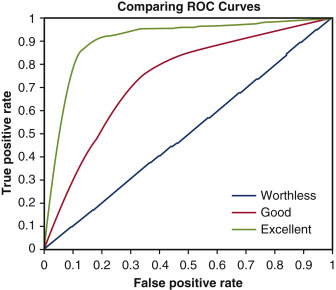
area under ROC curve (AUC) shows overall performance of a classifier summarized over all possible thresholds

ideal ROC curve will hug the top left corner and have large AUC

a classifier that performs no better than random chance will have AUC of 0.5

compares different classifiers because can consider all possible thresholds

use ROC analysis to choose threshold that has a good balance between sensitivity and specificity



**Methods to Generate Classification Rules**

Logistic Regression

estimate the probability of a binary outcome as a function of covariates

choose a threshold on the predicted value to decide how to classify

each threshold can be associated with a misclassification error table

Classification Trees

provides classification rule using a partition of the space of covariates

in each terminal node, the “default” classification rule is to assign the most common label, in the dataset

ROC analysis of the best classification tree identified by cross-validation

K-Nearest Neighbor

identify k closest observations in training set

assign x to the most common class of the k nearest neighbors

non-parametric approach

no assumptions are made about the shape

doesn’t identify which predictors are important

standardize observations because the method is distance-based

choose k by ROC analysis

Discriminant Analysis

model the distribution of the predictors separately in each of the response classes

use Bayes’ theorem to flip them around into estimates for

linear discriminant analysis assumes equal variance between groups

quadratic discriminant analysis assumes unequal variances between groups, so each class has its own covariance matrix