Classification

Classification is often used to describe modeling of a categorical outcome.

In binary classification the outcome is two possible values/classes and the goal is the predict the correct class using covariates.

Classification rule: A mathematical function to predict the outcome of a new sample unit when the values of the covariates are known.

Common Classification Problems

- Molecular diagnostic: Using values of gene products (such as biomarkers) develops a diagnostic tool for early detection of diseases.
- Classification of cancer type: Different sub-types of cancer are characterized by a combination of specific markers that are on/off. Gene based classification rules can be used to increase the specificity of the diagnosis.
- Classification of drug response
- Risk prediction

Types of Classifiers

- Regression-based:
  - Logistic regression, CART
- Example-based:
  - KNN
- Based on Bayes theorem:
  - Discriminant analysis
- Ensemble of classifiers

To evaluate a classifier, split the data into a training and test set. Use the training set to build the classification rule and the test set to evaluate how the classification rule labels new cases with known label.

Logistic Regression

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k\, X_k$$

b_i, i = 0,1... k can be estimated using Maximum Likelihood

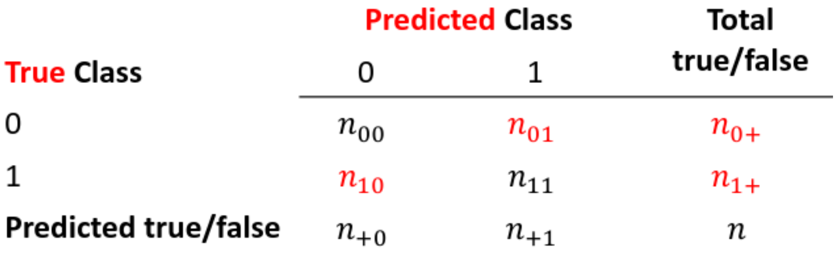
To estimate the probability of a binary outcome as a function of covariates with logistic regression:

$$p(y=1|x) = \frac{e^{b_0+b_1X_1+b_2X_2+\cdots+b_kX_k}}{1+e^{b_0+b_1X_1+b_2X_2+\cdots+b_kX_k}}$$

How to Pick Classification Rule

We can use the above formula for x in [0,1] to generate thresholds based on the predicted value to decide how to classify.

Accuracy is the rate of correctly classified labels in the test set:



	Predicted Class		Total
True Class	0	1	**true/false**
0	n_{00}	n_{01}	n_{0+}
1	n_{10}	n_{11}	n_{1+}
Predicted true/false	n_{+0}	n_{+1}	n

Misclassification error:
False positive: prediction 1 and true is 0; n01 / n0+
False negative: predict 0 and true is 1; Fn10 / n1+
Accuracy: (n00 + n11) / (n00 + n01 + n10 + n11)

Sensitivity (recall) the metric of true positive detection and shows whether the rule is sensitive to identify positive outcomes: 1 - n10 / n1+ = n11/n1+ = 1 - FNR

Specificity is the measure of true negative detection and shows whether the rule is specific in detection of negative outcomes: 1 - n01 / n0+ = n00 / n0+ = 1 - FPR

You CANNOT maximize sensitivity and specificity simultaneously. Maximum sensitivity test always says 1, maximum sensitivity always says 0.

Positive/Negative predicted values: The number of correct all predicted positive/negative values

ROC (Receiver Operating Characteristics) Analysis

The ROC curve is a population graphic for simultaneously displaying the two types of errors for all possible thresholds. They are useful for comparing different classifiers since they take into account all possible thresholds.

The overall performance of a classifier, summarized over all possible thresholds is given by the area under the ROC curve (AUC). An ideal ROC curve will hug the top-left corner, so the larger the AUC the better the classifier.

We expect a classifier that performs no better than chance to have an AUC of .5

Steps to Build and Evaluate Classification Rule:

1. Generate training and test set
2. Generate the classification rule using the training set;
3. Generate the predicted rules in the test set;
4. Use ROC analysis to decide the threshold that gives a good balance between sensitivity and specificity.
5. The next examples will show a variety of methods to generate classification rules (so step 2)

The previous chapter describes methods to create classification trees and K-nearest neighbor.

Discriminant Analysis

Logistic regression involves directly modeling P(Y = k | X = x) using logistic function. An alternative approach will model the distribution of the predictors X separately in each of the response classes, then use Bayes theorem to flip these around into estimates for P(Y = k | X =x)

Assuming the covariate x is normally distributed:

$$p(C=1\,|\,x)=\frac{p(x\,|\,C=1)p(C=1)}{p(x)}=\frac{p(x\,|\,C=1)p(C=1)}{\sum_j p(x\,|\,C=j)p(C=j)}$$

$$p(x\,|\,C=j)=\sqrt{\frac{1}{2\pi\sigma^2{}_j}}\exp\left(-\frac{1}{2\sigma^2{}_j}(x-\mu_j)^2\right);X\,|\,C=j\sim N(\mu_j;\sigma^2{}_j)$$

Binary outcome: Classify as 1 if p(C = 1 | x) > p(C = 0 | x)

$$\boxed{\begin{aligned}&-\frac{1}{2}\log(\sigma^2{}_1)-\frac{1}{2\sigma^2{}_1}(x-\mu_1)^2+\log(p(C=1))>\\&-\frac{1}{2}\log(\sigma^2{}_1)-\frac{1}{2\sigma^2{}_0}(x-\mu_0)^2+\log(p(C=0))\end{aligned}}$$

Classification rule

A special case when the variances of X in the groups are the same:

$$\boxed{x>\frac{(\mu_1+\mu_0)}{2}+\frac{\sigma^2(\log(p(C=0))-\log(p(C=1)))}{(\mu_1-\mu_0)}}$$

More than One Feature

To extend this approach to multiple covariates, one needs to decide:

1. How to model the correlation of the covariates within each group defined by the outcome
2. If the correlation of the covariates changes in different groups

LDA - assumes equal variance-covariance structure between groups
QDA - assumes group specific variance covariance matrices
LDA/QDA models all covariates as normal distributions

Summary

For each method:

1. Fit the classification model using training data
2. Evaluate the classification accuracy in test data using ROC analysis
3. Can also choose a “best” threshold to optimize sensitivity/specificity
4. Compare different classifiers by their AUC

The final classifier should be trained on all data to be used for future applications

There is no single "Best Classification Method". There is clear evidence that different methods work better in some data and worse in others. Thus, we often use the prediction probabilies of various classifiers to build an ensemble, using prediction probabilities from various rules. This is limited as there is no description of mechanism, useful only for prediction.

Revision #2
Created 17 November 2022 23:04:15 by Elkip
Updated 18 November 2022 00:31:48 by Elkip