Homework 2: Classification

MACS 30100: Perspectives on Computational Modeling
University of Chicago

Overview

For each of the following prompts, produce responses *with* code in-line. While you are encouraged to stage and draft your problem set solutions using any files, code, and data you'd like within the private repo for the assignment, *only the final, rendered PDF with responses and code in-line will be graded.*

A Theoretical Problem

1. (25 points) In classification problems, we minimize the generalization ("test") error rate by a simple classifier that assigns each observation to the most likely class given some set of input/predictor features,

$$\Pr(Y = j|X = x_0),$$

where x_0 is the test observation and each possible class is represented by $j \in \{1, \ldots, J\}$, which in the binary context is $\{0, 1\}$. The formula above is the **Bayes classifier**, which represents the conditional probability that $Y = j$, given the observed predictor value x_0. In the binary context, the Bayes classifier corresponds to predicting $j = 1$ if $\Pr(Y = 1|X = x_0) > 0.5$, else $j = 0$.

If the Bayes decision boundary is non-linear, then, would we expect LDA or QDA (both based on the Bayes classifier) to perform better on the training set? What about on the test set?

Answer this question with a simulation exercise. That is, follow the steps below and be sure to **numerically** and **visually** present error rates for both classifiers. Use this evidence to support your answer.

Repeat (simulate) the following process 1000 times (hint: I'd consider writing a function to make your simulation simpler to run, but of course this is up to you.):

a. Create a dataset with $n = 1000$, with two input features, $X_1, X_2 \sim \text{Uniform}(-1, +1)$. Also, create a response, Y, and let it be binary defined by $f(X) = X_1 + X_1^2 + X_2 + X_2^2$, where values 0 or greater are coded `TRUE` and values less than 0 or coded `FALSE`. Note: your Y is a function of the Bayes decision boundary ($X_1 + X_1^2 + X_2 + X_2^2$, as this non-linear model defines separation between the two classes), plus some error.
b. Randomly split your data into 80/20% training/test sets, respectively.
c. Train LDA and QDA classifiers.
d. Calculate each model's training and test error rate, based on your trained model from the previous step.
e. Present results (error rates for both sets of data and both classifiers) **visually** and **numerically**
f. Offer a *few sentence discussion* after results both answering the question and discussing differences in LDA and QDA approaches to classification in non-linear contexts like this.

An Applied Problem

For this applied problem, we will return to the 2016 ANES pilot study. Using these data, you will solve the classic political problem: predict party affiliation as a function of feelings toward a variety of things, concepts, and people as well as respondents' self-reported ideologies. The theoretical assumption here is that concepts indirectly related to one's party affiliation actually drive their party affiliation. Thus, we should be able to predict party affiliation as a function of non-partisan concepts. This is certainly debatable, and thus it's a perfect task for classification.

2. Answer the following questions, taking care to *discuss results and output at a technical and substantive level throughout* (e.g., what do ROC curves tell us and why? What are their relationships to AUC? What do the functional forms of different classifiers tell us about the quality of the solutions we get? What do the patterns *substantively* mean for our goal of predicting party affiliation? And so on.)

a. Load the data.
b. Preprocess the data to:

   - keep only four feeling thermometers for major 2016 politicians (2 extreme and 2 moderate from each party: `fttrump`, `ftobama`, `fthrc`, `ftrubio`), ideology on a five point scale (`ideo5`) party id (`pid3`)
   - recode party to a dichotomous feature where `1 = democrat` and `0 = all others`
   - drop NAs
   - make the response (democrat) a factor

c. Set the seed, and split the data into training (0.8) and testing (0.2) sets.
d. Fit the following classifiers using 10-fold cross-validation:

   - Logistic regression
   - Linear discriminant analysis
   - Quadratic discriminant analysis
   - K-nearest neighbors with $k = 1, 2, \ldots, 10$ (that is, 10 *separate* models varying k for each) and Euclidean distance metrics

e. Evaluate each model's performance using the test set. Select the best model based on the test set performance via:

   - Error rate
   - ROC curve
   - Area under the curve (AUC)

f. Once you select the best model (from your perspective)...

   - calculate your final estimate of the test error rate using the test set. In other words, take your best model and re-fit it using the entire training set (i.e. no cross-validation)
   - calculate performance metrics using the original test set
   - report (numerically *and* visually) and discuss results