Classification of Phoneme Data Using Quadratic Discriminant Analysis

Lucy L.

2024-01-06

This write-up is derived from the textbook "The Elements of Statistical Learning" from Chapter 5, exercise 5.5: "Write a program to classify the phoneme data using a quadratic discriminant analysis (Section 4.3). Since there are many correlated features, you should filter them using a smooth basis of natural cubic splines (Section 5.2.3). Decide beforehand on a series of five different choices for the number and position of the knots, and use tenfold cross-validation to make the final selection."

The dataset was extracted from the TIMIT database, a resource for speech recognition research. The dataset includes log-periodograms computed from 4509 speech frames, each 32 milliseconds long, representing five phonemes ("sh," "dcl," "iy," "aa," and "ao") from 50 male speakers.

```r
# Loading and subsetting data
phoneme_orig <- read.csv("phoneme.csv")
values <- c('aa', 'ao')
phoneme <- subset(phoneme_orig, g %in% values)
phon_subset <- phoneme[, -which(names(phoneme) == 'speaker')]
X <- phon_subset[, 1:256]
Y <- data.frame(phon_subset[,257])

# Choose 5 different degree of freedoms
# (Internal knots are uniformly distributed by default)
dfs <- c(5, 11, 50, 100, 200)
frequencies = 1:256
Y_fac = data.frame(phoneme$g)
Y_fac[, 1] <- factor(Y_fac[, 1])

for (df in dfs) {

  # Calculate H and X*
  H <- ns(frequencies, df = df)
  H <- as.data.frame(H)
  X_ast <- as.matrix(X) %*% as.matrix(H)

  # Perform QDA
  ctrl <- trainControl(method = "cv", number = 10)
  qda_model <- train(x = X_ast, y = Y_fac$phoneme.g,
                     method = "qda", trControl = ctrl)

  # Calculate mean error rate
  mean_err_rate <- mean(1 - qda_model$results$Accuracy)
  print(paste('When df is', df,
              ', mis-classification error rate is', mean_err_rate))
}
```

```
## [1] "When df is 5 , mis-classification error rate is 0.230652444953851"
## [1] "When df is 11 , mis-classification error rate is 0.193965048279614"
## [1] "When df is 50 , mis-classification error rate is 0.20737951321823"
## [1] "When df is 100 , mis-classification error rate is 0.233484858132696"
## [1] "When df is 200 , mis-classification error rate is 0.28302930108351"
```

In choosing the model with the best error rate, it appears that would be at $df = 11$.