DATA 590 | Homework 4 | Praveen Kumar Neelappa

1. **Short description of what your method does**

Penalized regression, minimize β {||y − Xβ||2 + P(β)}, especially in the setting where the number of features p exceeds the number of observations n.

P is a penalty function. Could be chosen to promote sparsity (e.g. the lasso,P(β) = ||β||1), smoothness and piecewise constancy.

Source: <http://www.stat.rice.edu/~jrojo/4th-Lehmann/slides/Witten.pdf>

1. **Code**

> Howells <- read.csv('http://math.mercyhurst.edu/~sousley/STAT\_139/data/Howells.csv', as.is = T);

> attach(Howells);

> HBNMF <- Howells[which(Pop == 'NORSE' | Pop == 'BERG'),];

> H4A <- na.omit(HBNMF[,c(5:61,63,67:80)])

> H4A$PopSex <- as.factor(H4A$PopSex)

> table(H4A$PopSex)

BERGF BERGM NORSEF NORSEM

53 56 55 55

>

> library(MASS)

> library(caret)

> Accuracies <- c(0.00)

> for (i in seq(500))

+ {

+ inTrain <- createDataPartition(y = H4A$PopSex, p = .70, list = FALSE)

+ training <- H4A[inTrain,]

+ testing <- H4A[-inTrain,]

+ pda4 <- train(PopSex ~ ., data = training, method = "pda",

+ preProcess = c("center", "scale"), tuneLength = 10,

+ trControl = trainControl(method = "cv"))

+ update(pda4, list(.lambda = 3))

+ pda4\_pred <- predict(pda4,newdata = testing)

+ Accuracies[i] <- confusionMatrix(knn4\_pred,testing$PopSex)$overall["Accuracy"]}

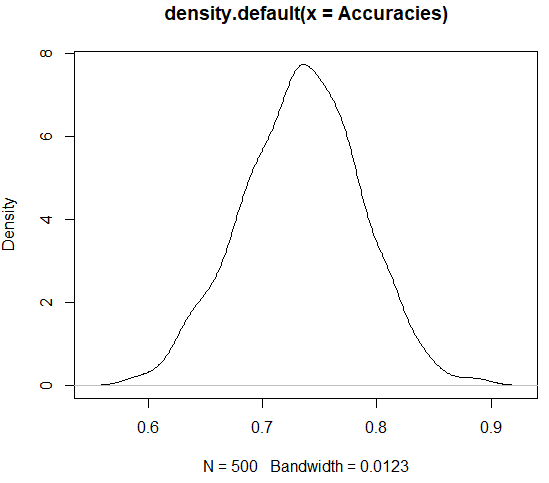
1. **Result**

> summary(Accuracies)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.5873 0.6984 0.7302 0.7341 0.7619 0.8889

> plot(density(Accuracies))



> pda4

Penalized Discriminant Analysis

156 samples

71 predictor

4 classes: 'BERGF', 'BERGM', 'NORSEF', 'NORSEM'

Pre-processing: centered (71), scaled (71)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 140, 140, 140, 140, 141, 140, ...

Resampling results across tuning parameters:

lambda Accuracy Kappa

0.0000000000 0.6812500 0.5751916

0.0001000000 0.6812500 0.5751916

0.0002371374 0.6812500 0.5751916

0.0005623413 0.6879167 0.5841202

0.0013335214 0.6879167 0.5841202

0.0031622777 0.6879167 0.5841202

0.0074989421 0.6879167 0.5841202

0.0177827941 0.6816667 0.5757869

0.0421696503 0.6820833 0.5761683

0.1000000000 0.6820833 0.5761683

Accuracy was used to select the optimal model using

the largest value.

The final value used for the model was lambda = 0.0005623413.

> confusionMatrix(pda4\_pred,testing$PopSex)

Confusion Matrix and Statistics

Reference

Prediction BERGF BERGM NORSEF NORSEM

BERGF 12 0 0 0

BERGM 3 15 0 2

NORSEF 0 0 15 2

NORSEM 0 1 1 12

Overall Statistics

Accuracy : 0.8571

95% CI : (0.7461, 0.9325)

No Information Rate : 0.254

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8093

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: BERGF Class: BERGM Class: NORSEF

Sensitivity 0.8000 0.9375 0.9375

Specificity 1.0000 0.8936 0.9574

Pos Pred Value 1.0000 0.7500 0.8824

Neg Pred Value 0.9412 0.9767 0.9783

Prevalence 0.2381 0.2540 0.2540

Detection Rate 0.1905 0.2381 0.2381

Detection Prevalence 0.1905 0.3175 0.2698

Balanced Accuracy 0.9000 0.9156 0.9475

Class: NORSEM

Sensitivity 0.7500

Specificity 0.9574

Pos Pred Value 0.8571

Neg Pred Value 0.9184

Prevalence 0.2540

Detection Rate 0.1905

Detection Prevalence 0.2222

Balanced Accuracy 0.8537

> confusionMatrix(pda4)

Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

Reference

Prediction BERGF BERGM NORSEF NORSEM

BERGF 17.3 5.1 2.6 0.6

BERGM 1.9 16.7 0.0 3.2

NORSEF 4.5 0.6 18.6 5.1

NORSEM 0.6 3.2 3.8 16.0

Accuracy (average) : 0.6859