Homework 4: Expected Prediction Error

Di Gravio, Chiara

Wed Mar 18 18:12:12 2020

Consider a 1-nearest neighbor classifier applied to a two-class classification problem, where the marginal probability associated with either class is one half, and where the distribution of a univariate predictor is standard normal, independent of class (i.e., not a very good predictor).

a) Show that the expected prediction error (EPE; HTF expression 7.3) is equal to 0.5.

If we assume the 0-1 loss function, the expected prediction error is given by:

$$
\begin{aligned}
EPE = E[\text{Err}_\tau] = E[L(Y,\hat{f}(X))|\tau] &= P(Y=\hat{f}(X))L(Y,\hat{f}(X)|Y=\hat{f}(X)) + P(Y\neq\hat{f}(X))L(Y,\hat{f}(X)|Y\neq\hat{f}(X)) \\
&= \frac{1}{2}\times 0 + \frac{1}{2}\times 1 = \frac{1}{2}
\end{aligned}
$$

We can show that the expected prediction error is 0.5 using a simulation study:

```
library(caret) # to do knn


set.seed(144)

# generate test data (always the same)
test <- data.frame(y = rbinom(1, n = 200, prob = 0.5), x = rnorm(200))
clas.err <- c()
B <- 1000

# compute over B training set
for(i in 1:B){
  # generate 2 classes with probability 1/2
  y <- rbinom(1, n = 200, prob = 0.5)
  # generate x
  x <- rnorm(200)
  dat <- data.frame(y = y, x = x)
  # fit KNN on training data and compute EPE
  knn_fit <- knnreg(x = data.frame(dat$x), y = dat$y, k = 1)
  knn_class <- predict(knn_fit, newdata = test$x)
  clas.err[i] <- mean(knn_class != test$y)
}
mean(clas.err)
```

```
## [1] 0.49759
```

b) Show that $E_z[\hat{\text{Err}}_{\text{boot}}]$ (expectation of HTF expression 7.54) is approximately equal to 0.184, where z represents the training sample of N class and predictor pairs. Thus, demonstrate that the bootstrap estimate of EPE is optimistic.

The idea is to fit the 1-nearest neighbour on a set of bootstrap samples, and then keep track of how well it predicts the original training set. The expected prediction error is:

$$\widehat{Err} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

Using a simulation study we have:

```r
set.seed(144)

# generate train data (always the same)
train <- data.frame(y = rbinom(1, n = 500, prob = 0.5), x = rnorm(500))
clas.err <- c()
B <- 1000

# compute over B training set
for(i in 1:B){
  # resample
  boot.dat <- train[sample(nrow(train), replace=T), ]
  # fit KNN on bootstrap sample and compute error on train sample
  knn_fit <- knnreg(x = data.frame(boot.dat$x), y = boot.dat$y, k = 1)
  knn_class <- predict(knn_fit, newdata = train$x)
  clas.err[i] <- mean(knn_class != train$y)
}
mean(clas.err)
```

```
## [1] 0.18402
```

c) Compute or approximate $E_z[\hat{\text{Err}}^{(1)}]$

The expected prediction error we want to compute is:

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

Using a simulation study we have:

```r
set.seed(144)

# generate train data (always the same)
# add id to keep track of observations
train <- data.frame(id = 1:500, y = rbinom(1, n = 500, prob = 0.5), x = rnorm(500))
# number of bootstrap (do only 100 since the code is slow)
B <- 100
# boostrap error
b.err <- c()

for (b in 1:B) {
  boot.dat <- train[sample(nrow(train), replace=T), ]
  knn_fit <- knnreg(x = data.frame(boot.dat$x), y = boot.dat$y, k = 1)
```

```r
  # for each i not in the data keep track of the error
  err <- c()
  for (i in 1:nrow(train)) {
    # we only want those replicates without i
    if (!(i %in% boot.dat$id)) {
      test <- train[i, ]
      knn_pred <- predict(knn_fit, test$x)
      err <- c(err, mean(knn_pred != test$y))
    }
    }
  b.err <- c(b.err, mean(err))
}
mean(err)
```

```
## [1] 0.4972376
```

d) Compute or approximate $E_z[\hat{\text{Err}}^{(0.632)}]$

We need to compute:

$$\widehat{Err}^{(0.632)} = 0.368 \times \overline{\text{err}} + 0.632 \times \widehat{\text{Err}}^{(1)}$$

where $\overline{\text{err}}$ is the training error. When using 1-nearest neighbour in the training sample, the training erro is 0. From the question above, $\widehat{\text{Err}}^{(1)}$ is approximately 0.497. Thus:

$$\widehat{Err}^{(0.632)} = 0.368 \times \overline{\text{err}} + 0.632 \times \widehat{\text{Err}}^{(1)} = 0.368 \times 0 + 0.632 \times 0.497 = 0.314$$