Project 3 - Part 2

Cameron Sims

6/9/2021

In this analysis, I will use my_penguins.csv data to predict mody_mass_g using covariates bill_length_mm, bill_depth_mm, and flipper_length_mm. I will use fold sizes of 2, 5, and 10 and run 30 iterations with each size. I will then make 3 boxplots of the data, a table displaying the average CV estimate and the standard deviation of the CV estimates for each fold size, and discuss the observed results. The figures and results will be saved in respectively-named folders within the “Output” subfolder.

```r
# Vector to hold CV estimates MSEs
mses <- c()

# Using k = c(2, 5, 10), perform random forest 30 times each
for (k in c(2, 5, 10)) {
  for (i in 1:30) {
    mses <- append(mses, my_rf_cv(k))
  }
}

# Save MSEs as 30 row, 3 column csv
write.csv(data.frame(k_1 = mses[1:30], k_5 = mses[31:60], k_10 = mses[61:90]),
                     "../Output/Results/simulation_results.csv")

# Separate MSEs by k for boxplot
dat <- data.frame(mse = mses, k = rep(c(2, 5, 10), each = 30))
```

```r
# Create a boxplot for each value of k
chart <- ggplot(data = dat, mapping = ggplot2::aes(x = as.factor(k), y = mse)) +
  geom_boxplot() +
  labs(title = "Distribution Of MSE For Different Numbers Of Folds") +
  xlab("Number of Folds") +
  ylab("CV-Estimated MSE")

# Save chart in "figures" sub-subfolder
ggsave("mse_boxplot.pdf", chart, path = "../Output/Figures")

# Calculate average CV estimate and standard deviation for k = 2
two_avg <- mean(subset(dat, k == 2)$mse)
two_sd <- sd(subset(dat, k == 2)$mse)

# Calculate average CV estimate and standard deviation for k == 5
five_avg <- mean(subset(dat, k == 5)$mse)
five_sd <- sd(subset(dat, k == 5)$mse)

# Calculate average CV estimate and standard deviation for k == 10
ten_avg <- mean(subset(dat, k == 10)$mse)
```

```
ten_sd <- sd(subset(dat, k == 10)$mse)

# Save average CV estimates and standard deviations in a table
summary_stats <- data.frame(Mean_CV = c(two_avg, five_avg, ten_avg),
           SD_CV = c(two_sd, five_sd, ten_sd))

# Save summary statistics in "results" sub-subfolder
saveRDS(summary_stats, file = "../Output/Results/summary_stats.rds")
```

From the boxplots, it appears the median CV-Estimated MSE decreases as the number of folds increases. 5 folds has the smallest interquartile range of the three fold amounts and exists almost entirely within the range of the 10-fold boxplot. The 2-fold boxplot has the largest range and is centered higher than the other boxplots. Numerically, the table shows 10 folds has the lowest average CV estimate whereas five folds has the lowest CV standard deviation. Two folds has both the highest mean CV estimate and the highest CV standard deviation.

Generally, test error (CV MSE) will decrease as the number of folds increases. This is because more data is being used to calculate the predicted value(s). However, after a certain point, this value can decrease as the model becomes less reliable (training error increases). As such, it is possible the best number of folds is close to five, or between five and ten. While ten folds might have the lowest mean CV-Estimated MSE, it has higher standard deviation than five-folds and thus is not conclusively better.