Answer 3

Alvee Chowdhury

28/01/2021

Set up:

```r
library(FNN)
library(MASS)
library(flextable)

train_data <- read.csv("h1q3-train-data.csv")
test_data <- read.csv("h1q3-test-data.csv")

X_train <- train_data["x"]
X_test <- test_data["x"]

y_train <- train_data["y"]
y_test <- test_data["y"]

testing_data <- data.frame(test = seq(min(X_train), max(X_train),
                                      by = 0.01))

k = seq(5, 50, by = 5)
```

Fitting 10 nearest neighbors models:

```r
for (i in c(1:10)){
  Model_name <- paste("Model_", i, sep = "")
  assign(Model_name, knn.reg(train = X_train, test = testing_data, y = y_train, k = k[i]))
}
```

Hence we have 10 nearest neighbors models. Now we need to calculate test RMSE and train RMSE for each:

```r
rmse = function(actual, predicted) {
  sqrt(mean((actual - predicted) ^ 2))
}

# define helper function for getting knn.reg predictions
# note: this function is highly specific to this situation and dataset
make_knn_pred = function(k = 1, training, predicting) {
  pred = FNN::knn.reg(train = training["x"],
                      test = predicting["x"],
                      y = training$y, k = k)$pred
  act = predicting$y
  rmse(predicted = pred, actual = act)
}

# get requested train RMSEs

knn_trn_rmse = sapply(k, make_knn_pred,
                      training = train_data,
                      predicting = train_data)

# get requested test RMSEs
knn_tst_rmse = sapply(k, make_knn_pred,
                      training = train_data,
                      predicting = test_data)

# determine "best" k
best_k = k[which.min(knn_tst_rmse)]

# find overfitting, underfitting, and "best"" k
fit_status = ifelse(k < best_k, "Over", ifelse(k == best_k, "Best", "Under"))
```

Now to tabulate our findings and conclusion:

```r
# summarize results
knn_results = data.frame(
  k,
  round(knn_trn_rmse, 2),
  round(knn_tst_rmse, 2),
  fit_status
)


colnames(knn_results) = c("k", "Train RMSE", "Test RMSE", "Fit?")

t<-flextable(knn_results)


m<- colformat_num(t,j=1,digits = 0)
j<- colformat_num(m,j=c(2,3),digits = 2)



autofit(align(j,align = "center", part = "all"))
```

k	Train RMSE	Test RMSE	Fit?
5	1.65	2.16	Over
10	1.70	2.08	Over
15	1.79	2.05	Best
20	1.93	2.06	Under
25	2.02	2.14	Under
30	2.28	2.36	Under
35	2.60	2.67	Under

k	Train RMSE	Test RMSE	Fit?
40	2.96	2.99	Under
45	3.27	3.29	Under
50	3.58	3.57	Under