K-Fold CV Simulation

William Chiu

3/14/2022

Intro

A study to examine the trade-offs associated with increasing the K in K-Fold cross-validation. Sometimes referred to as V-fold cross-validation.

Values of K: 3-fold, 5-fold, 10-fold, and leave-one-out.

Simulating one data set

```
nobs <- 200
K_param <- 5
seedVal <- 123

set.seed(seedVal)

x1 <- rnorm(nobs, 1, 2)
x2 <- rnorm(nobs, 2, 4)

y <- 5 + 5*x1 + 2*x2 - 0.8*x2^2 + rnorm(nobs,0,6)

full_data <- data.frame(y=y, x1=x1, x2=x2)

rm(y, x1, x2)

trainID <- sample(1:nobs, size=150)

trainData <- full_data[trainID,]
testData <- full_data[-trainID,]

lm_mod <- glm(y ~ x1 + x2, data=trainData, family=gaussian)

test_preds <- predict(lm_mod, newdata=testData)

test_MSE <- mse(testData$y, test_preds)

cv_MSE <- cv.glm(trainData, lm_mod, K=K_param)$delta[1]

outdf <- data.frame(Iter=seedVal, K=K_param, test_MSE=test_MSE, cv_MSE=cv_MSE)

knitr::kable(outdf)
```

Iter	K	test_MSE	cv_MSE
123	5	431.3462	325.8719

Simulate many data sets

```
doOne <- function(K_param, seedVal) {
  nobs <- 200

  set.seed(seedVal)

  x1 <- rnorm(nobs, 1, 2)
  x2 <- rnorm(nobs, 2, 4)

  y <- 5 + 5*x1 + 2*x2 - 0.8*x2^2 + rnorm(nobs,0,6)

  full_data <- data.frame(y=y, x1=x1, x2=x2)

  rm(y, x1, x2)

  trainID <- sample(1:nobs, size=150)

  trainData <- full_data[trainID,]
  testData <- full_data[-trainID,]

  lm_mod <- glm(y ~ x1 + x2, data=trainData, family=gaussian)

  test_preds <- predict(lm_mod, newdata=testData)

  test_MSE <- mse(testData$y, test_preds)

  cv_MSE <- cv.glm(trainData, lm_mod, K=K_param)$delta[1]

  outdf <- data.frame(Iter=seedVal, K=K_param, test_MSE=test_MSE, cv_MSE=cv_MSE)

  outdf
}

grid <- expand.grid(K_param=K_try, seedVal=seq(p_sims))

allSims_df <- future_map2_dfr(.x=grid$K_param, .y=grid$seedVal, .f=doOne,
                              .options=furrr_options(seed=NULL))

future:::ClusterRegistry("stop")

knitr::kable(head(allSims_df))
```

Iter	K	test_MSE	cv_MSE
1	3	321.0966	419.2220
1	5	321.0966	384.1935
1	10	321.0966	401.9093

Iter	K	test_MSE	cv_MSE
1	150	321.0966	404.8134
2	3	488.0906	432.8900
2	5	488.0906	380.3881

```
allSim_df_wide <- pivot_wider(allSims_df, id_cols=c(Iter, test_MSE), names_from=K,
                              values_from=cv_MSE, names_prefix="cv_")

knitr::kable(head(allSim_df_wide))
```

Iter	test_MSE	cv_3	cv_5	cv_10	cv_150
1	321.0966	419.2220	384.1935	401.9093	404.8134
2	488.0906	432.8900	380.3881	376.2437	376.1140
3	734.5094	354.4563	385.3012	365.7957	360.1479
4	329.9043	336.6535	351.5706	349.8909	329.2857
5	172.4584	496.1130	566.5655	535.6773	519.1317
6	410.0495	439.2408	379.5364	388.4821	402.4820

```
allSim_df_long <- pivot_longer(allSim_df_wide, -Iter, names_to='Method', values_to = 'MSE')

knitr::kable(head(allSim_df_long))
```

Iter	Method	MSE
1	test_MSE	321.0966
1	cv_3	419.2220
1	cv_5	384.1935
1	cv_10	401.9093
1	cv_150	404.8134
2	test_MSE	488.0906

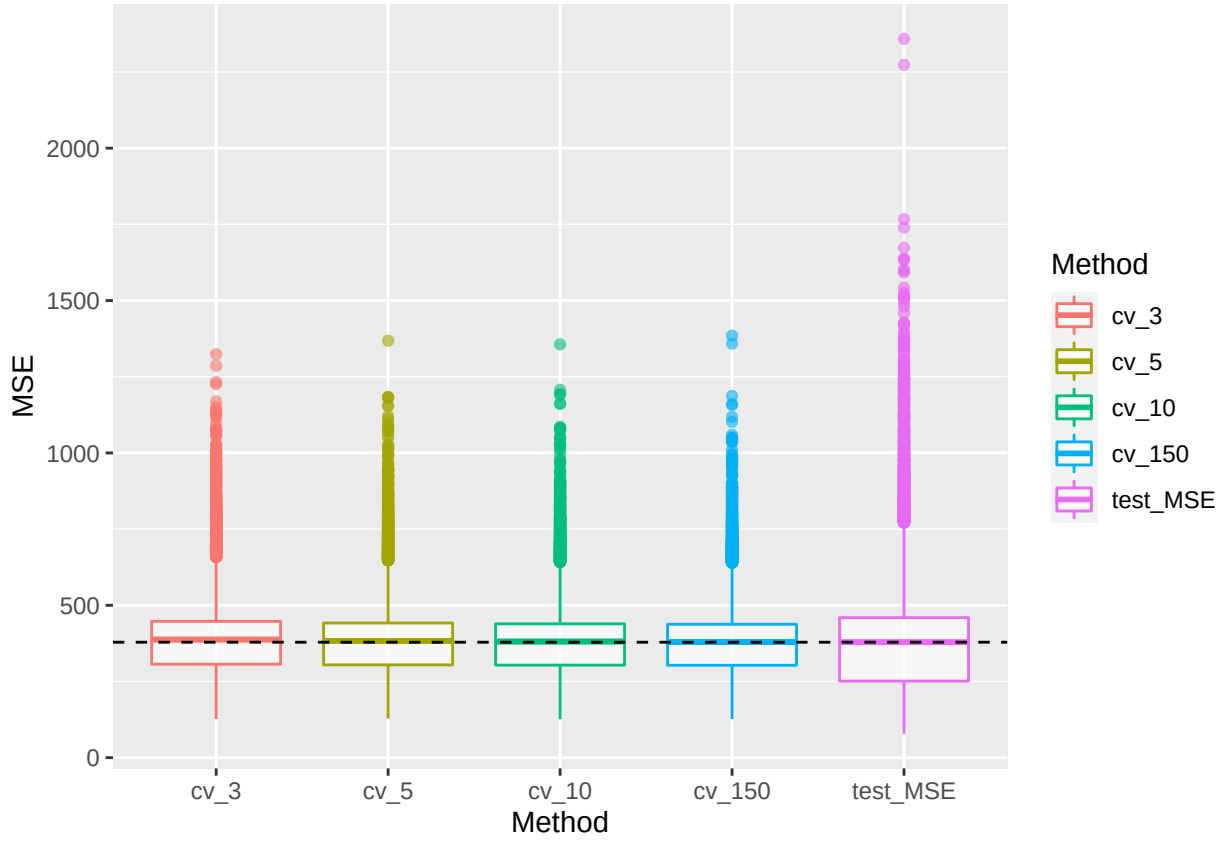
```
allSim_df_long$Method <- factor(allSim_df_long$Method, levels=c("cv_3", "cv_5",
                                                "cv_10", "cv_150", "test_MSE"))
```

```
unbiased_MSE <- allSim_df_long %>%
  filter(Method=='test_MSE') %>%
  summarize(meanTestMSE = mean(MSE), .groups = 'drop') %>%
  as.numeric()

ggplot(allSim_df_long, aes(x=Method, y=MSE, color=Method)) +
  geom_boxplot(fatten = NULL, alpha=0.6) +
  stat_summary(fun = mean, geom = "errorbar", aes(ymax = ..y.., ymin = ..y..),
               width = 0.75, size = 1, linetype = "solid") +
  geom_hline(yintercept=unbiased_MSE, linetype="dashed", color = "black")
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
## Removed 1 rows containing missing values (geom_segment).
## Removed 1 rows containing missing values (geom_segment).
```

```
## Removed 1 rows containing missing values (geom_segment).
## Removed 1 rows containing missing values (geom_segment).
```




```
out <- allSim_df_long %>%
  group_by(Method) %>%
  summarize(AvgMSE=mean(MSE), SD=sd(MSE), .groups = 'drop')

knitr::kable(out)
```

Method	AvgMSE	SD
cv_3	387.4832	115.6571
cv_5	383.2223	111.7662
cv_10	381.1764	110.1112
cv_150	379.5376	108.8682
test_MSE	378.8028	181.5732