Read Data

11/23/2021

```
library(foreign)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(e1071)
sesame <- read.dta("sesame.dta")
sesame <- sesame %>%
  mutate(site=factor(site)) %>%
  mutate(bodyDiff = postbody - prebody,
         letDiff = postlet - prelet,
         formDiff = postform - preform,
         numbDiff = postnumb - prenumb,
         relatDiff = postrelat - prerelat,
         clasfDiff = postclasf - preclasf)
```

Q.2 Classification Question: Can we use the pre-test scores and other demographic variables to predict which region the children came from?

```
set.seed(3241)

n <- nrow(sesame)
train.index <- sample(1:n, size = floor(0.7*n), replace=FALSE)
train.data <- sesame[train.index,]
test.data <- sesame[-train.index,]
```

```
# Response: site (categorical)
set.seed(315)
costs <- c(0.001, 0.01, 0.1, 1, 5, 10, 100)
gammas <- c(0.1, 0.5, 1, 2, 3, 4)

linear.tune <- tune(svm, site~sex+age+prebody+prelet+preform+prenumb+prerelat+preclasf,
                    data=train.data, kernel="linear",
                    ranges=list(cost=costs))
radial.tune <- tune(svm, site~sex+age+prebody+prelet+preform+prenumb+prerelat+preclasf,
                    data=train.data, kernel="radial",
                    ranges=list(cost=costs,
                                gamma=gammas))
```

```
linear.conMatrix <- table(true=test.data[, "site"],
                          pred=predict(linear.tune$best.model, newdata=test.data))

radial.conMatrix <- table(true=test.data[, "site"],
                          pred=predict(radial.tune$best.model, newdata=test.data))

linear.conMatrix
```

```
##     pred
## true  1  2  3  4  5
##    1  0 10 10  0  0
##    2  0  8  5  0  0
##    3  0  3 13  0  0
##    4  0  7 11  0  0
##    5  0  1  4  0  0
```

```
radial.conMatrix
```

```
##     pred
## true  1  2  3  4  5
##    1  6  7  7  0  0
##    2  1  8  4  0  0
##    3  0  3 13  0  0
##    4  1  7 10  0  0
##    5  0  1  4  0  0
```

Radial kernel improves prediction on class 1.

Questions for OH:

Both linear and radial kernels never output predictions for 4 & 5?

polynomial kernel? Which variables to give polynomial terms

use PCA to perform feature selection?

feature selections for SVM in general?