Week4, Assignment 1

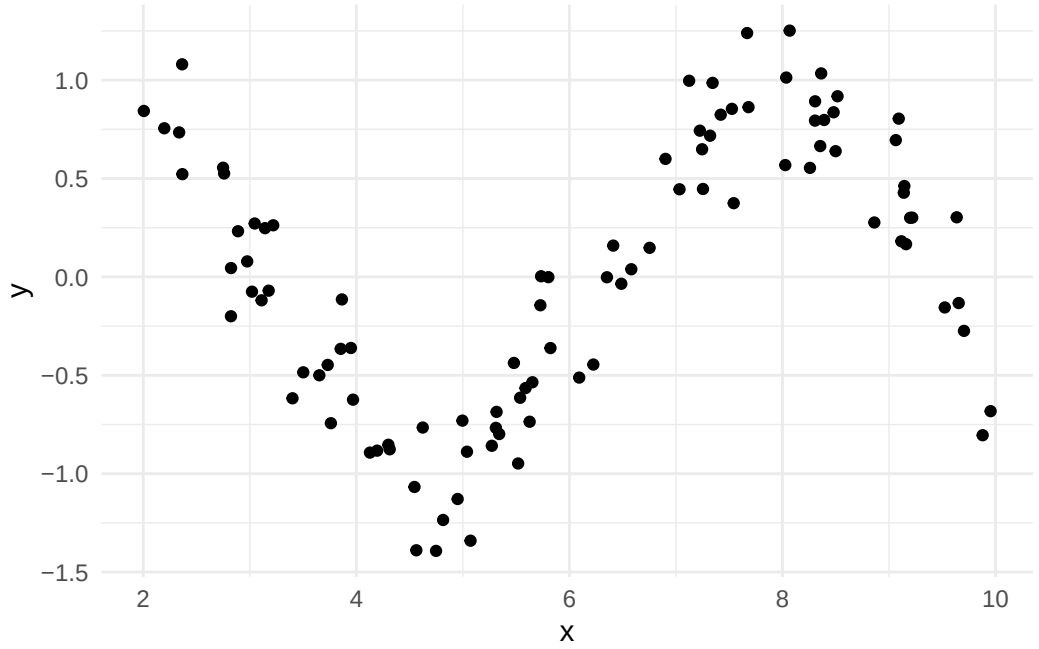
Student Name

```
library(caret)
# ... add additional libraries here ...
library(tidyverse)
# ...

# ggplot
theme_set(theme_minimal())

seed <- 123
```

Problem 4.1 (20 points)

Simulate a single predictor and a nonlinear relationship, such as a sin wave shown in Fig. 7.7 of textbook, and investigate the relationship between the cost, , and kernel parameters for a support vector machine model

```
set.seed(seed)
sinData <- tibble(
    x = runif(100, min = 2, max = 10),
    y = sin(x) + rnorm(length(x)) * .25
)
sinData |>
    ggplot(aes(x,y)) +
    geom_point()
```




```
## Create a grid of x values to use for prediction
dataGrid <- tibble(x = seq(2, 10, length = 100))
```

4.1.a (12 points)

Fit different models using a radial basis function and different values of the cost (the C parameter) and assigning sigma a constant value of 1. To study the impact of each parameter, keep all other parameters constant and vary only one parameter at a time. Plot the fitted curves and discuss what happens when cost and epsilon parameters are varied in terms of overfitting/underfitting.

```
# library(kernlab)

SVM_explore <- function(cost, epsilon,
                        data_grid = dataGrid, sin_data = sinData) {
#... code ..
    tibble(
        # x = ... ,
        # y = ... ,
        # label = ...
    )
}
```

```
cost_varies <- 10^seq(-2,1, length = 4)
eps01_fixed <- rep(0.1, 4)
cost_curves <- map2_dfr(cost_varies,
                        eps01_fixed,
                        ~ SVM_explore(.x, .y))
#  ggplot(cost_curves ...)

eps_varies <- c(0.01, 0.1, 0.5, 1.0)
cost_fixed <- rep(0.1, 4)
eps_curves <- map2_dfr(cost_fixed,
                        eps_varies,
                        ~ SVM_explore(.x, .y))
#  ggplot(eps_curves ...)
```

...Observations on varying Cost...

...Observations on varying Epsilon...

4.1.b (8 points)

The parameter can be adjusted using the kpar argument, such as kpar = list(sigma = 1). Try different values of to understand how this parameter changes the model fit. How do the cost, , and values affect the model?

Hint: you'll need to modify the `SVM_explore` function above

```
SVM_explore3 <- function(cost = 0.1, epsilon = 0.1, sigma = 1,
                        data_grid = dataGrid, sin_data = sinData) {
#... code ..
    tibble(
        # x = ... ,
        # y = ... ,
        # label = ...
    )
}

sigma_varies <- 10^seq(-2,1, length = 4)
sigma_curves <- map_dfr(sigma_varies,                    #note switch to map2... from map...
                        ~ SVM_explore3(sigma = .x))

#  ggplot(sigma_curves ...) ...
```

...Observations on varying Epsilon...

Problem 4.2 (30 points)

Returning to the chemical manufacturing process data (from Module 3 Assignment Problem 3.3):

```
library(AppliedPredictiveModeling)
data(ChemicalManufacturingProcess)
```

The ChemicalManufacturingProcess data frame contains 57 predictors (12 describing the input biological material and 45 describing the process predictors) and a yield column which is the percent yield for each run for the 176 manufacturing runs. **Split the data into a training (80%) and a test set (20%). A small percentage of cells in the predictor set contain missing values. Use an imputation function to fill in these missing values in both training and test data sets, also perform centering and scaling. Build and tune nonlinear regression models Neural network, MARS, SVM and KNN.**

Hint: Reuse code from Module 3 Problem 3.3.b

```
#...code for data prep (reuse from 3.3 (b)...
```

Model fits

```
#... code for non-linear models...
```

4.2.a (20 point)

Which nonlinear regression model gives the optimal resampling and test set performance (use RMSE metric to make this determination)?

```
#... summarize results ...#
```

...Answer...

4.2.b (5 point)

What are the ten most important predictors in the optimal nonlinear regression model? Do either the biological or process variables dominate the list?

```
#... code ...#
```

...Answer...

4.2.c (5 point)

How do the important predictors compare to the important predictors from the optimal linear model you built in module 3 assignment? (complete the table below)

Ranked importance of predictors by model

Predictor	SVM Rank (1-10)	Lasso Rank (1-5)
VarNN	1	*<Rank or empty>*
VarNN	2	*<Rank or empty>*
...	...	...
VarNN	10	*<Rank or empty>*
...	...	*<Rank>*

Hint: *The total number of rows in the table above should be 10 + any variables from Assignment 3 top 5 that are NOT in the top 10 for your non-linear model*

Problem 4.3 (30 points)

Recreate the simulated data as shown below:

```
library(mlbench)
set.seed(200)
simulated <- mlbench.friedman1(200, sd = 1)
simulated <- cbind(simulated$x, simulated$y)
simulated <- as.data.frame(simulated)
colnames(simulated)[ncol(simulated)] <- "y"
```

In this data set V6 to V10 are columns that are random noise and have no relation to the response.

4.3.a (10 points)

Fit a random forest model to all of the predictors, then estimate the variable importance scores

```
library(randomForest)
```

```
randomForest 4.7-1.1
```

```
Type rfNews() to see new features/changes/bug fixes.
```

```
Attaching package: 'randomForest'
```

```
The following object is masked from 'package:dplyr':

    combine
```

```
The following object is masked from 'package:ggplot2':

    margin
```

```
library(caret)
set.seed(seed)
model1 <- randomForest(y ~ ., data = simulated, importance = TRUE, ntree = 1000)
importance1 <- varImp(model1)

# ... show resulting Importance values ...
```

Did the random forest model significantly use the uninformative predictors (V6 – V10)? What do you think the impact of having uninformative predictors will be on random forest model performance?

...Answer...

4.3.b (10 points)

Now add an additional predictor that is highly correlated with one of the informative predictors, V1.

```
simulated$duplicate1 <- simulated$V1 + rnorm(200) * .1
cor(simulated$duplicate1, simulated$V1)
```

```
[1] 0.9340411
```

Fit another random forest model to these data. Did the importance score for V1 change?

```
# ... code ...
```

...Answer...

What happens if you add one more predictor to the data set that is also highly correlated with V1 (after this you should have 12 predictors, 10 original and two added correlated predictors) and fit another random forest model again to this data. Did the importance score for V1 change?

```
# ... code ...
```

...Answer...

Problem 4.4 (20 points)

In stochastic gradient boosting the bagging fraction and learning rate will govern the construction of the trees as they are guided by the gradient. Although the optimal values of these parameters should be obtained through the tuning process, it is helpful to understand how the magnitudes of these parameters affect magnitudes of variable importance. Figure 8.24 provides the variable importance plots for boosting using two extreme values for the bagging fraction (0.1 and 0.9) and the learning rate (0.1 and 0.9) for the solubility data. The left-hand plot has both parameters set to 0.1, and the right-hand plot has both set to 0.9:

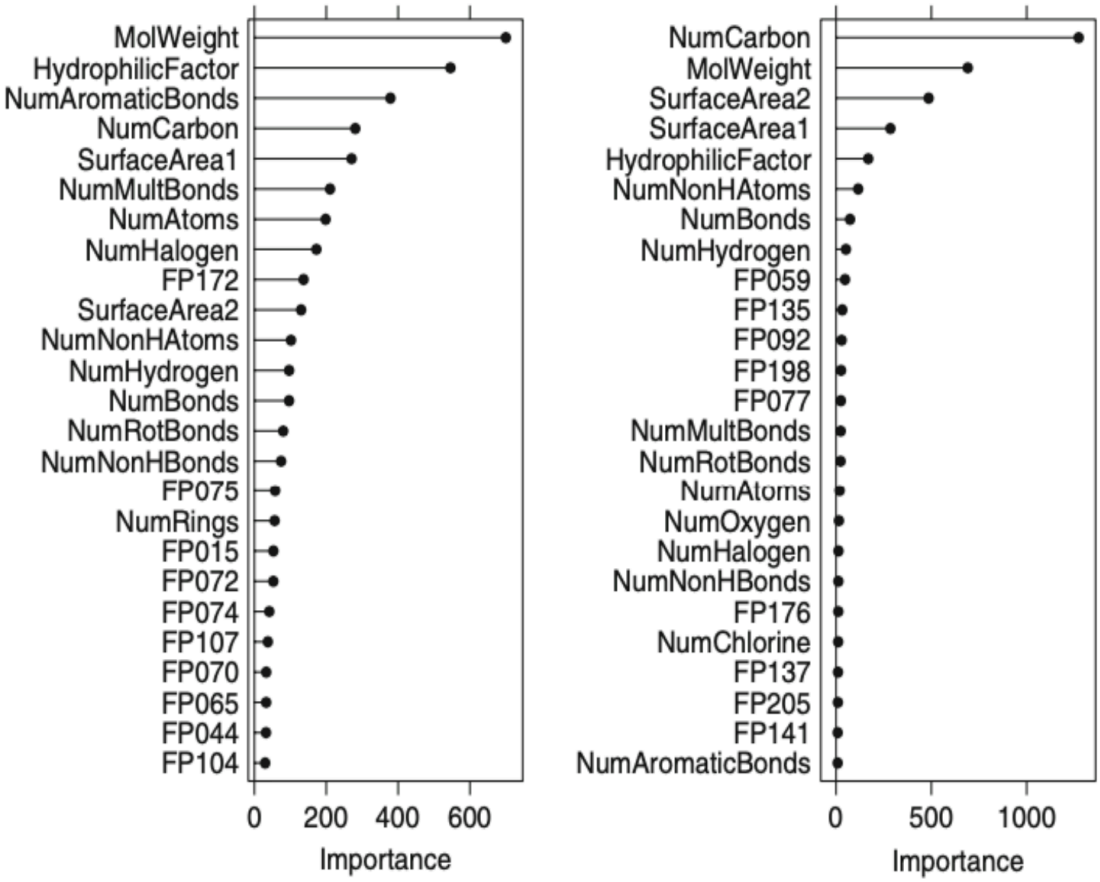


Fig. 8.24: A comparison of variable importance magnitudes for differing values of the bagging fraction and shrinkage parameters. Both tuning parameters are set to 0.1 in the *left* figure. Both are set to 0.9 in the *right* figure

4.3.a (7 points)

Why does the model on the right focus its importance on just the first few of predictors, whereas the model on the left spreads importance across more predictors?

...Answer...

4.3.b (7 points)

Which model do you think would be more predictive of other samples?

...Answer...

4.3.c (6 points)

How would increasing interaction depth affect the slope of predictor importance for either model in the figure below?

...Answer...