Support Vector Regression

Author : Rose Ellison

Given a set of data points, $\{(x_1, z_1), \dots, (x_l, z_l)\}$, such that $x_i \in R^n$ is an input and $z_i \in R^1$ is a target output, the standard form of support vector regression is:

I will be building support vector regression(SVR) model based off the 'Position_Salaries' data to determine 1.) determine if the model fits the data and 2.) to determine if a particuilar new hire's past salary was possibly $160,000 as a region manager. In this dataset there are three columns *Position*, *Level*, and *Salary*. Salary is our dependent variable while the other two are our independent variables.

Preparing the Data

```
# Set the seed
set.seed(1)

# Importing the data
positions <- read.csv('../../data/Position_Salaries.csv')

# Examine the Data
dim(positions)
```

```
## [1] 10  3
```

```
positions
```

```
##             Position Level  Salary
## 1   Business Analyst     1   45000
## 2  Junior Consultant     2   50000
## 3  Senior Consultant     3   60000
## 4            Manager     4   80000
## 5    Country Manager     5  110000
## 6     Region Manager     6  150000
## 7            Partner     7  200000
## 8     Senior Partner     8  300000
## 9            C-level     9  500000
## 10               CEO    10 1000000
```

From the table we can see there is some redundancy between the *Position* and *Level* column. Therefor it would make sense to drop the *Position* column and just use the numeric *Level* and *Salary* columns. Since we only have 10 observations, it would not be useful to split the data into a training and test set.

```
# Saving the dataset with only the two necessary columns
positions <- positions[, 2:3]
```

SVR Regressor

```
regressor <- svm(formula = Salary ~ Level,
                 data = positions,
                 type = 'eps-regression')
```

Predicting a Result

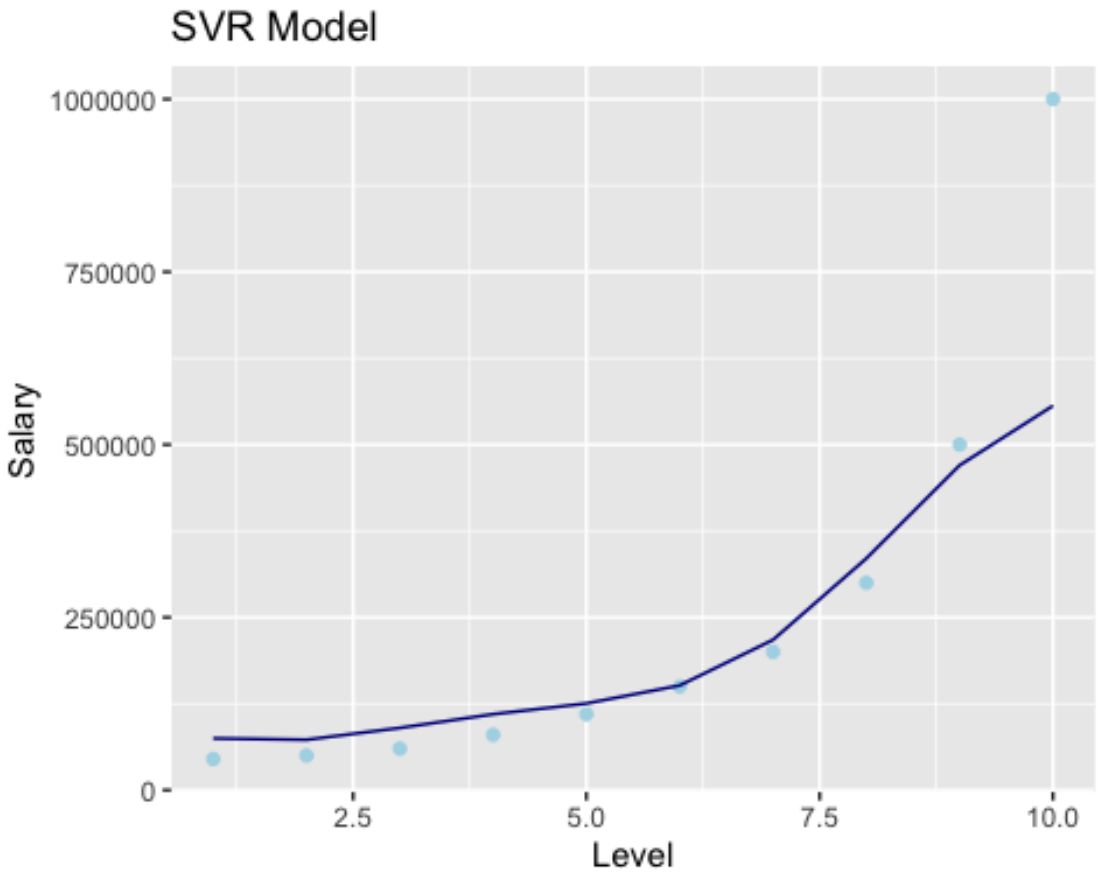
Is it likely that the new hire's past salary was actually $160,000 as a level 6.5?

```
y.pred <- predict(regressor, data.frame(Level = 6.5))
```

Visualizing the SVR

Visualizing the data will allow us to see if the SVR is a good model for the data.

```
# Visualizing the Support Vector Regression
ggplot() +
  geom_point(data = positions, aes(x = Level, y = Salary), col = 'lightblue')
+
  geom_line(aes(x = positions$Level, y = predict(regressor, newdata =
positions)), col = 'darkblue') +
  ggtitle('SVR Model') +
  xlab('Level') +
  ylab('Salary')
```



Conclusion

This model fits well with all of the data EXCEPT the CEO level. The CEO can also be categorized as an outlier. It is clear that the SVR model is not calculating any outliers in it's model.