Problem Set 2
Fitting Models to Data

Christoph Riedl

Business Analytics

To complete this homework assignment, download the homework template `PS2_Template.Rmd` and complete the whole assignment using R Markdown. Submit the final PDF you created using `Knit PDF` or `Knit HTML`. Blackboard does not accept HTML files. Just rename your file to ".txt" and upload it that way.

1 Linear Regression 1 [*Applied*]

The data folder on Blackboard contains a dataset called `ps2_ex1`. The dataset contains outcome y and inputs x_1, x_2 for 40 data points. There are an additional 20 obersvations with inputs but no observed outcome. Download the dataset to your working directory and read it into `R` using the `read.csv()` function.

1. Use `R` to fit a linear regression model predicting y from x_1, x_2, using the first 40 data points in the file. Summarize the inferences and check the fit of your model. Explain your findings. Hint: `R` will automatically detect that y is `NA` for the last 20 observations and will automatically drop them while running `lm()` - so there is nothing specific you need to do. You can confirm this by looking at the "number of observations" shown in the `summary()` output.
2. Display the estimated model graphically (Hint: numerical variables make great candidates for for x/y axes; categorical variables can be mapped to colors.)
3. Make predictions for the remaining 20 data points in the file that have missing observations. How confident do you feel about these predictions?

2 Linear Regression 2 [*Applied*]

In this exercise you will simulate two variables that are statistcally independent of each other to see what happens when we run a regression of one on the other

- First, generate 1,000 data points from a normal distribution with mean 0 and standard deviaiton 1 by typing `var1 <- rnorm(1000, 0, 1)`. Generate a second variable in the same way (call it `var2`). Run a regression of one variable on the other. Is the slope coeffcient significant? Interpret and explain. How much variation does the model explain? Why is this?

3 Linear Regression 3 [*Applied*]

In this exercise you will be working with New York City (Manhattan) housing sales data. The data is taken from NYC.gov but I created a simplified, cleaned-up dataset for you to download on Blackboard: `rollingsales_manhattan.csv`.

0. Load the data in `R` and explore the overall structure of the dataset.
1. Analyze sales using regression with any predictors you feel are relevant. Hint: `SALE.PRICE` is the target variable of interest and should hence be on the left-hand side of the regressions. Since the sales data is very skewed (you can verify this by plotting the data as a histogram), you should perform a `log()` transformation of the sales variable first.
2. Interpret your findings. Where would you recommend we look for affordable housing in Manhattan?
3. Justify why regression was appropriate to use.
4. Describe some decisions that a city planning official might make based on your analysis.

4 Support-Vector Machines [*Applied*]

In this exercise you will train a spam classifier using support vector machines. We will use the `spam` dataset which comes with the `{kernlab}` package. First, we will split the `spam` data randomly into two halves: one half we will use as the training data, the other half we will use as the test data. The target variable is `type` which is a binary class `spam` and `nospam`.

0. Look at the help page for the dataset to find out what the different columns mean (hint: `?spam`).
1. Fit a support vector classifier using `svm()` on the training data. `type` is the target and all other variables can be used as predictors (hint: you can use the `.` notation which automatically includes all columns of the `data.frame` as predictors except the target variable).
2. Predict `spam`/`nonspam` classes for the data in the test dataset. How does the predicted classification compare with the true classes? What is the classification error?
3. Can you improve the classification accuracy? (Hint: Start by exploring different settings for the `cost` attribute and using different predictors.)
4. How easy is it to interpret the classification performed using svm? Compare the interpretability of the svm model to that of a regression model (e.g., like the one from the exercises above).

Use the following code fragment to get you started.

```
# install.packages("kernlab")
library(e1071)
library(kernlab)
data(spam)
dim(spam)
head(spam)

set.seed(02115)
```

```
sample <- sample( c(TRUE, FALSE), nrow(spam), replace=TRUE)
train <- spam[sample,]
test <- spam[!sample,]
...
```