Variable Selection

Brooke Anderson

February 29, 2016

```
knitr::opts_knit$set(root.dir = "..") # Reset root directory for analysis
library(lubridate) # To help handle dates
library(dplyr) # Data wrangling
library(ggplot2) # Plotting
```

Read in and clean up the data:

```
train <- read.csv("data/train.csv", as.is = TRUE) # `as.is` so `datetime` comes in as
                                                  # character, not factor
test <- read.csv("data/test.csv", as.is = TRUE)

train <- mutate(train,
                datetime = ymd_hms(datetime),
                year = factor(year(datetime)),
                hour = factor(hour(datetime)),
                month = month(datetime),
                yday = yday(datetime),
                weather = factor(weather, levels = c(1, 2, 3, 4),
                                 labels = c("Clear", "Mist", "Light Precip",
                                            "Heavy Precip")),
                season = factor(season, levels = c(1, 2, 3, 4),
                                labels = c("Spring", "Summer", "Fall", "Winter")),
                workingday = factor(workingday, levels = c(0, 1),
                                    labels = c("Holiday / weekend",
                                               "Working day")))
test  <- mutate(test,
                datetime = ymd_hms(datetime),
                year = factor(year(datetime)),
                hour = factor(hour(datetime)),
                month = month(datetime),
                yday = yday(datetime),
                weather = factor(weather, levels = c(1, 2, 3, 4),
                                 labels = c("Clear", "Mist", "Light Precip",
                                            "Heavy Precip")),
                season = factor(season, levels = c(1, 2, 3, 4),
                                labels = c("Spring", "Summer", "Fall", "Winter")),
                workingday = factor(workingday, levels = c(0, 1),
                                    labels = c("Holiday / weekend",
                                               "Working day")))
```

Add some derived variables related to weather and temperature, using `dplyr` for data wrangling (here's a good cheatsheet):

```
#train <- dplyr(train, )
```