Modeling in Data Science

Course: Introduction to Data Science

Unit Outline

01 Introduction to Linear Regression

02 Linear Regression Coding

03 Linear Regression Practice

04 Introduction to KNN

05 KNN Coding

06 KNN Practice

07 KNN Practice

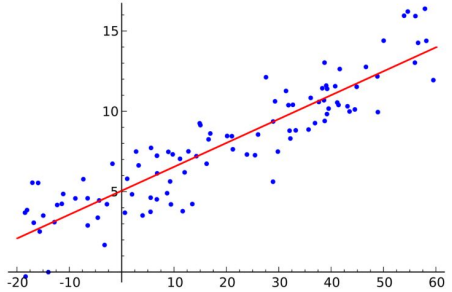
08 Unit Project

09 Unit Project

10 Unit Project

Linear Regression

- Students will learn about linear regression from a mathematical perspective
- Students will learn how to code a linear regression fit in R
- Students will practice coding linear regression fits with various datasets to help them see both the power and limitations of linear regression



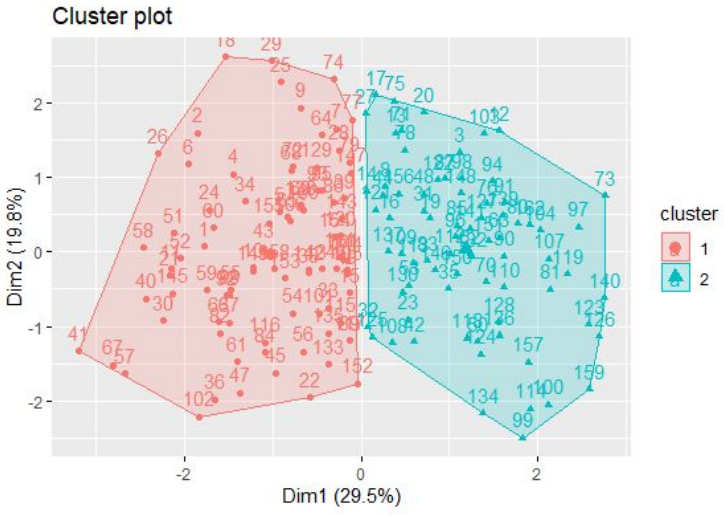
Linear Regression

FROM SCRATCH

```
123     fc <- lm(mean~Grade_12_Year, data = subjectTrend)
124     slope <- as.numeric(fc$coefficients[2])
```

KNN (k-nearest neighbors)

- Students will learn about KNN from a non-mathematical perspective
    - This will be done from a more conceptual level, focusing on groupings of similarities (similar to the movie experiment we did)
- Students will learn how to write code that creates KNN clusters in R
- Students will practice KNN regression on different datasets to get a feel for how they can manipulate the input parameters to get different groupings




```r
buildKNNModel <- function(splits, train, test, val_set, method){
  if(method == 'regression'){
    metrics <- metric_set(rmse)
    best_metric <- 'rmse'
    knn_rec <- recipe(target ~., data = train) %>%
      step_zv(all_predictors()) %>%
      step_normalize(all_predictors())
  } else{
    metrics <- metric_set(accuracy)
    best_metric <- 'accuracy'
    knn_rec <- recipe(target ~., data = train) %>%
      step_dummy(all_nominal_predictors()) %>%
      step_zv(all_predictors()) %>%
      step_normalize(all_predictors())
  }

  knn_mod <- nearest_neighbor(mode = method,
                              engine = "kknn",
                              neighbors = tune(),
                              weight_func = tune(),
                              dist_power = tune())

  knn_workflow <- workflow() %>%
    add_model(knn_mod) %>%
    add_recipe(knn_rec)
```

Unit Project

- Students will complete a Kaggle competition
- The competition includes a great real life dataset of homes and their selling price
- Students must attempt to create the most accurate model to predict the selling price of the homes in the dataset
- They may use either linear regression or KNN regression to complete the task

kaggle

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.