STAT 380 – kNN Regression Part 1 (Lecture 8)

IDEA: Suppose we observe a quantitative response, Y, and p predictors, $X_1, X_2, \dots, X_p$. Assuming there is a relationship between Y and $X = (X_1, X_2, \dots, X_p)$, we describe the general form of the relationship as:

General Regression Form: $Y = f(X) + \epsilon$

Linear Regression

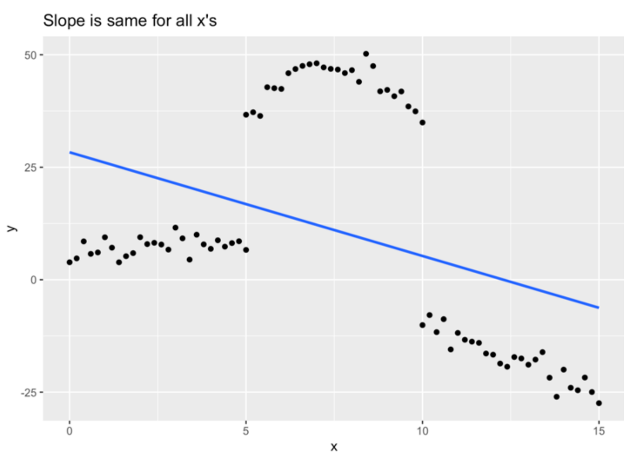
- Assumes a linear form for $f(X)$, namely

$$f(X) = \beta_0 + \beta_1 x_{i1} + \cdots$$

- Estimate $f(X)$ with

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots$$

- Not flexible (e.g., $\hat{\beta}_1$ is same for all values of x_{i1}, linear form is same everywhere)



k-Nearest Neighbors (kNN)

- Estimate $f(X)$ with

$$\hat{f}(X_0) = \frac{1}{k} \sum_{X_i \in N_0} y_i$$

where

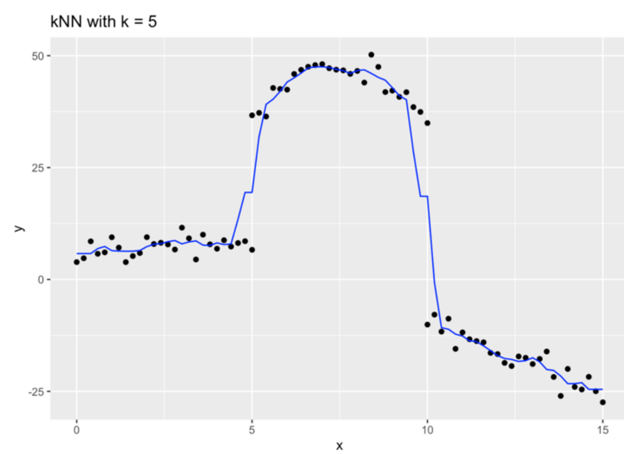
$X_0 =$ set of inputs for new obs.

$X_i =$ set of inputs for i$^{\text{th}}$ obs. in Train

$N_0 =$ collection of obs in NN set

$k =$ number of NN

- Flexible

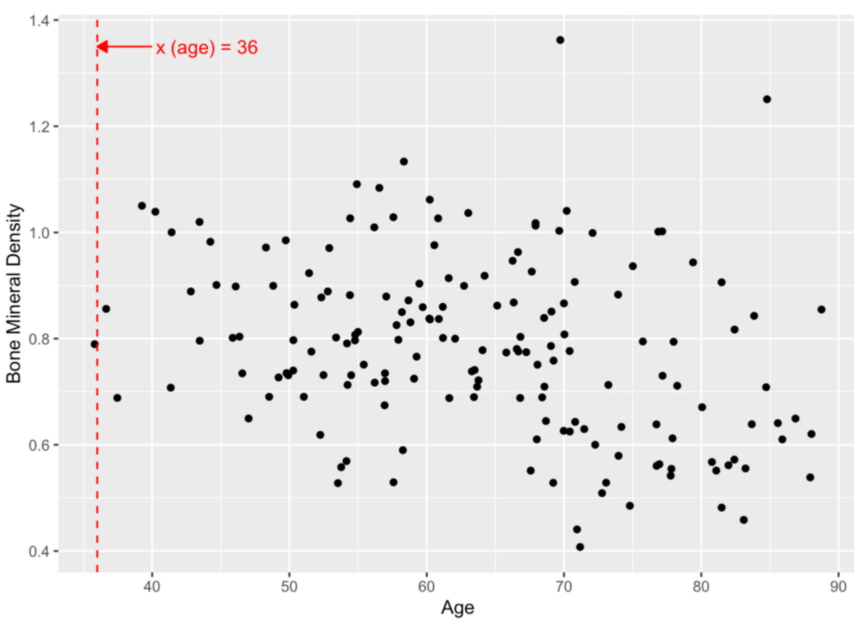


IDEA: The k-nearest neighbors (kNN) algorithm is a simple approach that can be used for both regression and classification. The intuition is that if we want to predict the response (continuous or categorical) for a new observation (i.e., for a new set of inputs/x's), we find the k nearest neighbors.

NOTE: Throughout this lecture, we will use the file L8_bmd.csv which contains 169 records of bone densitometries (measurement of bone mineral density). The following variables are included:

Variable	Meaning
id	Patient's identification number
age	Patient's age
sex	Patient's sex
fracture	A categorical variable indicating whether the patient has had a hip fracture
weight_kg	Patient's weight in kilograms
height_cm	Patient's height in centimeters
medication	Patient's medication status
waiting_time	Minutes patient spent waiting for the densitometry
bmd	Bone mineral density measurement in the hip

EXAMPLE 1: Consider the illustration shown below for the L8_bmd.csv dataset. Suppose we wish to predict the bone mineral density (bmd) using the person's age. The red (dashed) vertical line corresponds to a person who is 36 years old. Explain how you would predict the bmd for a 36 year old using the kNN algorithm, where k=3.



NOTE: In order to find the "nearest" neighbors, we must select a distance metric to determine closeness. Euclidean distance is commonly used. Recall that in $\mathbb{R}^2$, for two points $i = (x_{i1}, x_{i2})$ and $j = (x_{j1}, x_{j2})$, the Euclidean distance is given by:

$$d = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

In $\mathbb{R}^3$, suppose we have two points $i = (x_{i1}, x_{i2}, x_{i3})$ and $j = (x_{j1}, x_{j2}, x_{j3})$, the Euclidean distance is given by:

$$d = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2}$$

In $\mathbb{R}^k$, suppose we have two points $i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{ik})$ and $j = (x_{j1}, x_{j2}, x_{j3}, \dots, x_{jk})$, the Euclidean distance is given by:

$$d = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2 + \cdots + (x_{ik} - x_{jk})^2}$$

NOTE: In Example 1, the distance would be given by: $d = \sqrt{(x_{i1} - x_{j1})^2}$.

EXAMPLE 2: Although not a regression problem, suppose we wanted to use kNN to predict whether a person will default on their credit card payment using student, monthly credit card balance, and yearly income. A snapshot of a dataset is shown below:

	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494
5	No	No	785.65588	38463.496
6	No	Yes	919.58853	7491.559
7	No	No	825.51333	24905.227

a. How could we incorporate student in the kNN distance calculations?

b. Which predictor (x) variable will have the largest impact on the distance calculations in this situation? Explain.

NOTE: To reduce the effect of the scale of our data, we often standardize the variables. “Standardize” can mean a number of things, but we will use it to mean that the variable is made to have a mean of 0 and a standard deviation of 1 (i.e., a z-score). We can do this using the ‘scale’ function in R. (An example of how to do this is upcoming.)

EXAMPLE 3: Using the bmd data, suppose we want to predict bmd using age, weight, and medication using kNN with k = 5.

a. Prepare the data by 1) creating appropriate indicators, 2) scaling the input (x) data including indicators, and 3) performing a 85/15 training/testing split using a seed of 123.

```
#First, determine the levels of medication
bmd %>%
  group_by(medication) %>%
  summarize(N = n())
```

```
## # A tibble: 3 × 2
##   medication          N
##   <chr>           <int>
## 1 Anticonvulsant      9
## 2 Glucocorticoids    24
## 3 No medication     136
```

```r
#Create indicators - since medication has 3 levels, create 2 indicators
bmd <- bmd %>%
        mutate(Anti = ifelse(medication == "Anticonvulsant", 1, 0),
               Gluc = ifelse(medication == "Glucocorticoids", 1, 0))

#Scale my numeric variables
xvars <- c("age", "weight_kg", "Anti", "Gluc")
bmd[ , xvars] <- scale(bmd[ , xvars], center = TRUE, scale = TRUE)

#Training/Testing split
set.seed(123)
train_ind <- sample(1:nrow(bmd), floor(0.85 * nrow(bmd)))
set.seed(NULL)

Train <- bmd[train_ind, ]
Test <- bmd[-train_ind, ]
```

b. Build the kNN model with k = 5 and compute the RMSE for the testing set. Use the knn.reg() function from the FNN package.

```r
#Build Model
knn_res <- knn.reg(train = Train[ , xvars, drop = FALSE],
                   test = Test[ , xvars, drop = FALSE],
                   y = Train$bmd,
                   k = 5)

#Get Predictions
knn_res$pred
```

```r
#Find mSE
mse_knn5 <- mean((Test$bmd - knn_res$pred)^2)
mse_knn5
```

```
## [1] 0.03356567
```

```r
#Find RMSE
rmse_knn5 <- sqrt(mse_knn5)
rmse_knn5
```

```
## [1] 0.1832094
```

c. Using the same training set, build a multiple linear regression model using the same predictors age, weight, and medication. Compute the RMSE for the testing set.

d. Which model produces better predictions on new data?

```
#Compare RMSE's (or MSE's) - smaller is better
rmse_knn5
```

```
## [1] 0.1832094
```

```
rmse_reg
```

```
## [1] 0.1535161
```

e. Which model allows us to answer the question: How does the age affect bmd?