Lecture #4: Introduction to Regression

Data Science 1
CS 109A, STAT 121A, AC 209A, E-109A

Pavlos Protopapas   Kevin Rader
Margo Levine   Rahul Dave

Lecture Outline

Announcements

Data

Statistical Modeling

Regression vs. Classification

Error, Loss Functions

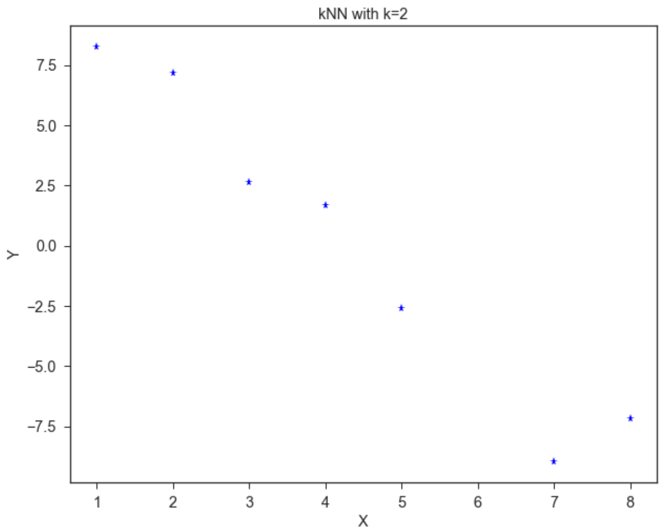
Model I: k-Nearest Neighbors

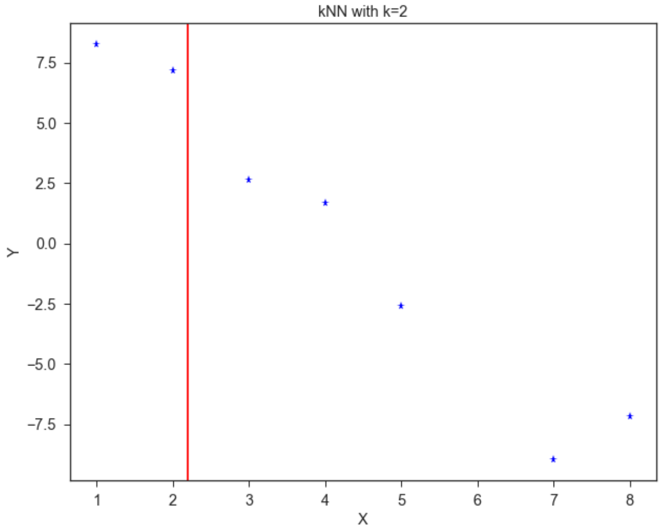
Model II: Linear Regression

Evaluating Model

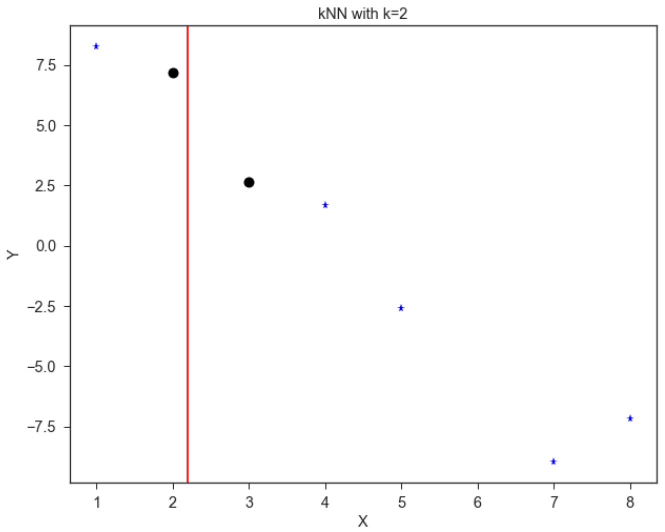
Comparison of Two Models

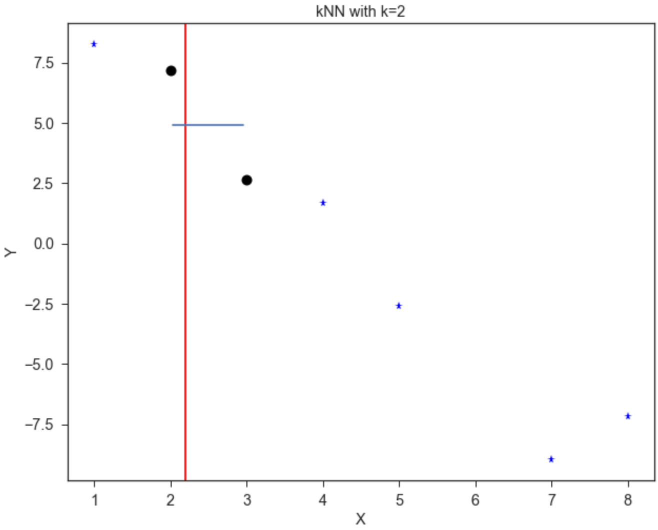
Model I: k-Nearest Neighbors


kNN with k=2
Y
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
1
2
3
4
5
6
7
8
X

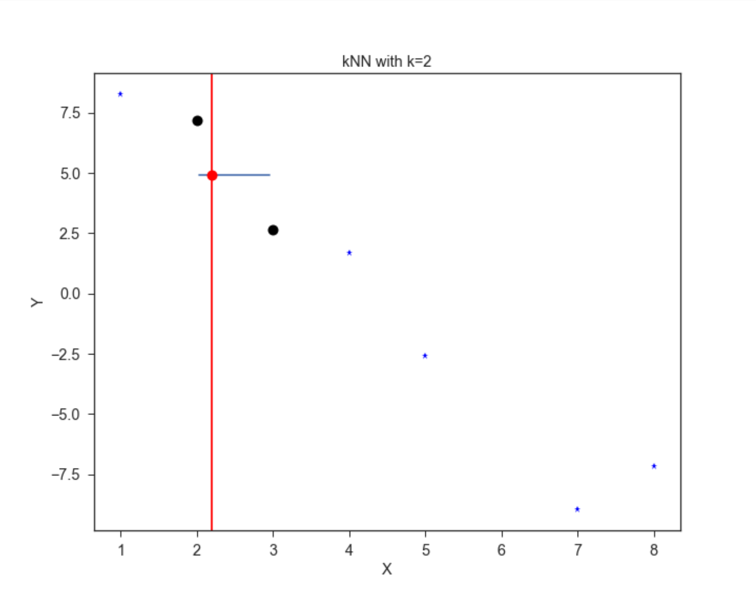


kNN with k=2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
Y
1
2
3
4
5
6
7
8
X


kNN with k=2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
Y
1
2
3
4
5
6
7
8
X



kNN with k=2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
Y
1
2
3
4
5
6
7
8
X


kNN with k=2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
Y
1
2
3
4
5
6
7
8
X

k-Nearest Neighbors

The k***-Nearest Neighbor (kNN) model*** is an intuitive way to predict a quantitative response variable:

> *to predict a response for a set of observed predictor values, we use the responses of other observations most similar to it!*

Note: this strategy can also be applied in classification to predict a categorical variable. We will encounter kNN again later in the semester in the context of classification.

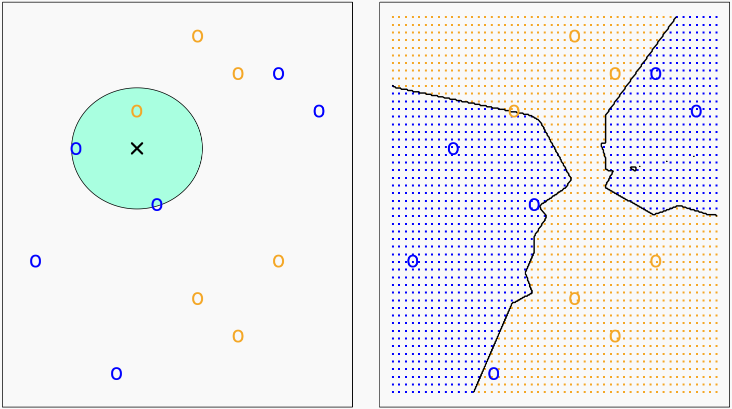
k-Nearest Neighbors

Fixed a value of k. The predicted response for the i-th observation is the average of the observed response of the k-closest observations

$$\widehat{y}_i = \frac{1}{k}\sum_{i=1}^{k} y_{n_i}$$

where $\{X_{n_1}, \ldots, X_{n_k}\}$ are the k observations most similar to X_i (similar refers to a notion of distance between predictors).

k-Nearest Neighbors for Classification



kNN Regression: A Simple Example

Suppose you have 5 observations of taxi cab pick ups in New York City, the response is the average cab fare (in units of $10), and the predictor is time of day (in hours after 7am):

X	1	2	3	4	5
Y	6	7	4	3	2

We calculate the predicted number of pickups using kNN for $k = 2$:

$$X = 1 \qquad \widehat{y}_1 = \frac{1}{2}(7 + 4) = 5.5$$

kNN Regression: A Simple Example

Suppose you have 5 observations of taxi cab pick ups in New York City, the response is the average cab fare (in units of $10), and the predictor is time of day (in hours after 7am):

X	1	2	3	4	5
Y	6	7	4	3	2

We calculate the predicted number of pickups using kNN for $k = 2$:

$$X = 2 \qquad \widehat{y}_2 = \frac{1}{2}(6+4) = 5.0$$

kNN Regression: A Simple Example

Suppose you have 5 observations of taxi cab pick ups in New York City, the response is the average cab fare (in units of $10), and the predictor is time of day (in hours after 7am):

X	1	2	3	4	5
Y	6	7	4	3	2

We calculate the predicted number of pickups using kNN for $k = 2$:

$$\widehat{Y} = (5.5, 5.0, 5.0, 3.0, 3.5)$$

kNN Regression: A Simple Example

Suppose you have 5 observations of taxi cab pick ups in New York City, the response is the average cab fare (in units of $10), and the predictor is time of day (in hours after 7am):

X	1	2	3	4	5
Y	6	7	4	3	2

We calculate the predicted number of pickups using kNN for $k = 2$:

$$\widehat{Y} = (5.5, 5.0, 5.0, 3.0, 3.5)$$

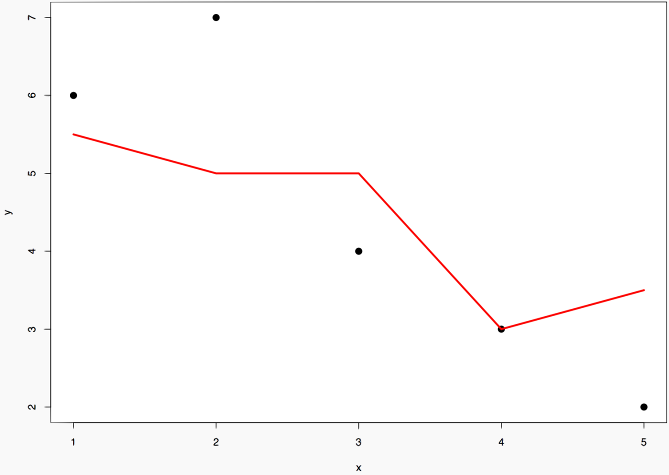
The MSE given our predictions is

$$MSE = \frac{1}{5}\left[(6-5.5)^2 + (7-5.0)^2 + \ldots + (3.5-2)^2\right] = 1.5$$

On average, our predictions are off by $15.

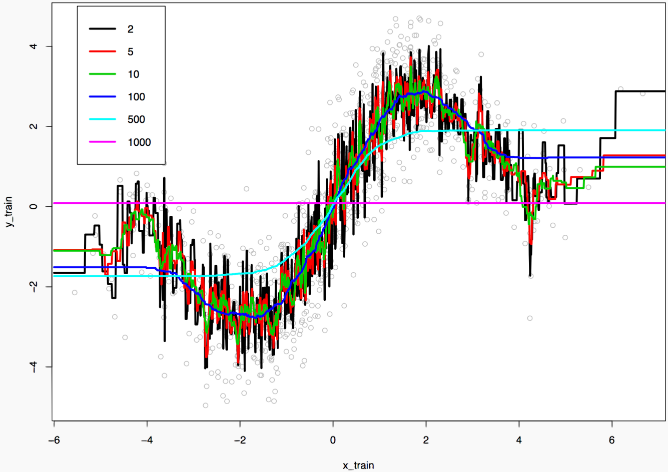
kNN Regression: A Simple Example

We plot the observed responses along with predicted responses for comparison:



Choice of k Matters

But what value of k should we choose? What would our predicted responses look like if k is very small? What if k is large (e.g. $k = n$)?



kNN with Multiple Predictors

In our simple example, we used absolute value to measure the distance between the predictors in two different observations, $|x_i - x_j|$.

When we have multiple predictors in each observation, we need a notion of distance between two **sets** of predictor values. Typically, we use Euclidean distance:

$$d(x_i - x_j) = \sqrt{(x_{i,1} - x_{j,1})^2 + \ldots + (x_{i,p} - x_{j,p})^2}$$

Caution: when using Euclidean distance, the scale (or units) of measurement for the predictors matter! Predictors with large values, comparatively, will dominate the distance measurement.