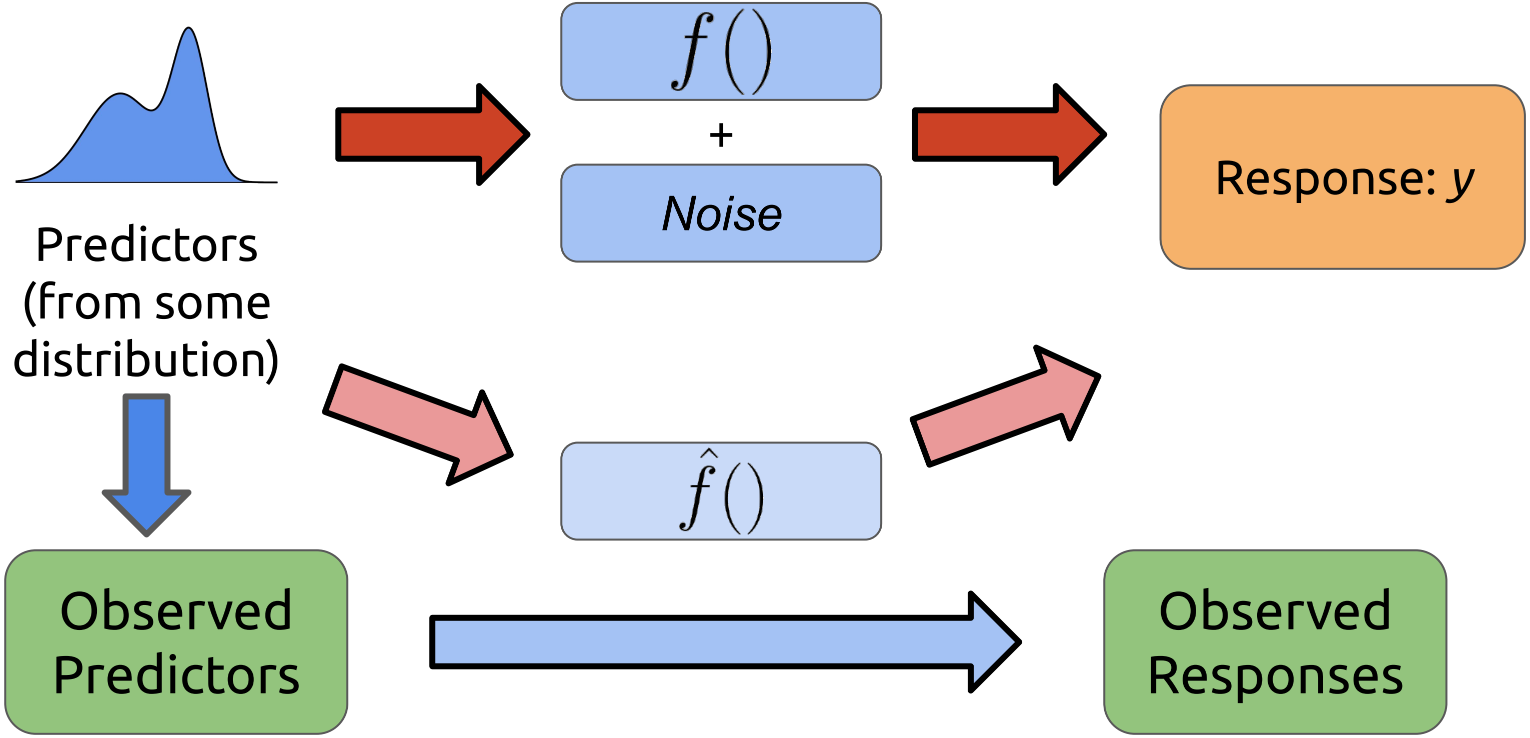
Instance-Based Learning and Decision Trees

Reminder to Michael - Hit record!

Supervised Learning



Supervised Learning - How?

We need to pick a way to make predictions from our available training data.

Supervised Learning - How?

We need to pick a way to make predictions from our available training data.

There are many, many ways to do this.

Supervised Learning - How?

We need to pick a way to make predictions from our available training data.

There are many, many ways to do this.

For example, we can pick a functional form for $\hat{f}()$

Supervised Learning - How?

We need to pick a way to make predictions from our available training data.

There are many, many ways to do this.

For example, we can pick a functional form for $\hat{f}()$

Linear regression use a particularly simple functional form to make predictions - a weighted sum of the predictor variables.

Linear Regression

Given *k* predictors $x^{(1)}$, $x^{(2)}$,...,$x^{(k)}$, linear regression uses the following equation to predict the target variable:

$$\hat{f}(\vec{x}) = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_k x^{(k)}$$

Here, β_0, β_1,...,β_k are constants that are determined by using the available training data.

Supervised Learning - How?

Linear regression can do quite well on a range of problems.

Supervised Learning - How?

Linear regression can do quite well on a range of problems.

However, since we are picking a functional form, we are still somewhat restricted in the types of predictions we can make.

Supervised Learning - How?

Linear regression can do quite well on a range of problems.

However, since we are picking a functional form, we are still somewhat restricted in the types of predictions we can make.

In these slides, we will look at some alternative types of models - ones that don’t rely on picking a particular type of function in order to make predictions.

K-Nearest Neighbors

Big Idea: Make predictions about new data by finding the most similar observations in the training data and using their target values.

K-Nearest Neighbors

Big Idea: Make predictions about new data by finding the most similar observations in the training data and using their target values.

Requires us picking how many training instances to look at (a **hyperparameter**).

K-Nearest Neighbors - King County Example

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: ?

K-Nearest Neighbors - King County Example

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: ?

We scan through our training data for the most similar and find this one:

bedrooms	4
bathrooms	2.5
sqft_living	2280
sqft_lot	7200
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2280
sqft_basement	0
lat	47.3829
long	-122.207
sqft_living15	2250
sqft_lot15	7200
age_at_sale	20
years_since_renovation	20
zipcode	98030

K-Nearest Neighbors - King County Example

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: ?

We scan through our training data for the most similar and find this one:

bedrooms	4
bathrooms	2.5
sqft_living	2280
sqft_lot	7200
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2280
sqft_basement	0
lat	47.3829
long	-122.207
sqft_living15	2250
sqft_lot15	7200
age_at_sale	20
years_since_renovation	20
zipcode	98030

Price: $322,000

K-Nearest Neighbors - King County Example

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: $322,000

We scan through our training data for the most similar and find this one:

bedrooms	4
bathrooms	2.5
sqft_living	2280
sqft_lot	7200
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2280
sqft_basement	0
lat	47.3829
long	-122.207
sqft_living15	2250
sqft_lot15	7200
age_at_sale	20
years_since_renovation	20
zipcode	98030

Price: $322,000

K-Nearest Neighbors

Pros: Super simple idea and easy to implement.

Cons:

- In this implementation, we aren't taking into consideration the importance of each feature.
- We have a lot of redundant features, so these are getting "double-counted"
- We are currently relying on a single observation to make predictions.

K-Nearest Neighbors - 3 neighbors

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: ?

We scan through our training data for the most similar and find these 3:

K-Nearest Neighbors - 3 neighbors

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

We scan through our training data for the most similar and find these 3:

bedrooms	4
bathrooms	2.5
sqft_living	2280
sqft_lot	7200
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2280
sqft_basement	0
lat	47.3829
long	-122.207
sqft_living15	2250
sqft_lot15	7200
age_at_sale	20
years_since_renovation	20
zipcode	98030

bedrooms	3
bathrooms	2.5
sqft_living	2200
sqft_lot	7201
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2200
sqft_basement	0
lat	47.3821
long	-122.207
sqft_living15	2250
sqft_lot15	7240
age_at_sale	20
years_since_renovation	20
zipcode	98030

bedrooms	4
bathrooms	2.5
sqft_living	2210
sqft_lot	17715
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2210
sqft_basement	0
lat	47.3818
long	-122.2
sqft_living15	2210
sqft_lot15	16907
age_at_sale	17
years_since_renovation	17
zipcode	98030

Prediction: ?

K-Nearest Neighbors - 3 neighbors

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

We scan through our training data for the most similar and find these 3:

bedrooms	4	bedrooms	3	bedrooms	4
bathrooms	2.5	bathrooms	2.5	bathrooms	2.5
sqft_living	2280	sqft_living	2200	sqft_living	2210
sqft_lot	7200	sqft_lot	7201	sqft_lot	17715
floors	2	floors	2	floors	2
waterfront	0	waterfront	0	waterfront	0
view	0	view	0	view	0
condition	3	condition	3	condition	3
grade	8	grade	8	grade	8
sqft_above	2280	sqft_above	2200	sqft_above	2210
sqft_basement	0	sqft_basement	0	sqft_basement	0
lat	47.3829	lat	47.3821	lat	47.3818
long	-122.207	long	-122.207	long	-122.2
sqft_living15	2250	sqft_living15	2250	sqft_living15	2210
sqft_lot15	7200	sqft_lot15	7240	sqft_lot15	16907
age_at_sale	20	age_at_sale	20	age_at_sale	17
years_since_renovation	20	years_since_renovation	20	years_since_renovation	17
zipcode	98030	zipcode	98030	zipcode	98030

Price: $322,000

Price: $302,495

Price: $360,000

Prediction: ?

K-Nearest Neighbors - 3 neighbors

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: ?

We scan through our training data for the most similar and find these 3:

bedrooms	4
bathrooms	2.5
sqft_living	2280
sqft_lot	7200
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2280
sqft_basement	0
lat	47.3829
long	-122.207
sqft_living15	2250
sqft_lot15	7200
age_at_sale	20
years_since_renovation	20
zipcode	98030

Price: $322,000

bedrooms	3
bathrooms	2.5
sqft_living	2200
sqft_lot	7201
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2200
sqft_basement	0
lat	47.3821
long	-122.207
sqft_living15	2250
sqft_lot15	7240
age_at_sale	20
years_since_renovation	20
zipcode	98030

Price: $302,495

bedrooms	4
bathrooms	2.5
sqft_living	2210
sqft_lot	17715
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2210
sqft_basement	0
lat	47.3818
long	-122.2
sqft_living15	2210
sqft_lot15	16907
age_at_sale	17
years_since_renovation	17
zipcode	98030

Price: $360,000

Average Price: $328,165

K-Nearest Neighbors - 3 neighbors

We want to predict the sales for this house:

bedrooms	4
bathrooms	2.5
sqft_living	2240
sqft_lot	7589
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2240
sqft_basement	0
lat	47.3824
long	-122.207
sqft_living15	2250
sqft_lot15	7300
age_at_sale	21
years_since_renovation	21
zipcode	98030

Prediction: $328,165

We scan through our training data for the most similar and find these 3:

bedrooms	4
bathrooms	2.5
sqft_living	2280
sqft_lot	7200
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2280
sqft_basement	0
lat	47.3829
long	-122.207
sqft_living15	2250
sqft_lot15	7200
age_at_sale	20
years_since_renovation	20
zipcode	98030

Price: $322,000

bedrooms	3
bathrooms	2.5
sqft_living	2200
sqft_lot	7201
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2200
sqft_basement	0
lat	47.3821
long	-122.207
sqft_living15	2250
sqft_lot15	7240
age_at_sale	20
years_since_renovation	20
zipcode	98030

Price: $302,495

bedrooms	4
bathrooms	2.5
sqft_living	2210
sqft_lot	17715
floors	2
waterfront	0
view	0
condition	3
grade	8
sqft_above	2210
sqft_basement	0
lat	47.3818
long	-122.2
sqft_living15	2210
sqft_lot15	16907
age_at_sale	17
years_since_renovation	17
zipcode	98030

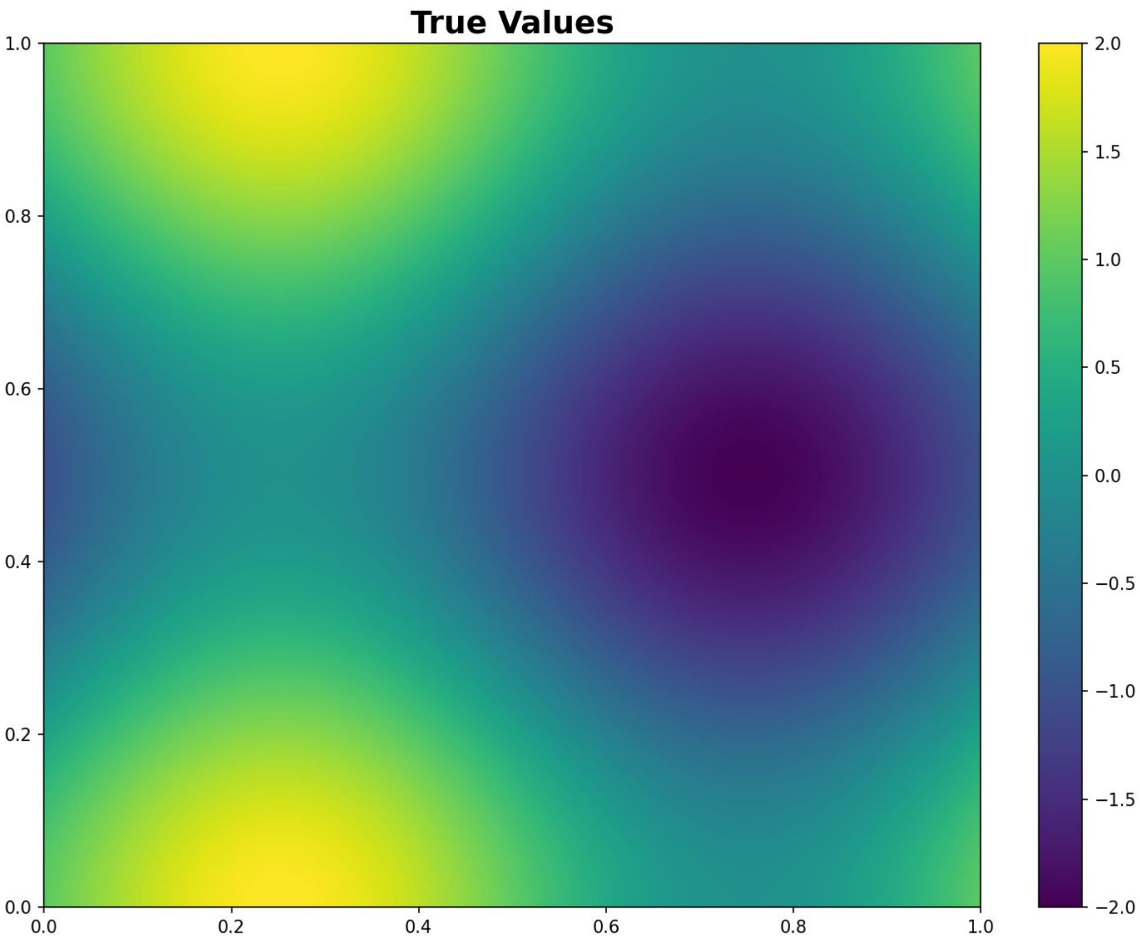
Price: $360,000

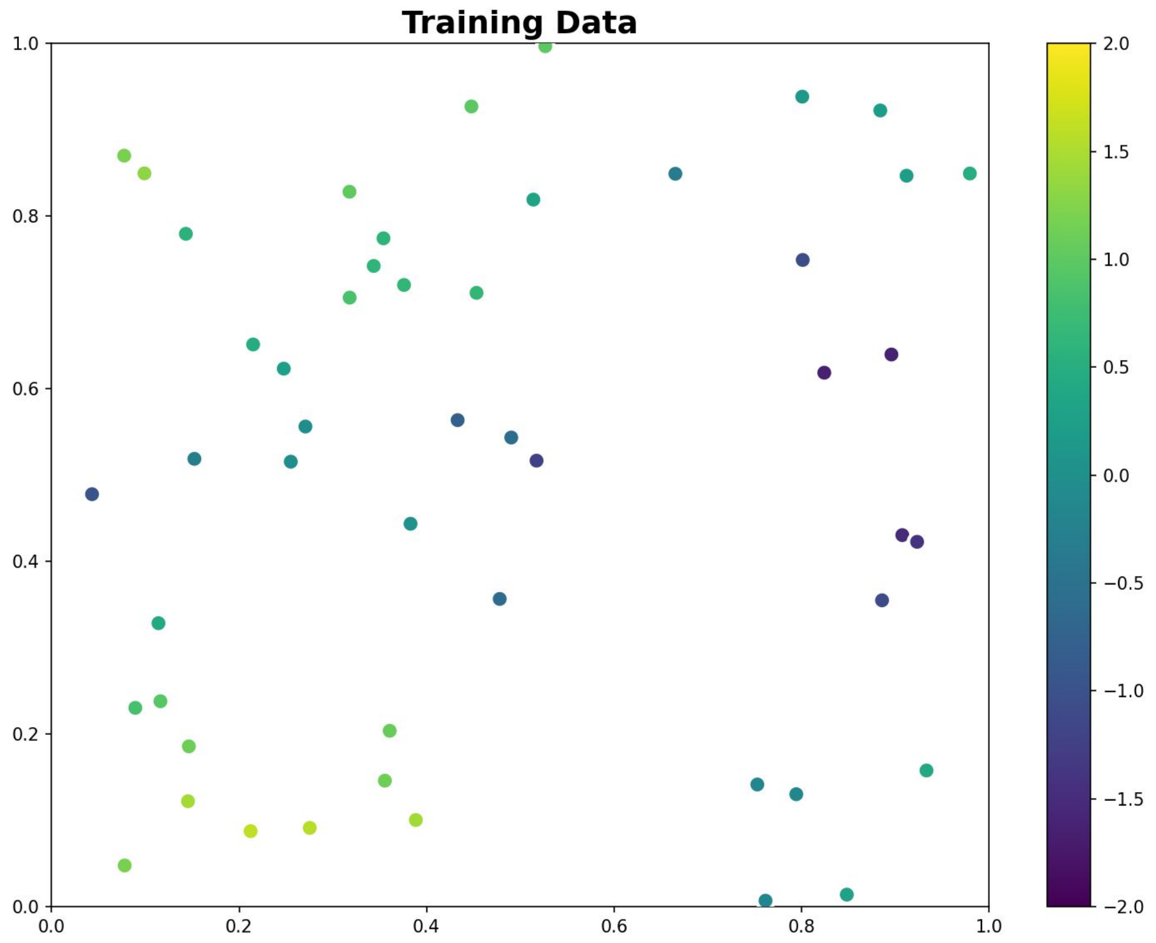
Average Price: $328,165

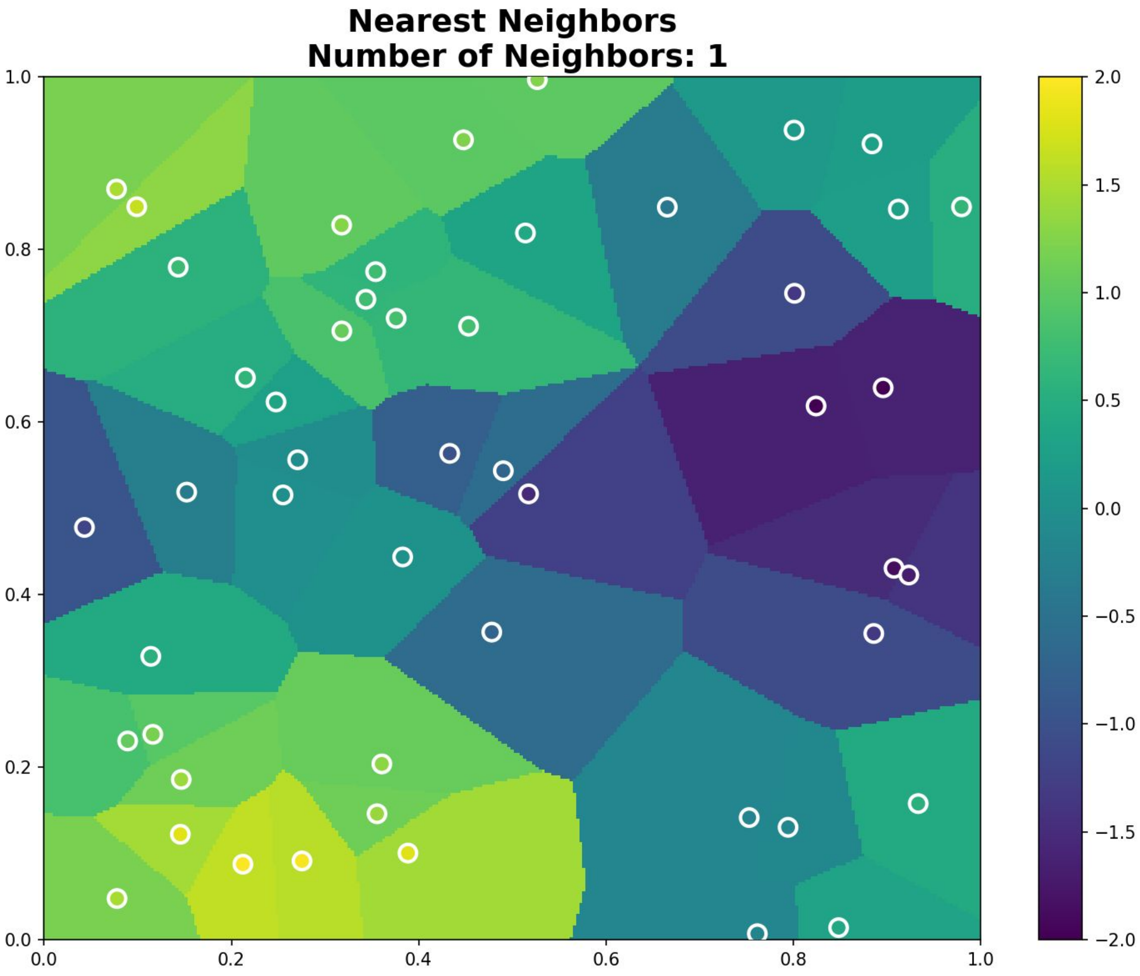
K-Nearest Neighbors - Example #2

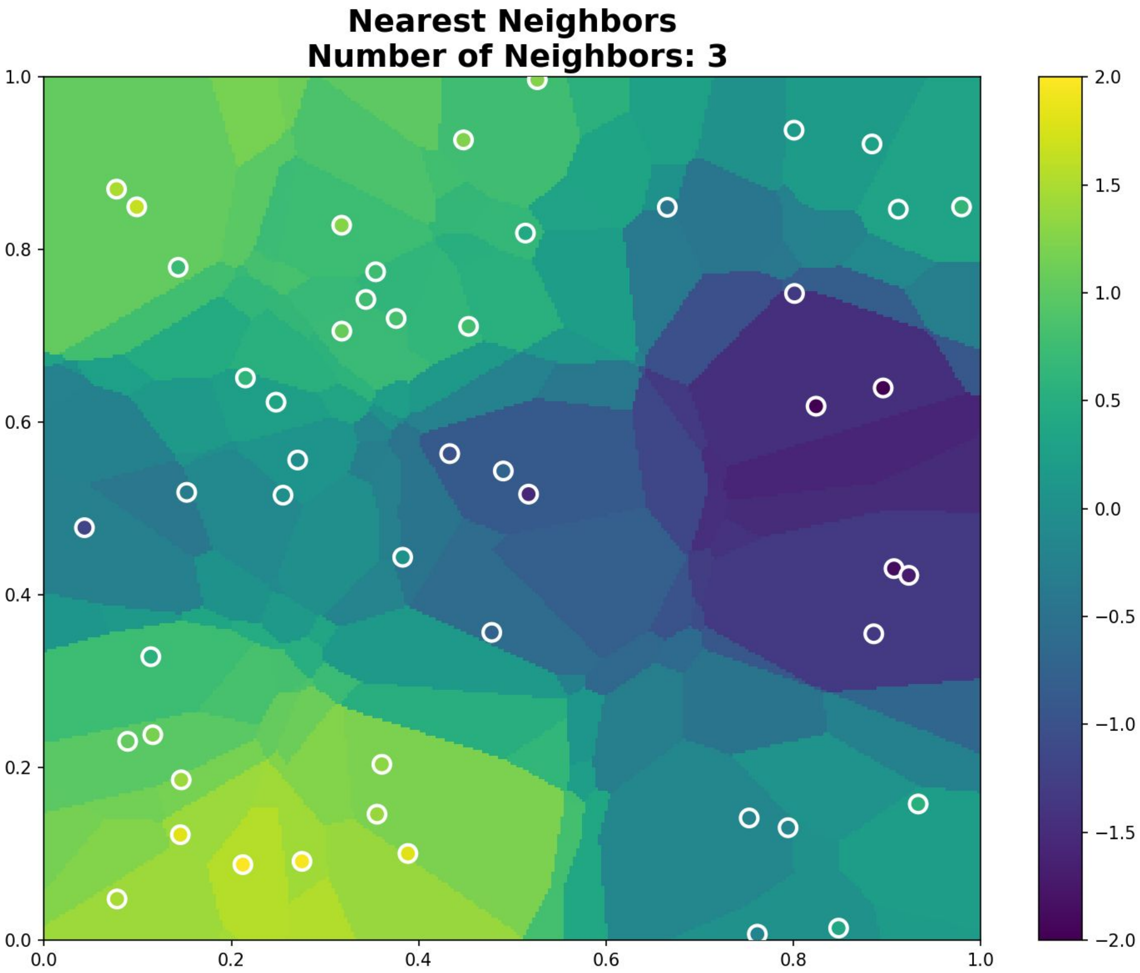
Let's say we are trying to make predictions on data that arose from

$$f(\vec{x}) = \sin(x_1) + \cos(x_2) + Noise$$

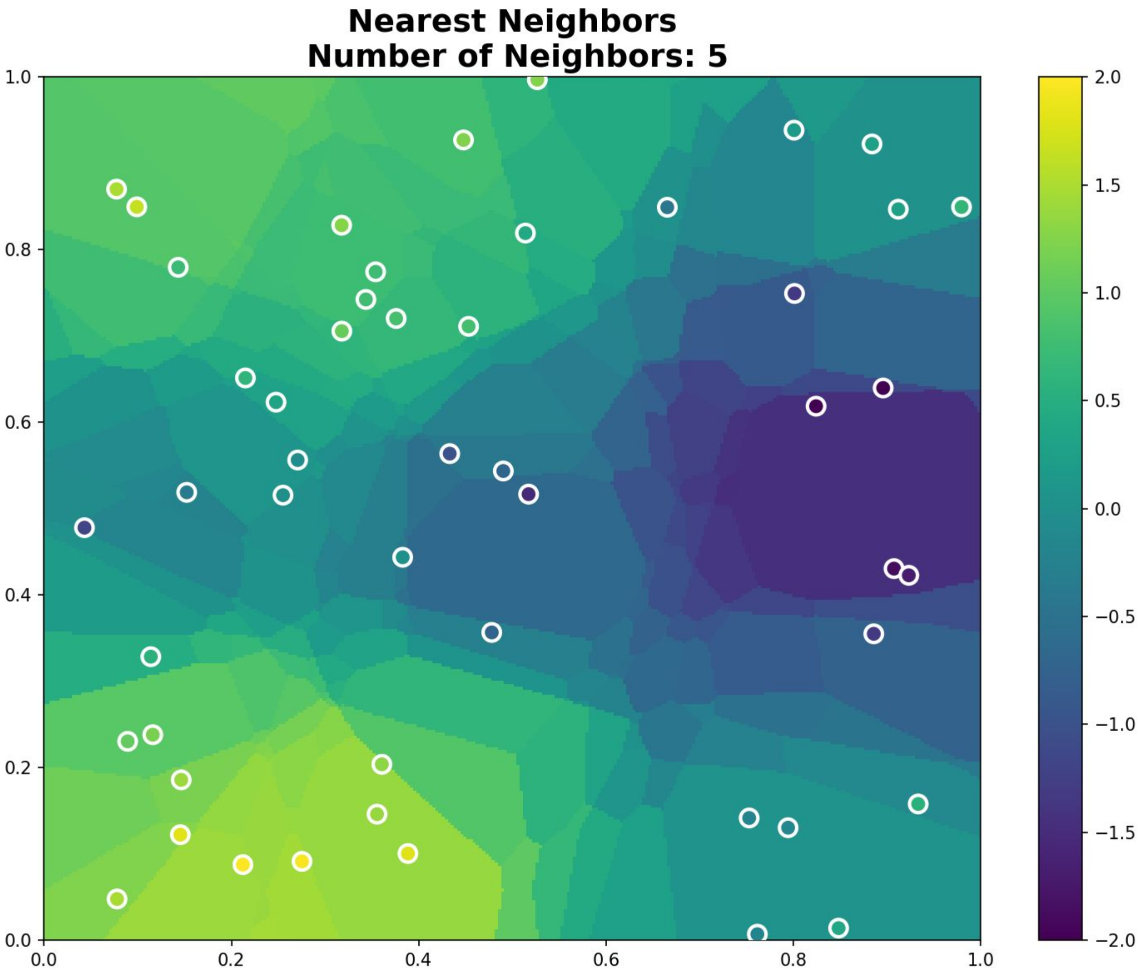

True Values
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0


Training Data
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0


Nearest Neighbors
Number of Neighbors: 1
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0



Nearest Neighbors
Number of Neighbors: 3
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0



Nearest Neighbors
Number of Neighbors: 5
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

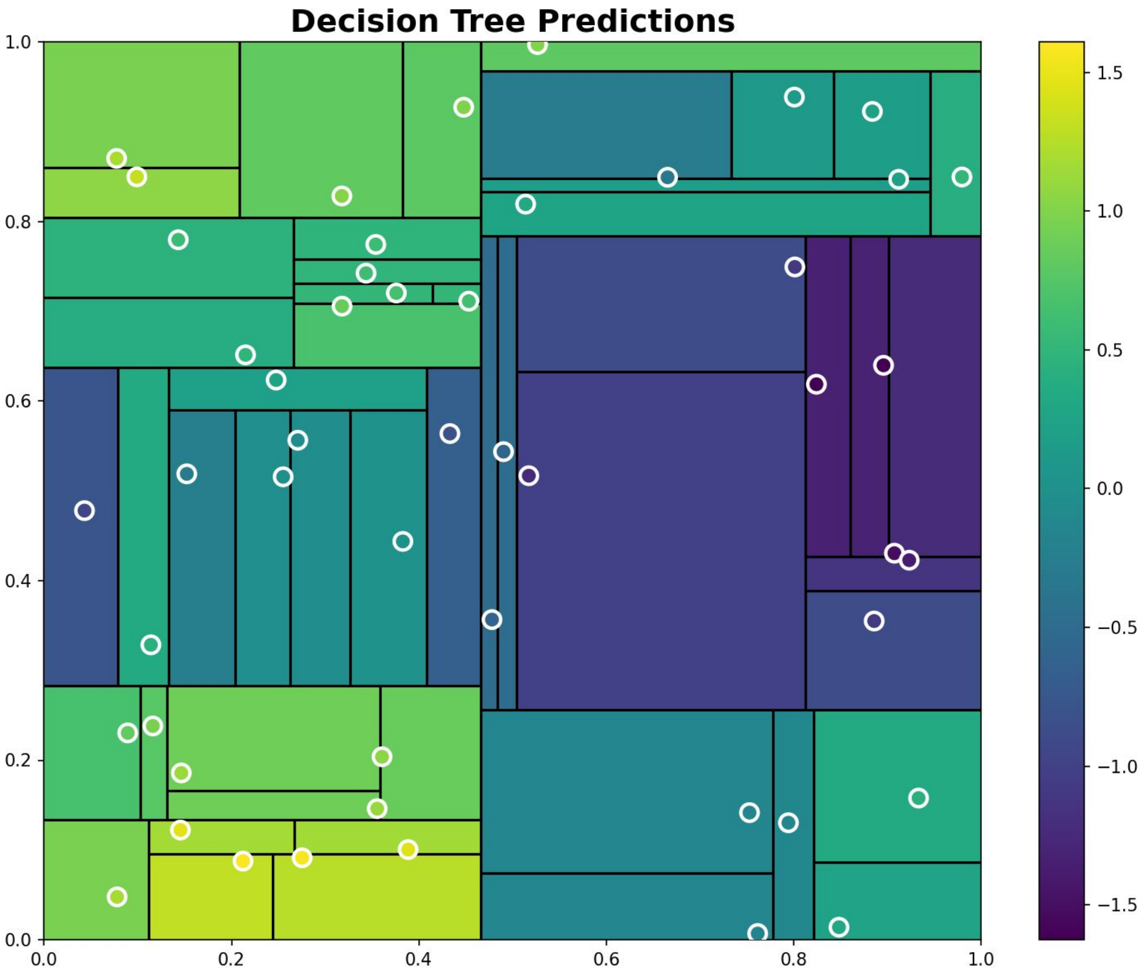
Decision Trees

Big Idea: Similar to K-Nearest Neighbors, in that it divides up the space of predictors and makes predictions by averaging the training data values within each region.

Decision Trees

Big Idea: Similar to K-Nearest Neighbors, in that it divides up the space of predictors and makes predictions by averaging the training data values within each region.

But, instead of using distances to divide up the region, we are only allowed to chop up the space using lines/planes/hyperplanes parallel to the axes.



Decision Tree Predictions
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Decision Trees

Big Idea: Similar to K-Nearest Neighbors, in that it divides up the space of predictors and makes predictions by averaging the training data values within each region.

But, instead of using distances to divide up the region, we are only allowed to chop up the space using lines parallel to the axes.

This is equivalent to making predictions by using a sequence of binary questions about the predictors. (More on this in the notebook).