Model Generalization

Dr. Mahmoud N Mahmoud
mnmahmoud@ncat.edu

North Carolina A & T State University

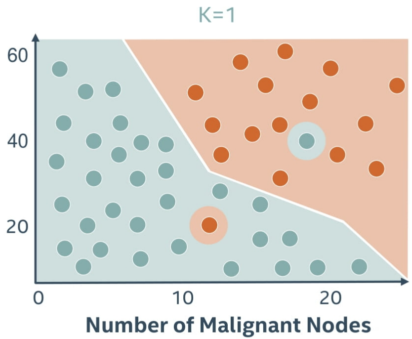
September 4, 2020

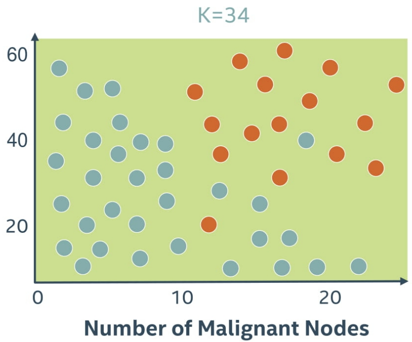
Talk Overview

1. Model Generalization
2. Introduction to Linear Regression
3. Advanced Linear Regression

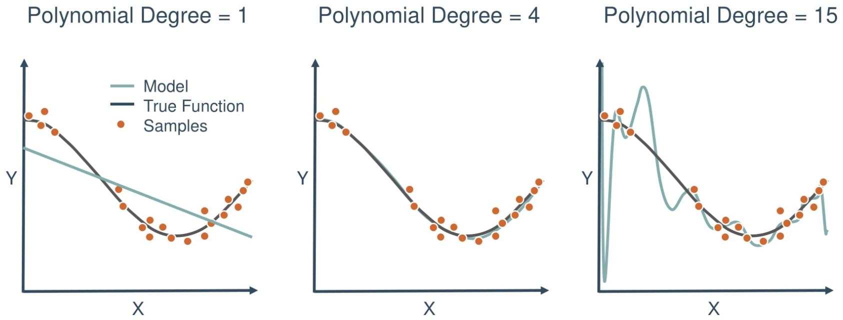
Outline

K VALUE AFFECTS DECISION BOUNDARY

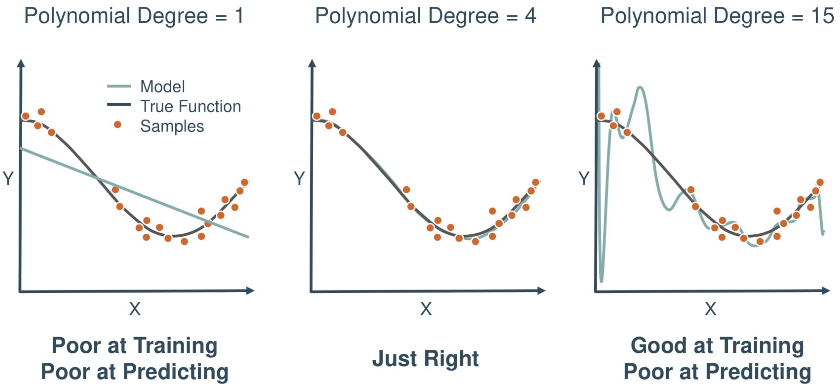




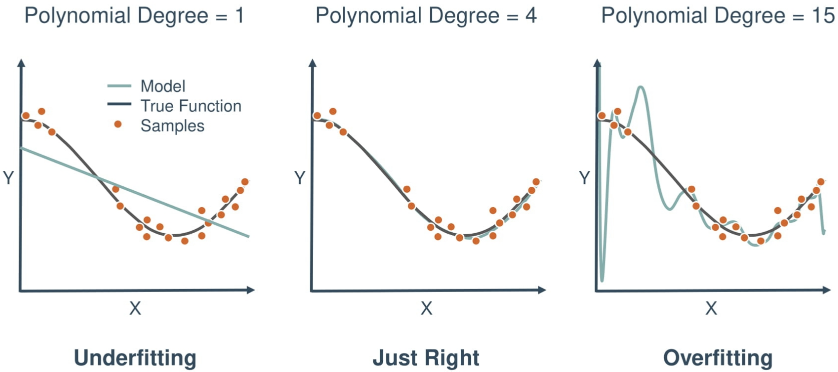
CHOOSING BETWEEN DIFFERENT COMPLEXITIES



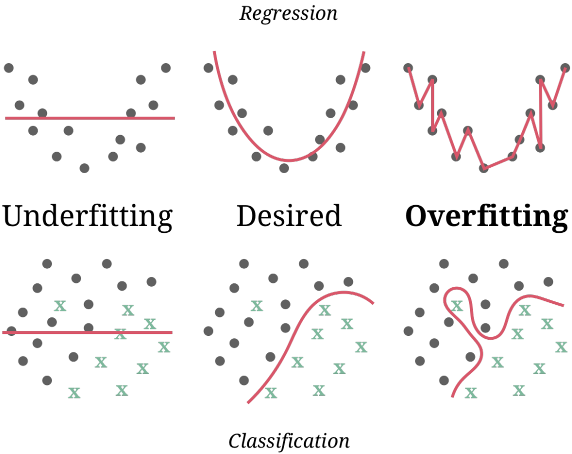
HOW WELL DOES THE MODEL GENERALIZE?



UNDERFITTING VS OVERFITTING



Overfitting and Underfitting (1/2)



Note

A good model (best fit) should be able to generalize to new (unseen) data. How?

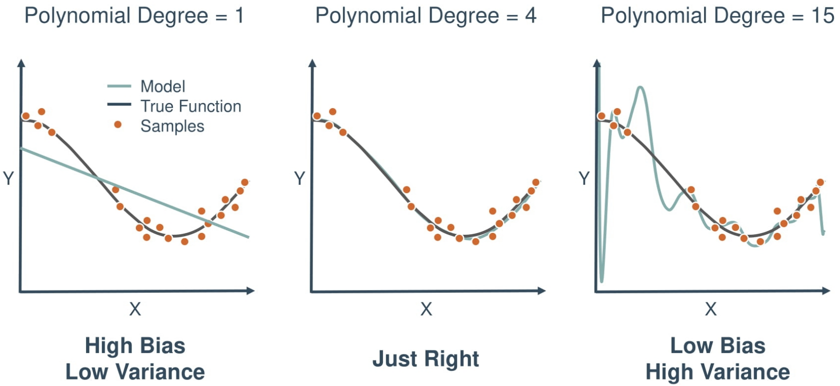
Overfitting and Underfitting (2/2)

- **Over-fitting:**
  - Model too complex (flexible)
  - Fits noise in the training data
  - High error is expected on the test data.
- **Under-fitting:**
  - Model too simplistic (too rigid)
  - Not powerful enough to capture salient patterns in training data and test data.

Note

A good model (best fit) should be able to generalize to new (unseen) data. How?

BIAS—VARIANCE TRADEOFF



What is Bias?

- Bias is the difference between the Predicted Value and the Expected Value of our training data.

What is Bias?

- Bias is the difference between the Predicted Value and the Expected Value of our training data.
- Weak models have High Bias as the error on the training set is expected to be high.

What is Bias?

- Bias is the difference between the Predicted Value and the Expected Value of our training data.
- Weak models have High Bias as the error on the training set is expected to be high.
- High bias means "Underfitting" and Low Bias means "Overfitting".

What is Bias?

- Bias is the difference between the Predicted Value and the Expected Value of our training data.
- Weak models have High Bias as the error on the training set is expected to be high.
- High bias means "Underfitting" and Low Bias means "Overfitting".
- In KNN, Describe the bias for k=1 and k=N

What is Variance?

- Variance is the difference between the Predicted Value and the Expected Value for future datasets.

What is Variance?

- Variance is the difference between the Predicted Value and the Expected Value for future datasets.
- Strong models in training phase have High Variance as they are not flexible for slight future changes.

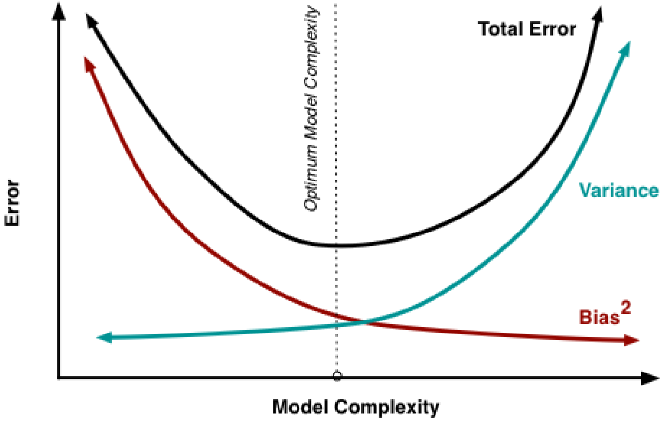
What is Variance?

- Variance is the difference between the Predicted Value and the Expected Value for future datasets.
- Strong models in training phase have High Variance as they are not flexible for slight future changes.
- High Variance means "Overfitting" and Low Variance means "Underfitting".

What is Variance?

- Variance is the difference between the Predicted Value and the Expected Value for future datasets.
- Strong models in training phase have High Variance as they are not flexible for slight future changes.
- High Variance means "Overfitting" and Low Variance means "Underfitting".
- In KNN, Describe the Variance for k=1 and k=N

Bias vs Variance Tradeoff



TRAINING AND TEST SPLITS

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

TRAINING AND TEST SPLITS

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

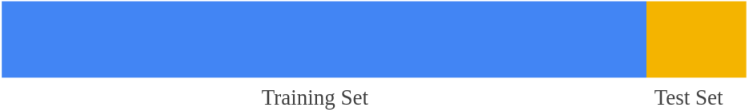
TRAINING DATA

TEST DATA

Training and Test Sets

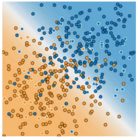
To measure how our model generalize, we split our data to

- **Training set** a subset to train a model.
- **Test set** a subset to evaluate the trained model Estimeate Generalization.

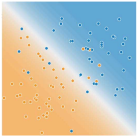


The test should:

- be large enough to yield statistically meaningful results.
- be representative of the data set as a whole.



Training Data



Test Data

USING TRAINING AND TEST DATA

TRAINING DATA

fit the model

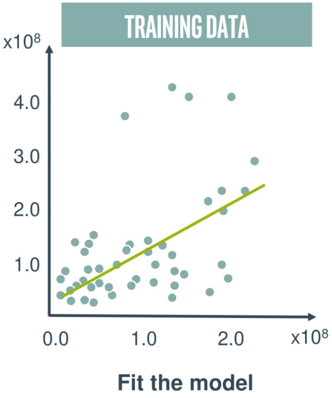
TEST DATA

measure performance

- predict label with model
- compare with actual value
- measure error

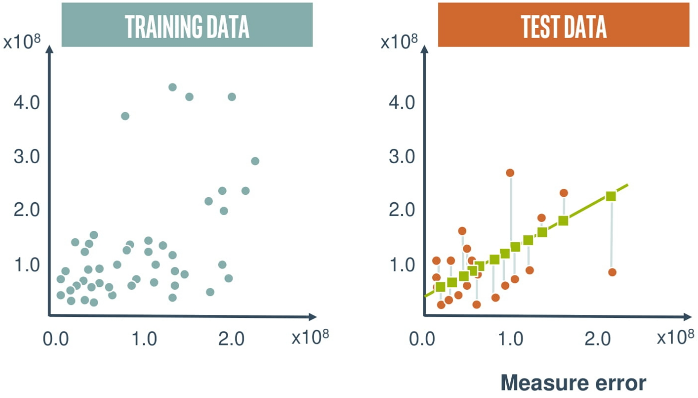
USING TRAINING AND TEST DATA

USING TRAINING AND TEST DATA

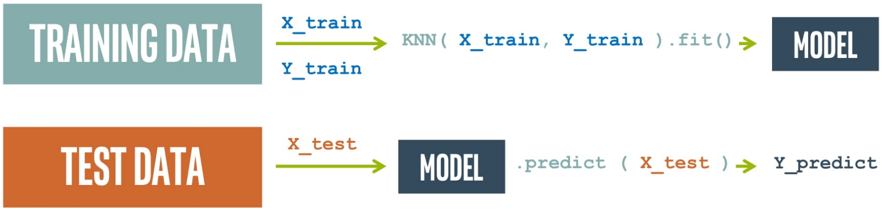


USING TRAINING AND TEST DATA

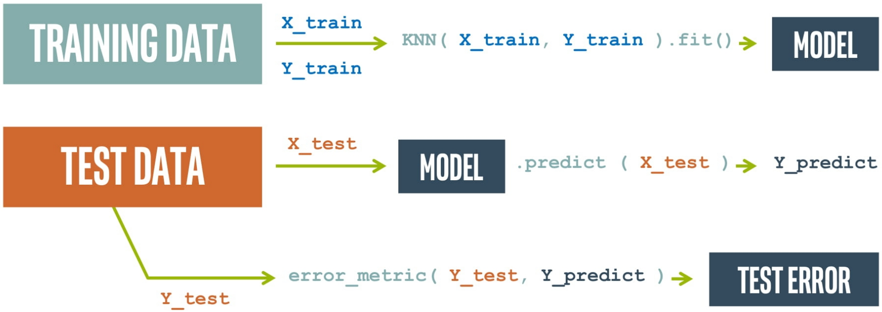
USING TRAINING AND TEST DATA



FITTING TRAINING AND TEST DATA



FITTING TRAINING AND TEST DATA



TRAIN AND TEST SPLITTING: THE SYNTAX

Import the train and test split function

```
from sklearn.model_selection import train_test_split
```

TRAIN AND TEST SPLITTING: THE SYNTAX

Import the train and test split function

```
from sklearn.model_selection import train_test_split
```

Split the data and put 30% into the test set

```
train, test = train_test_split(data, test_size=0.3)
```

TRAIN AND TEST SPLITTING: THE SYNTAX

Import the train and test split function

```
from sklearn.model_selection import train_test_split
```

Split the data and put 30% into the test set

```
train, test = train_test_split(data, test_size=0.3)
```

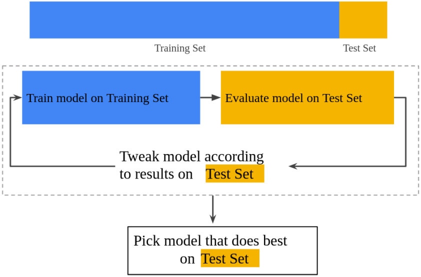
Other method for splitting data:

```
from sklearn.model_selection import ShuffleSplit
```

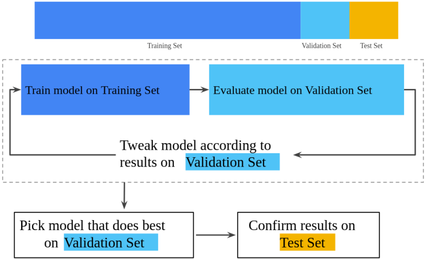
Beyond Test Set: Validation Set

What if we have several model to compare and pick only one?

- Adding or removing features
- Trying different model complexities (linear, quadratic, etc)
- ...



More chances to Overfit.



Less chances to Overfit.

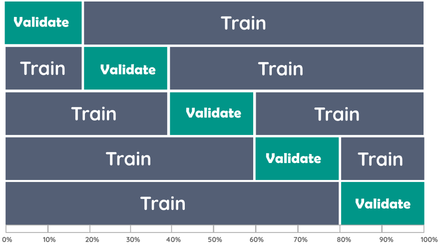
K-Cross Validation

Why?

- We can be exposed to the test set only once.
- We need to estimate future error as accurately as possible.

Ex.

- Randomly split the training into k sets.
- Validate on one in each turn (train on 4 others)
- Average the results over 5 folds



5-fold cross validation

BEYOND A SINGLE TEST SET: CROSS VALIDATION

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

TRAINING DATA 1

VALIDATION DATA 1

BEYOND A SINGLE TEST SET: CROSS VALIDATION

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

TRAINING DATA 2

VALIDATION DATA 2

BEYOND A SINGLE TEST SET: CROSS VALIDATION

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

VALIDATION DATA 3

TRAINING DATA 3

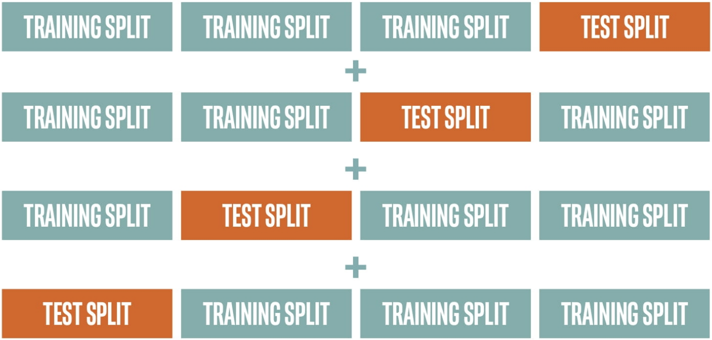
BEYOND A SINGLE TEST SET: CROSS VALIDATION

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

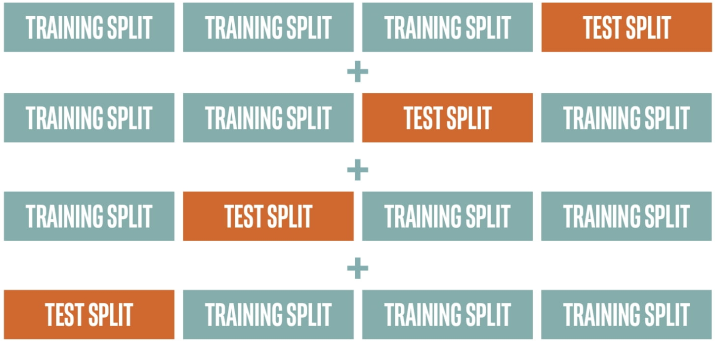
VALIDATION DATA 4

TRAINING DATA 4

BEYOND A SINGLE TEST SET: CROSS VALIDATION

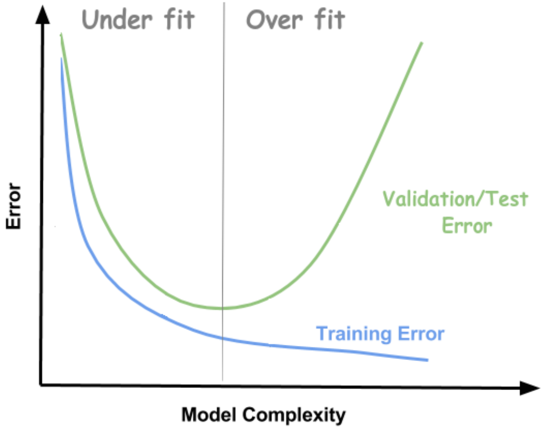


BEYOND A SINGLE TEST SET: CROSS VALIDATION



Average cross validation results

Training vs. Generalization Error (3/3)



Training vs. Generalization Error (1/3)

Training Error: It measures how we are performing on the training set (same as loss).

$$E_{train} = \frac{1}{|D_{train}|} \sum_{(\mathbf{x},y)\in D_{train}} error(f(\mathbf{x}), y)$$

Generalization Error:

- How well we will do on any kind future data from the same distribution.

$$E_{gen} = \int_{(\mathbf{x},y)\in D} error(f(\mathbf{x}), y) \underbrace{p(\mathbf{x}, y)}_{\text{How often we see (x,y) pair}} d\mathbf{x}$$

Can never compute generalization error practically

Training vs. Generalization Error (2/3)

Test Error:

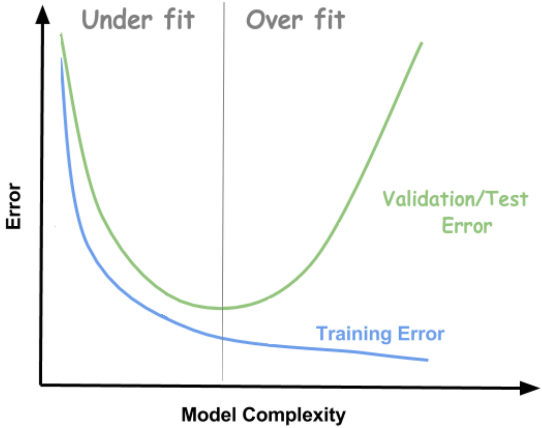
- Introduced to estimate the generalization error.
- That is why we should be exposed to test set only once.

$$E_{test} = \frac{1}{|D_{test}|} \sum_{(\mathbf{x},y)\in D_{test}} error(f(\mathbf{x}), y)$$

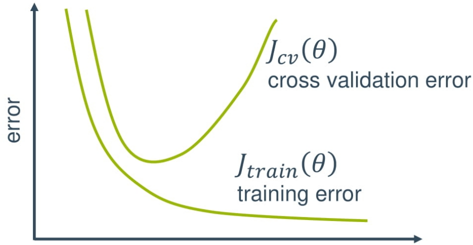
- How close E_{gen} to E_{test}? depends on $|D_{test}|$.

$$\lim_{|D_{test}|\to\infty} E_{test} \approx E_{gen}$$

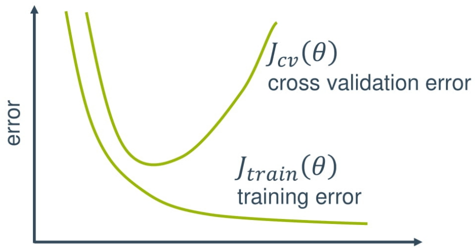
Training vs. Generalization Error (3/3)



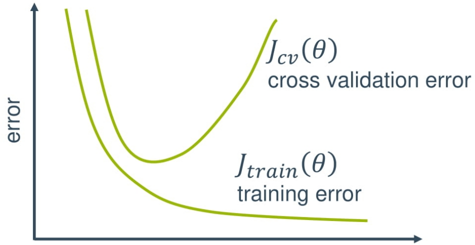
MODEL COMPLEXITY VS ERROR



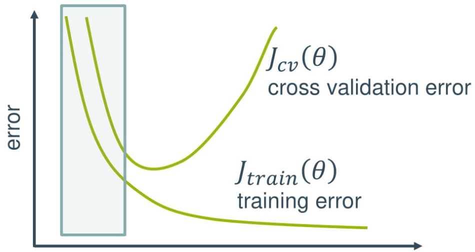
MODEL COMPLEXITY VS ERROR

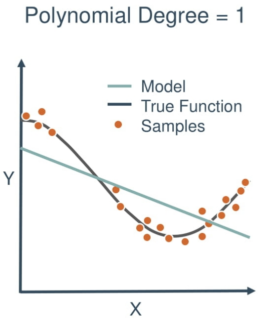


MODEL COMPLEXITY VS ERROR



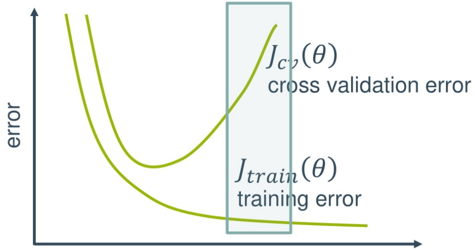
MODEL COMPLEXITY VS ERROR

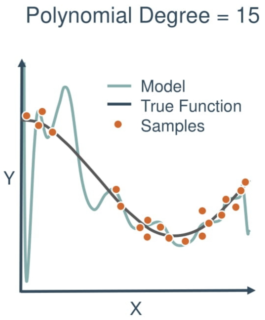




Underfitting: training and cross validation error are high

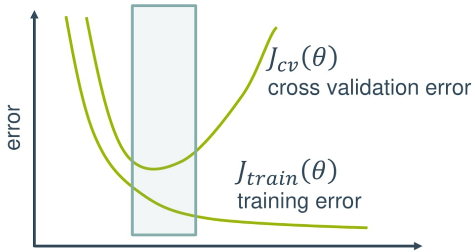
MODEL COMPLEXITY VS ERROR

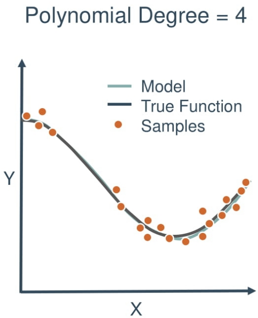




Overfitting: training error is low, cross validation is high

MODEL COMPLEXITY VS ERROR





Just right: training and cross validation errors are low

CROSS VALIDATION: THE SYNTAX

Import the train and test split function

```
from sklearn.model_selection import cross_val_score
```

CROSS VALIDATION: THE SYNTAX

Import the train and test split function

```
from sklearn.model_selection import cross_val_score
```

Perform cross-validation with a given model

```
cross_val = cross_val_score(KNN, X_data, y_data, cv=4,
                      scoring='neg_mean_squared_error')
```

CROSS VALIDATION: THE SYNTAX

Import the train and test split function

```
from sklearn.model_selection import cross_val_score
```

Perform cross-validation with a given model

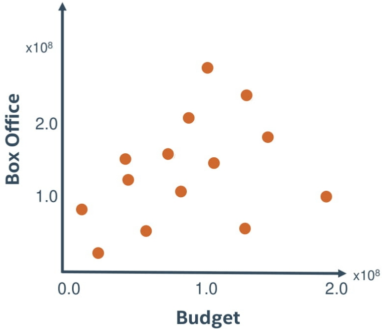
```
cross_val = cross_val_score(KNN, X_data, y_data, cv=4,
                      scoring='neg_mean_squared_error')
```

Other methods for cross validation:

```
from sklearn.model_selection import KFold, StratifiedKFold
```

Outline

INTRODUCTION TO LINEAR REGRESSION



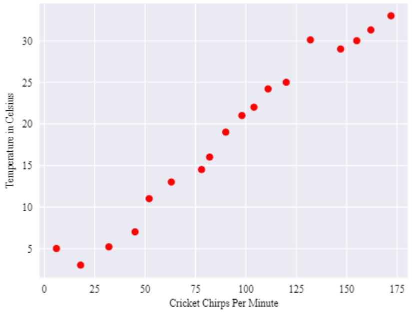
$$y_\beta(x) = \beta_0 + \beta_1 x$$

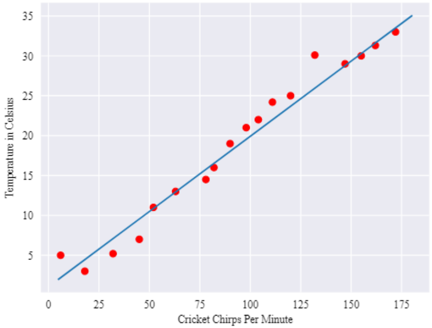
Linear Regression

Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data

Example: Scientists found that crickets (an insect species) chirp more frequently on hotter days than on cooler days.





Linear Regression

A linear relationship

- True, the line doesn't pass through every dot.
- However, the line does clearly show the relationship between chirps and temperature.

$$y = mx + b$$

where:

- y: is the temperature in Celsiusthe value we're trying to predict.
- m: is the slope of the line.
- x: is the number of chirps per minutethe value of our input feature.
- b: is the y-intercept.

Linear Regression

In machine learning, we'll write the equation for a model slightly differently:

$$y' = w_1 x_1 + w_0$$

where:

- y': is the predicted label (a desired output).
- w_1: is the weight of feature 1. Weight is the same concept as the "slope".
- x_1: is feature 1.
- w_0 or b: is the bias (the y-intercept).

Notethat

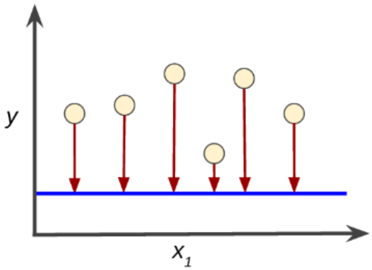
A model that relies on three features might look as follows:

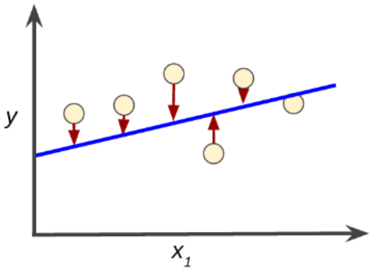
$$y' = w_3 x_3 + w_2 x_2 + w_1 x_1 + w_0$$

Training and Loss

- **Training** a model simply means learning (determining) good values for all the weights and the bias from labeled examples.
- **Loss** is the penalty for a bad prediction.
  - Perfect prediction means the loss is zero
  - Bad model have large loss.
- Suppose we selected the following weights and biases.

Which of them have lower loss?





Squared loss

- The linear regression models use a popular loss function called squared loss.
- Also known as L_2.
- Is represented as follows:

$$[obsevation(x) - prediction(x)]^2$$
$$= (y - y')^2$$

Squared loss

- The linear regression models use a popular loss function called squared loss.
- Also known as L_2.
- Is represented as follows:

$$[obsevation(x) - prediction(x)]^2$$
$$= (y - y')^2$$

Why squared loss?

Squared loss

- The linear regression models use a popular loss function called squared loss.
- Also known as L_2.
- Is represented as follows:

$$[obsevation(x) - prediction(x)]^2$$
$$= (y - y')^2$$

Why squared loss?
Can we do absolute loss?

Mean square error (MSE)

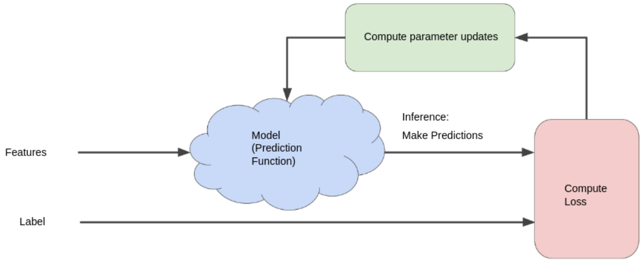
- Is the average squared loss per example over the whole dataset.

$$\text{MSE} = \frac{1}{N} \sum_{(x,y) \in D} (y - prediction(x))^2$$

- (x,y) is an example in which
  - y is the label
  - x is a feature
- prediction(x) is equal $y' = w_1 x + w_0$
- D is the dataset that contains all (x,y) pairs
- N is the number of samples in D

Reducing Loss

- **Training** is a feedback process that use the loss function to improve the model parameters.
- The **training** is an iterative process.



Two Questions

- What initial values should we set for w_1 and w_0?
- How to update w_1 and w_0?

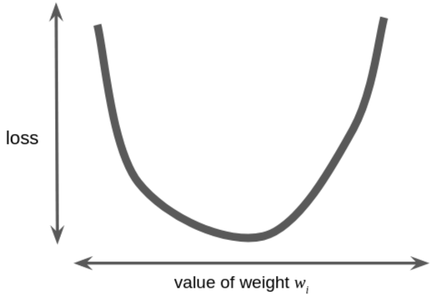
Gradient Descent (1/3)

- Assume (for symplicity) we are only concerned with finding w_1.
- Assume we had the time and the computing resources to calculate the loss for all possible values of w_1.

Gradient Descent (1/3)

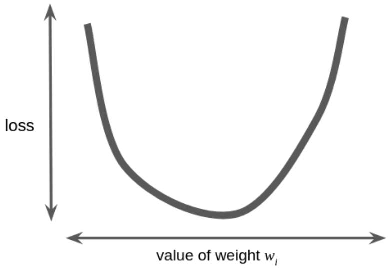
- Assume (for symplicity) we are only concerned with finding w_1.
- Assume we had the time and the computing resources to calculate the loss for all possible values of w_1.

Regression problems yield convex loss vs. weight plots.



Gradient Descent (2/3)

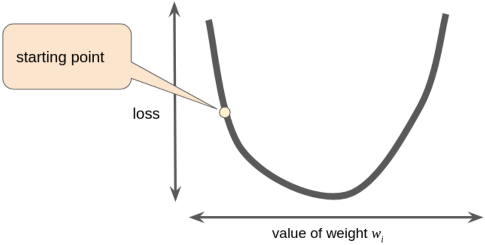
- Gradient descent enables you to find the optimal w without computing for all possible values.
- Gradient descent has the following steps
  1. Pick a random starting point for w
  2. Calculates the gradient of the loss curve at w.
  3. Update w
  4. go to 2, till convergence



Gradient Descent (3/3)

Note that a gradient is a vector, so it has both of the following characteristics:

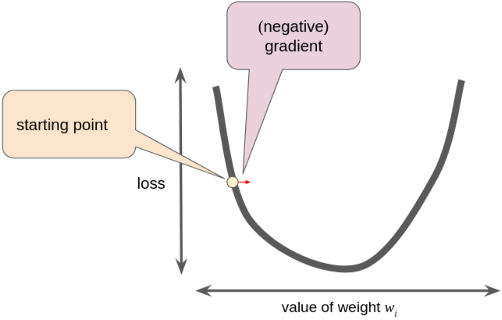
- Magnitude
- Direction



$$w_{new} = w_{old} - \eta * \frac{d\ \ loss}{dw}$$

Gradient Descent (3/3)

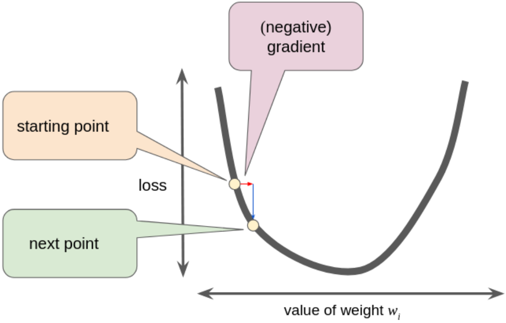
The gradient descent algorithm takes a step in the direction of the negative gradient



$$w_{new} = w_{old} - \eta * \frac{d\ \ loss}{dw}$$

Gradient Descent (3/3)

the gradient descent algorithm adds **some fraction** of the gradient's magnitude **(Learning Rate η)** to the previous point



$$w_{new} = w_{old} - \eta * \frac{d\ loss}{dw}$$

Convergence Criteria

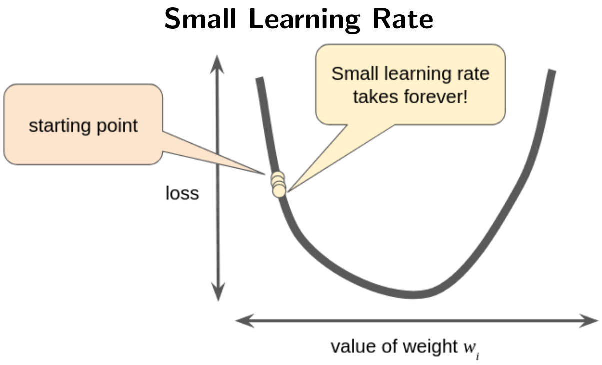
- For convex functions, optimum occurs when
  - $|\frac{d \ \ loss}{dw}| = 0$
- In practice, stop when
  - $|\frac{d \ \ loss}{dw}| \leq \epsilon$

Learning rate

- Gradient descent algorithms multiply the gradient by a scalar known as the learning rate (also sometimes called step size) .
- How can we choose the learning rate?

Learning rate

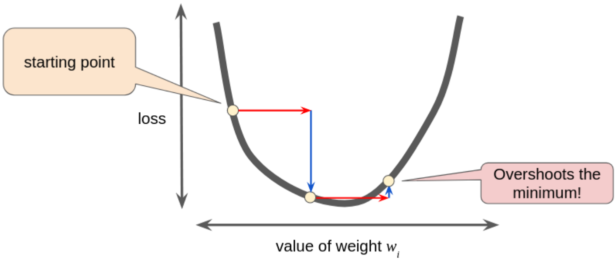
- Gradient descent algorithms multiply the gradient by a scalar known as the learning rate (also sometimes called step size) .
- How can we choose the learning rate?



Learning rate

- Gradient descent algorithms multiply the gradient by a scalar known as the learning rate (also sometimes called step size) .
- How can we choose the learning rate?

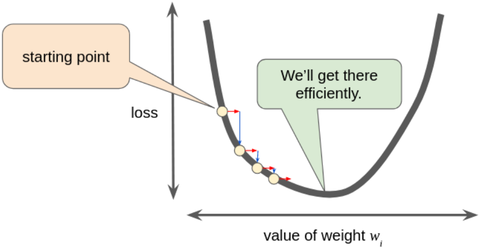
Large Learning Rate



Learning rate

- Gradient descent algorithms multiply the gradient by a scalar known as the learning rate (also sometimes called step size) .
- How can we choose the learning rate?

Optimal Learning Rate usually (0.01)



Generalization and Gradient

- For n features: $y' = \sum_{i=0}^{i=n} w_i x_i$
- Note w_0 is the bias (intercept), and $x_0 = 1$.
- vector representation y' $= \mathbf{w}^T \mathbf{x}$
- Loss $= \ell = (y - y')^2$
- Gradient derivation

$$\begin{aligned}\frac{d\ell}{dw_i} &= \frac{d\ell}{dy'}\frac{dy'}{dw_i} \\ &= [2(y - y') * x_i * (-1)]\end{aligned}$$

COMPARING LINEAR REGRESSION AND KNN

LINEAR REGRESSION	K NEAREST NEIGHBORS
Fitting involves minimizing cost function (slow)	**Fitting involves storing training data** (fast)
Model has few parameters (memory efficient)	**Model has many parameters** (memory intensive)

COMPARING LINEAR REGRESSION AND KNN

LINEAR REGRESSION	K NEAREST NEIGHBORS
Fitting involves minimizing cost function (slow)	**Fitting involves storing training data** (fast)
Model has few parameters (memory efficient)	**Model has many parameters** (memory intensive)
Prediction involves calculation (fast)	**Prediction involves finding closest neighbors** (slow)

LINEAR REGRESSION: THE SYNTAX

Import the class containing the regression method

```
from sklearn.linear_model import LinearRegression
```

LINEAR REGRESSION: THE SYNTAX

Import the class containing the regression method

```
from sklearn.linear_model import LinearRegression
```

Create an instance of the class

```
LR = LinearRegression()
```

LINEAR REGRESSION: THE SYNTAX

Import the class containing the regression method

```
from sklearn.linear_model import LinearRegression
```

Create an instance of the class

```
LR = LinearRegression()
```

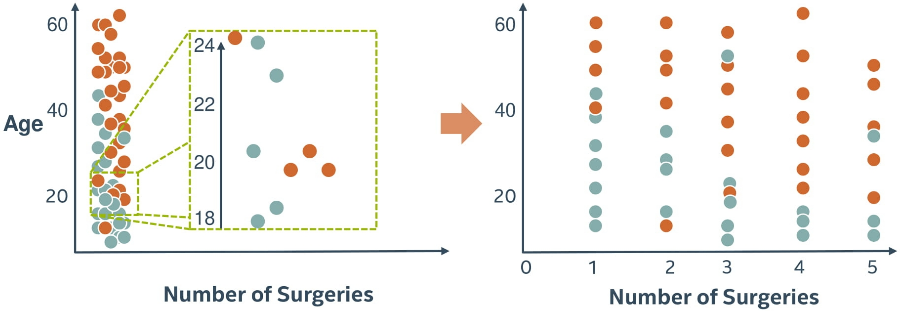
Fit the instance on the data and then predict the expected value

```
LR = LR.fit(X_train, y_train)

y_predict = LR.predict(X_test)
```

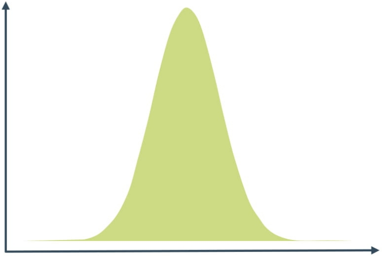
Outline

SCALING IS A TYPE OF FEATURE TRANSFORMATION



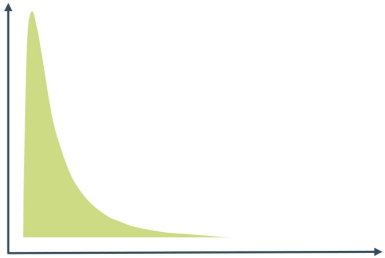
TRANSFORMATION OF DATA DISTRIBUTIONS

- Predictions from linear regression models assume residuals are normally distributed

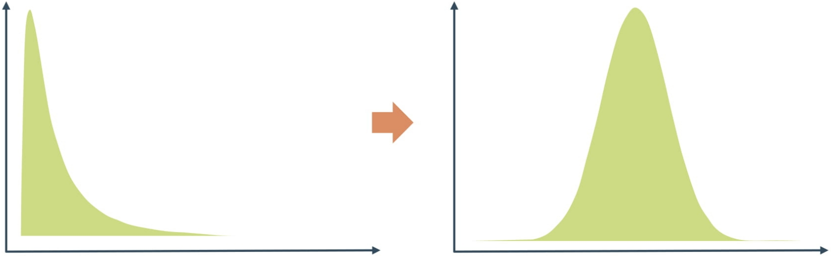


TRANSFORMATION OF DATA DISTRIBUTIONS

- Predictions from linear regression models assume residuals are normally distributed
- Features and predicted data are often skewed



TRANSFORMATION OF DATA DISTRIBUTIONS

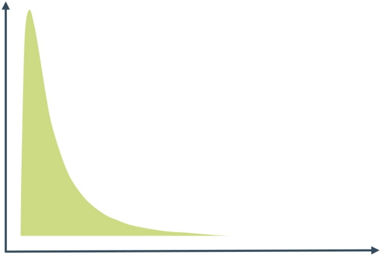


from numpy import log, log1p

from scipy.stats import boxcox

TRANSFORMATION OF DATA DISTRIBUTIONS

- Predictions from linear regression models assume residuals are normally distributed
- Features and predicted data are often skewed
- Data transformations can solve this issue



FEATURE TYPES

FEATURE TYPE	TRANSFORMATION

- **Continuous:** numerical values

FEATURE TYPES

FEATURE TYPE	TRANSFORMATION
Continuous: numerical values	Standard Scaling, Min-Max Scaling

FEATURE TYPES

FEATURE TYPE	TRANSFORMATION
Continuous: numerical values	Standard Scaling, Min-Max Scaling
Nominal: categorical, unordered features (True or False)	One-hot encoding (0, 1)

```
from sklearn.preprocessing import LabelEncoder, LabelBinarizer, OneHotEncoder
```

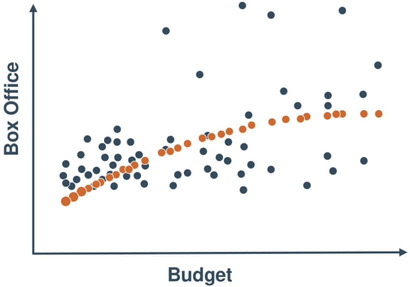
FEATURE TYPES

FEATURE TYPE	TRANSFORMATION
Continuous: numerical values	Standard Scaling, Min-Max Scaling
Nominal: categorical, unordered features (True or False)	One-hot encoding (0, 1)
Ordinal: categorical, ordered features (movie ratings)	Ordinal encoding (0, 1, 2, 3)

```
from sklearn.feature_extraction import DictVectorizer
from pandas import get_dummies
```

ADDITION OF POLYNOMIAL FEATURES

- Capture higher order features of data by adding polynomial features

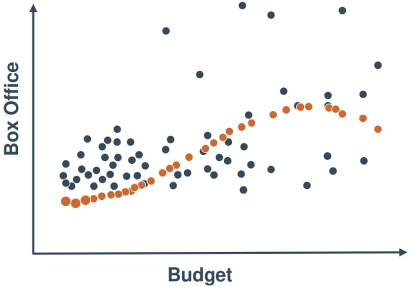


$$y_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

ADDITION OF POLYNOMIAL FEATURES

- Capture higher order features of data by adding polynomial features
- "Linear regression" means linear combinations of features

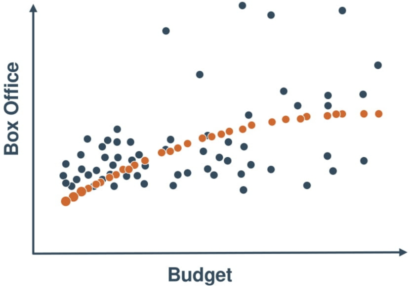
$$y_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



ADDITION OF POLYNOMIAL FEATURES

- Capture higher order features of data by adding polynomial features
- "Linear regression" means linear combinations of features

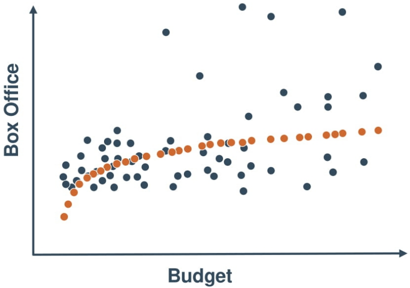
$$y_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$



ADDITION OF POLYNOMIAL FEATURES

- Capture higher order features of data by adding polynomial features
- "Linear regression" means linear combinations of features

$$y_\beta(x) = \beta_0 + \beta_1 \log(x)$$



ADDITION OF POLYNOMIAL FEATURES

- Can also include variable interactions

$$y_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

References

Intel Nervana AI Academy
https://software.intel.com/content/www/us/en/develop/training

Questions