TYPES OF ERRORS

UNDERFITTING

OVERFITTING

UNDERFITTING

- Does not do well in the training set.
- Error due to bias.

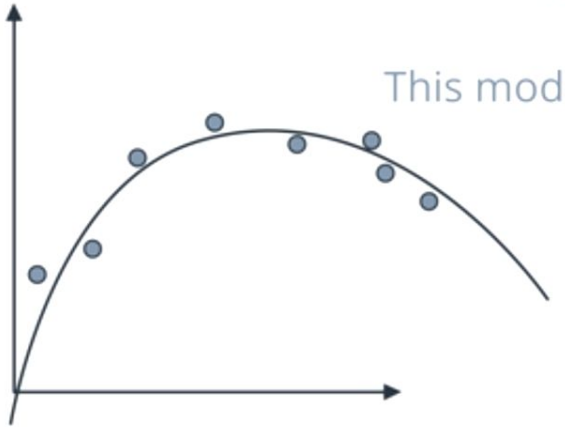
OVERFITTING

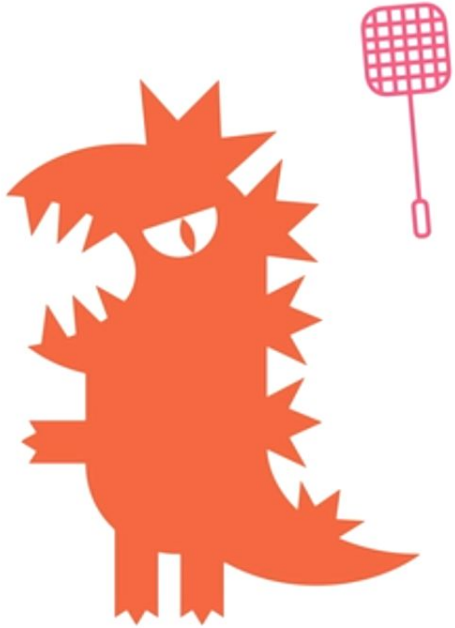
- Does well in the training set, but it tends to memorize it instead of learning the characteristics of it.
- Error due to variance.

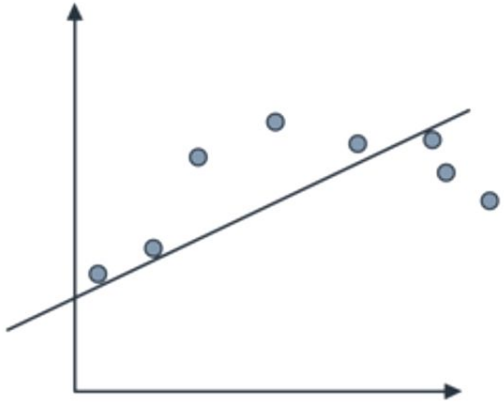
UNDERFITTING

Error due to bias

This model will not do well in the training set



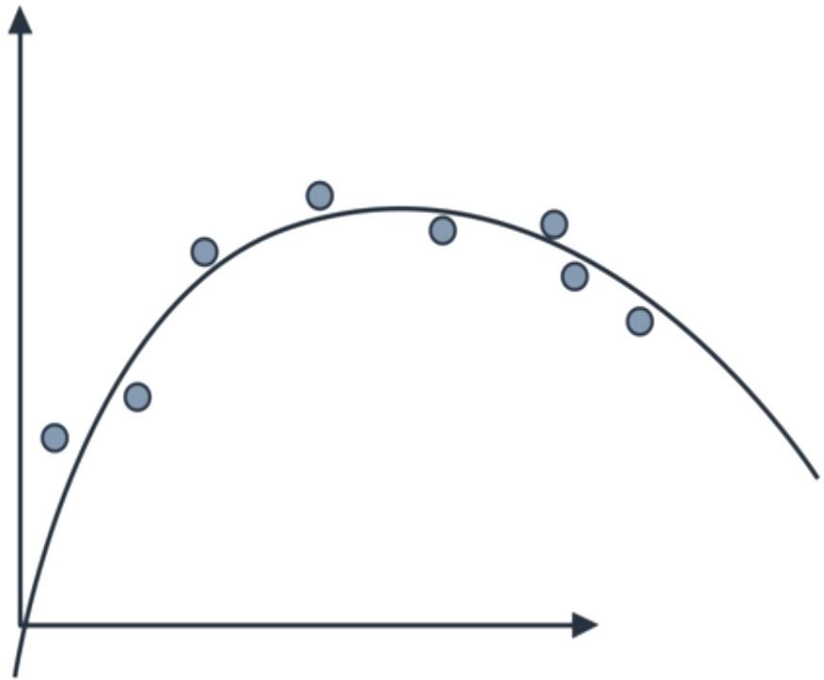


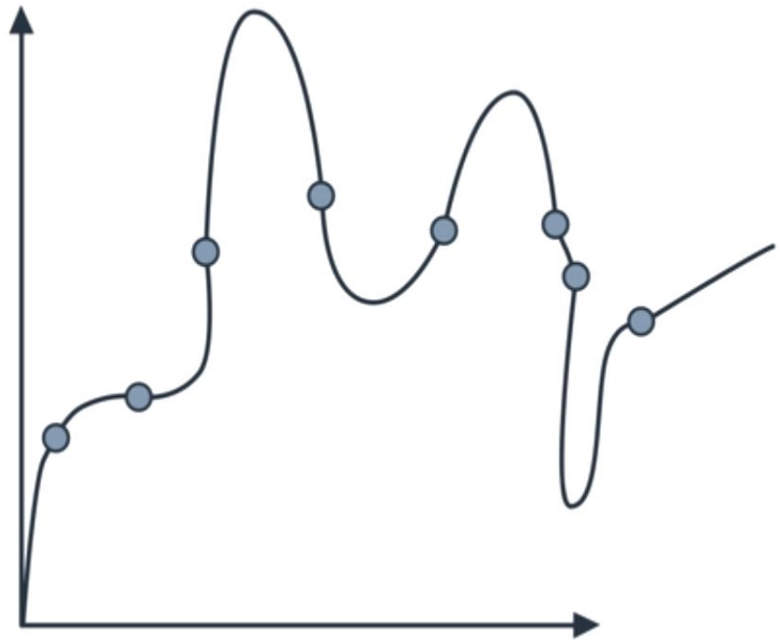


OVERFITTING

Error due to variance

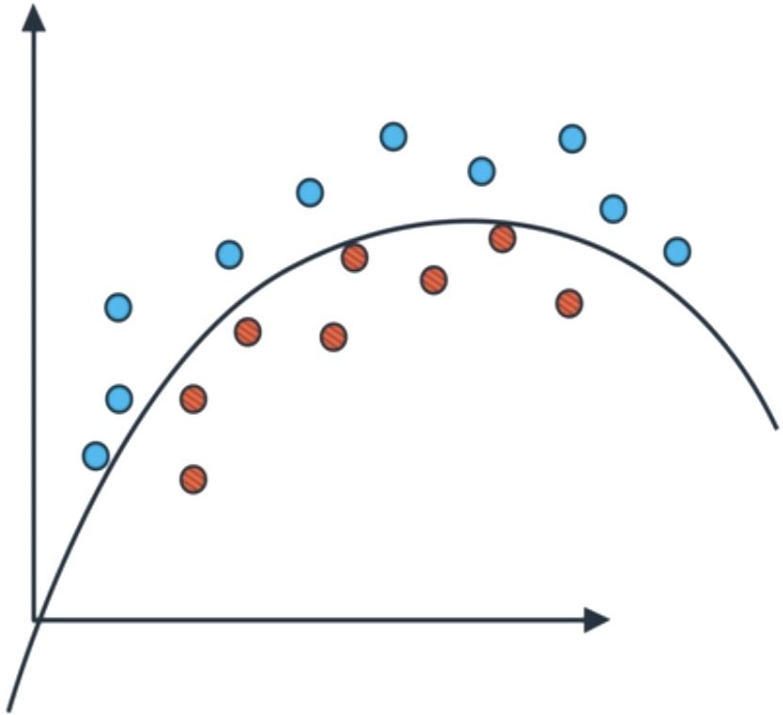
This model performs poorly in the testing set

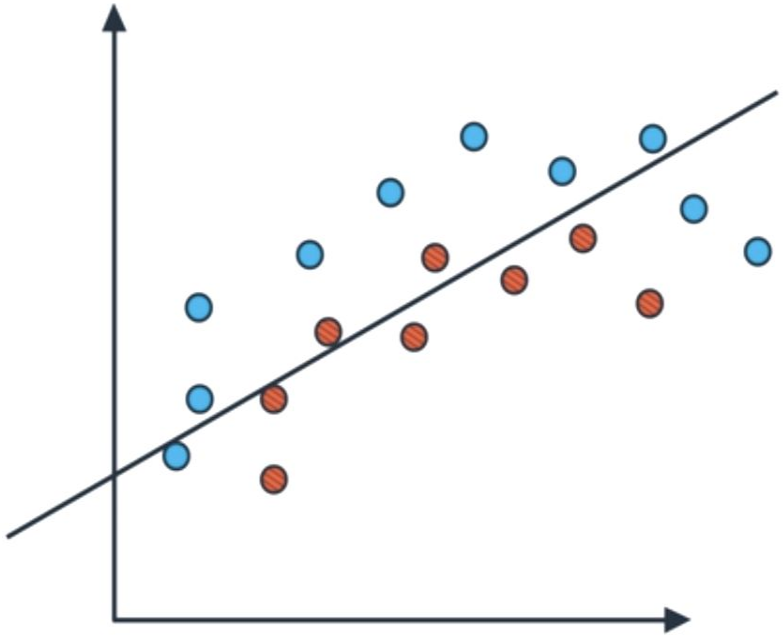




UNDERFITTING

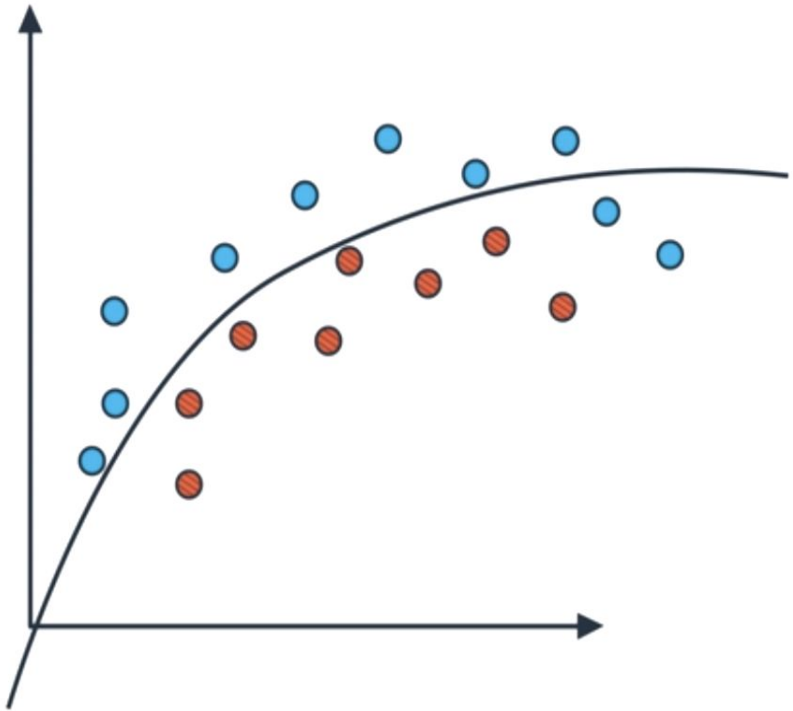
Error due to bias

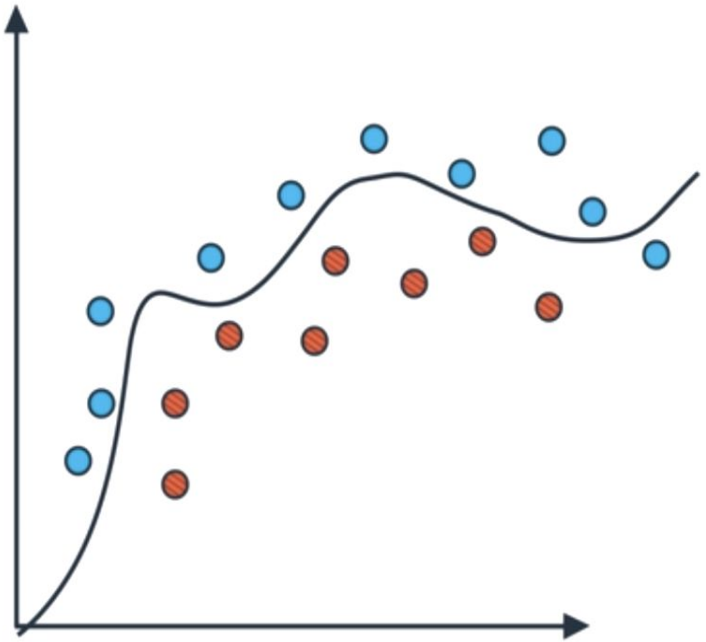




OVERFITTING

Error due to variance





TRADEOFF

High bias (underfitting)

Oversimplify the problem

Bad on training set

Bad on testing set

Good Model

Good model

Good on training set

Good on testing set

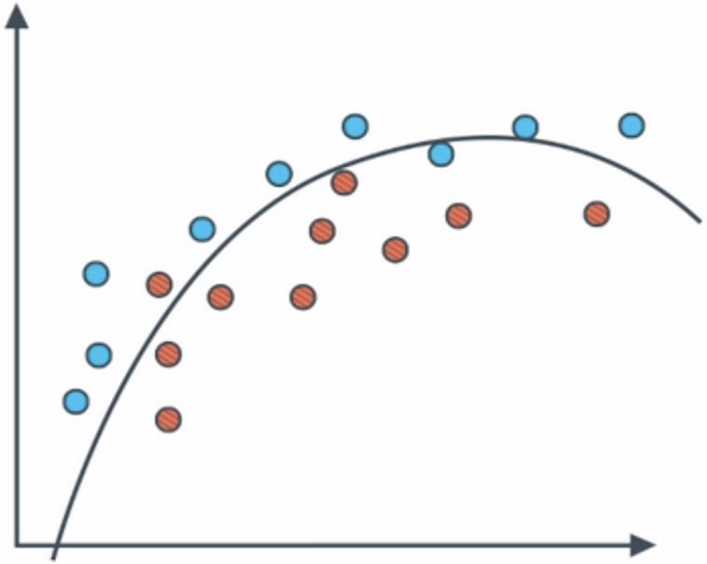
High variance (overfitting)

Overcomplicate the problem

Great on training set

Bad on testing set

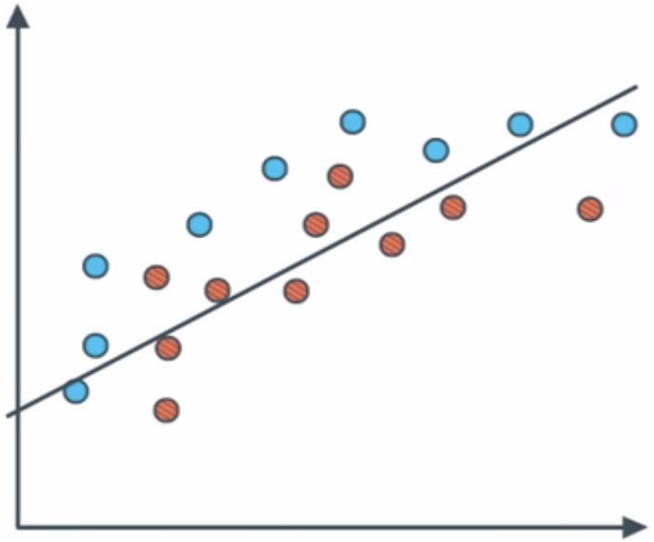
MODEL COMPLEXITY GRAPH



POLYNOMIAL

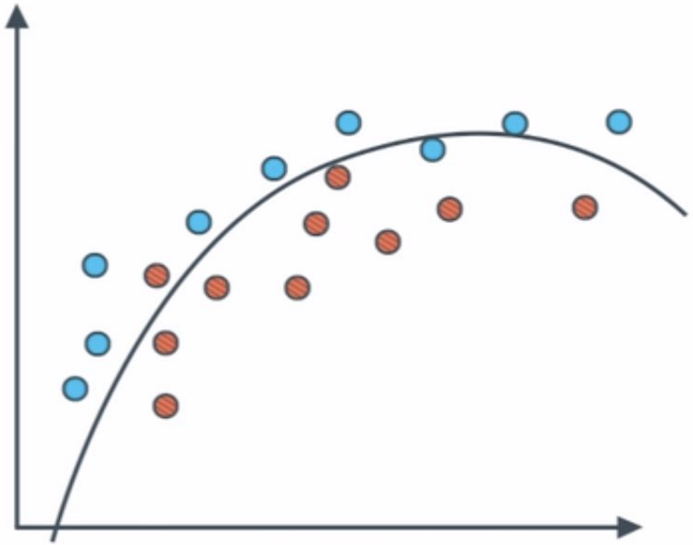
Degree = 2

MODEL COMPLEXITY GRAPH



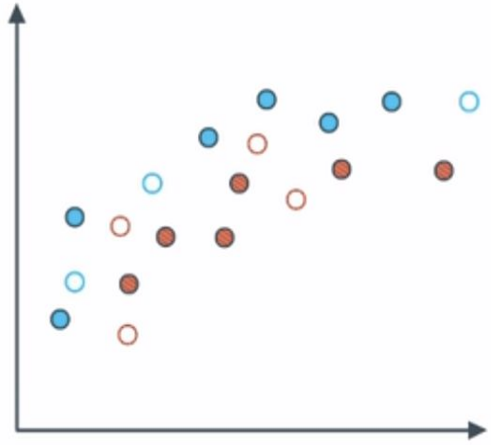
HIGH BIAS

Degree = 1



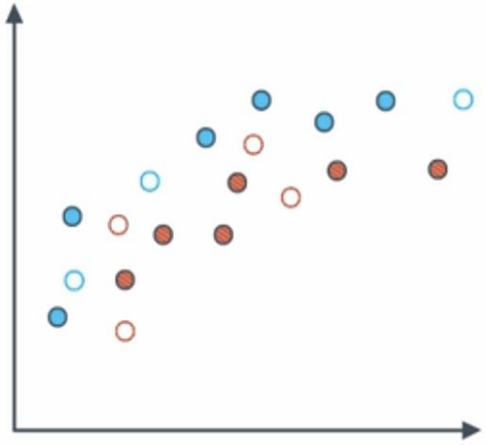
POLYNOMIAL

Degree = 2



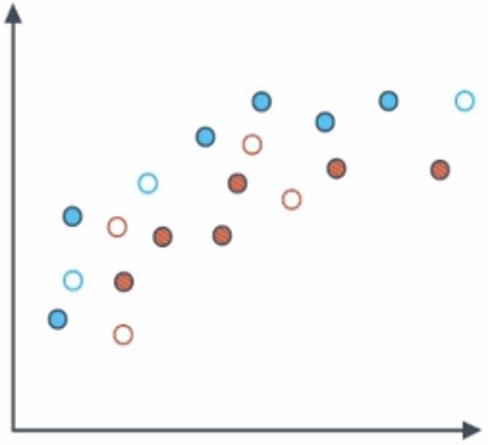
LINEAR MODEL

Degree = 1



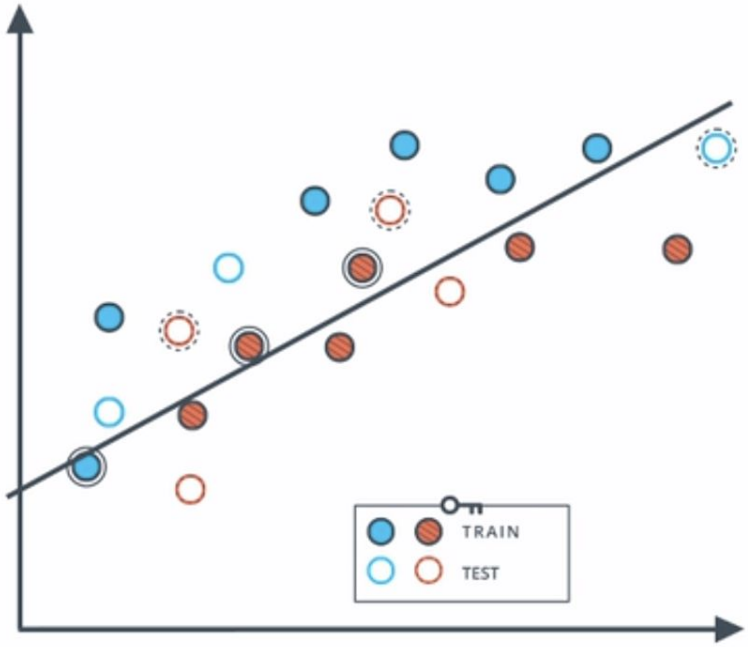
QUADRATIC MODEL

Degree = 2



POLYNOMIAL MODEL

Degree = 6

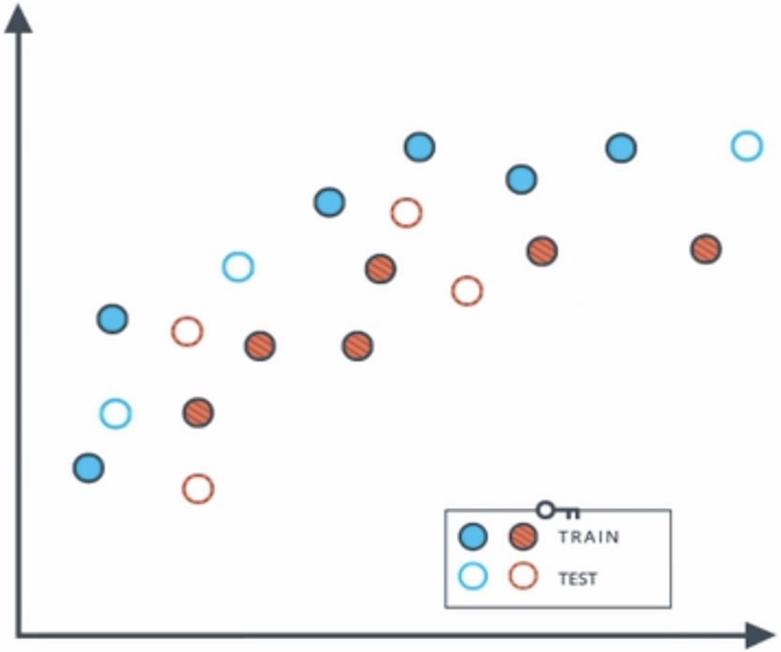


LINEAR MODEL

Degree = 1

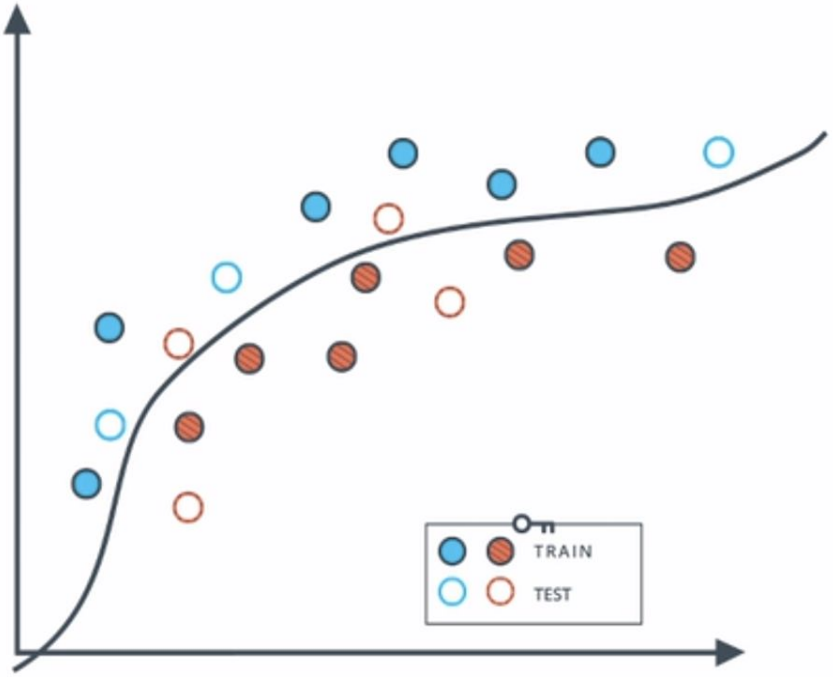
TRAINING ERRORS

TESTING ERRORS



QUADRATIC MODEL

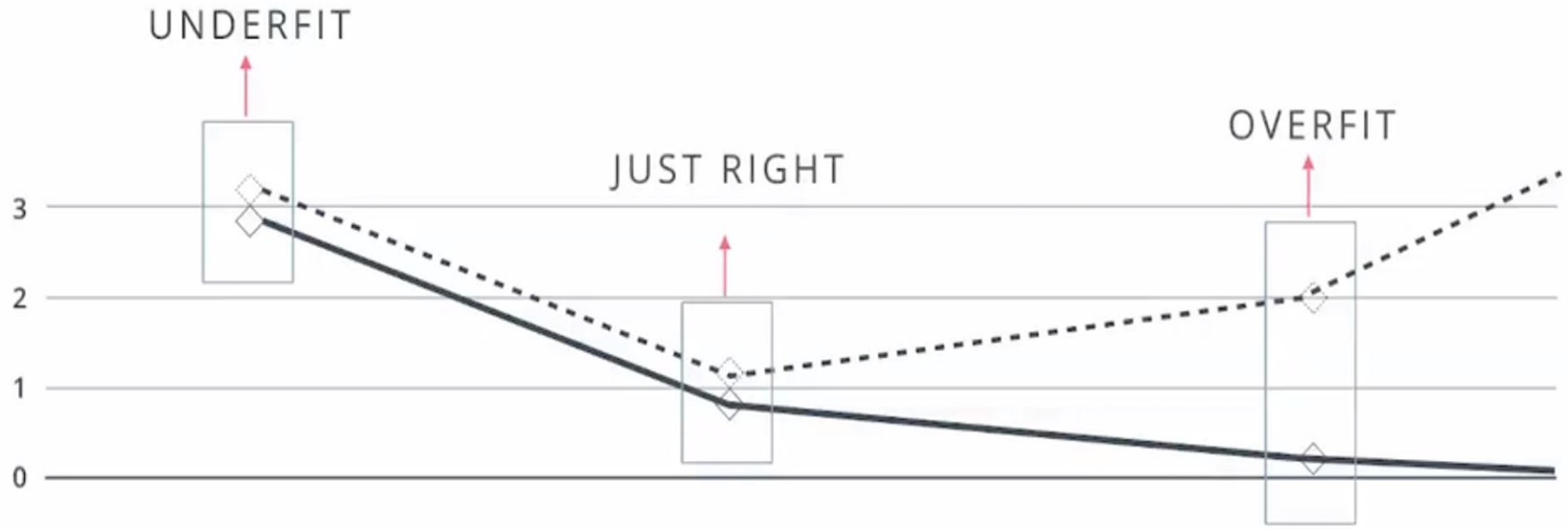
Degree = 2



POLYNOMIAL MODEL

Degree = 6

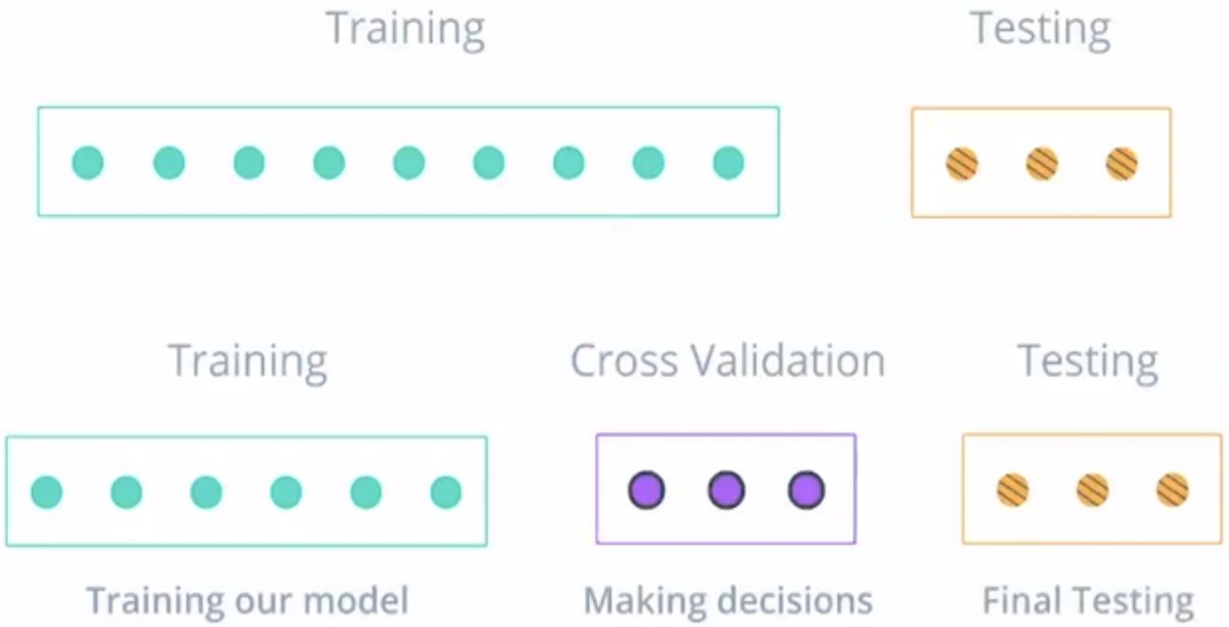
MODEL COMPLEXITY GRAPH

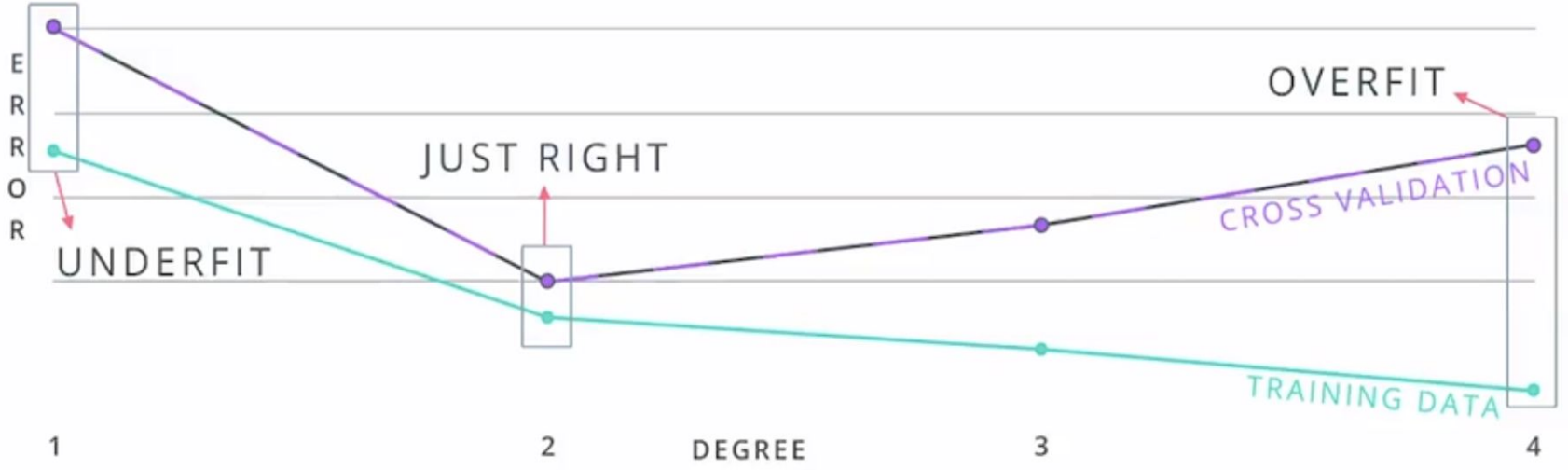


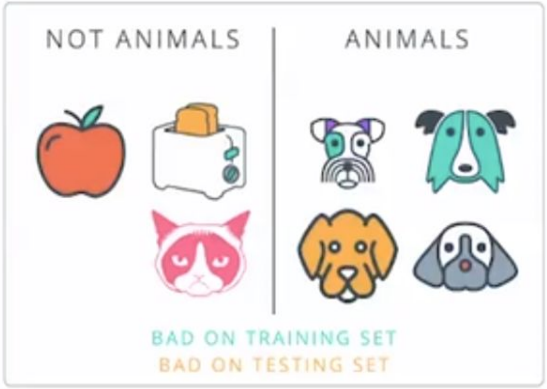
MODEL COMPLEXITY GRAPH

THOU SHALT NEVER
USE YOUR TESTING DATA
FOR TRAINING

SOLUTION: CROSS VALIDATION



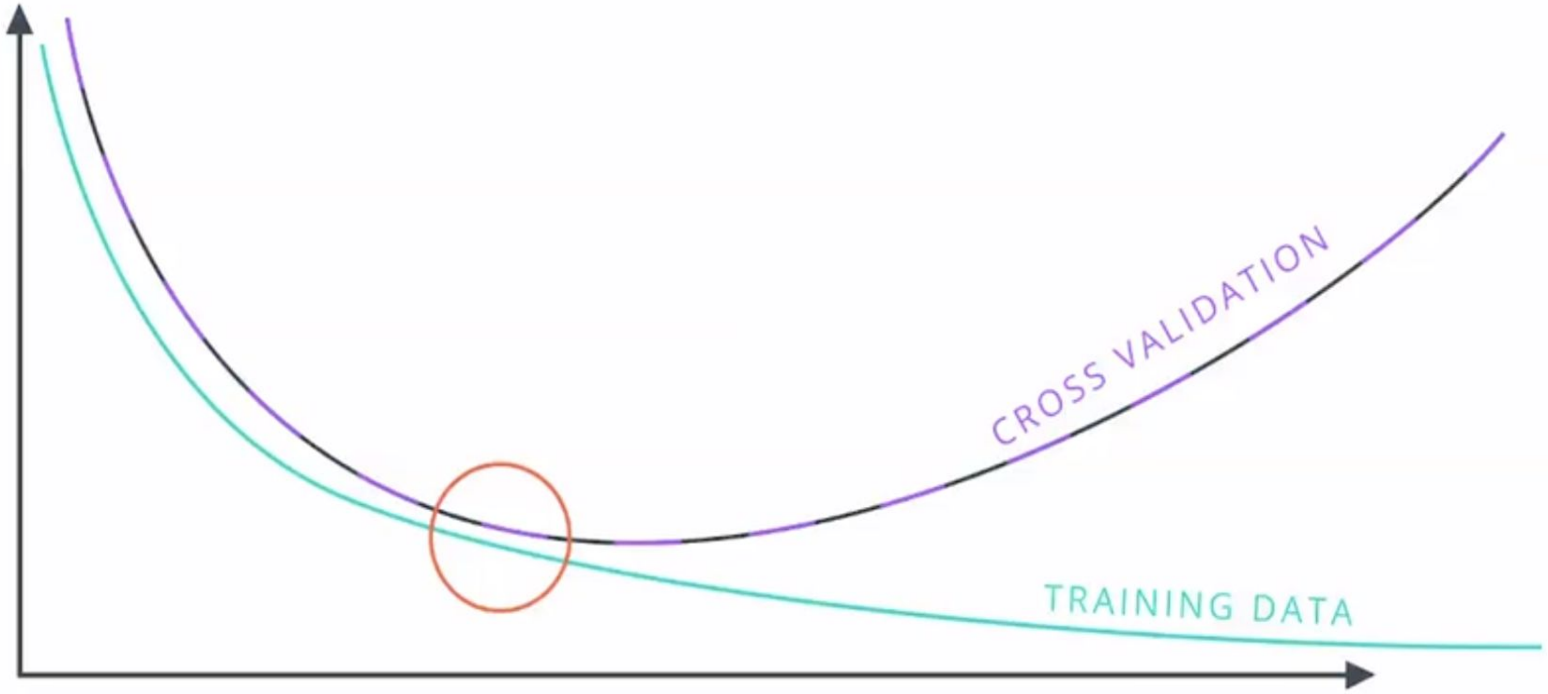

ERROR
OVERFIT
JUST RIGHT
CROSS VALIDATION
UNDERFIT
TRAINING DATA
1
2
DEGREE
3
4


NOT ANIMALS
ANIMALS
BAD ON TRAINING SET
BAD ON TESTING SET


NOT DOGS
DOGS
GOOD ON TRAINING SET
GOOD ON TESTING SET

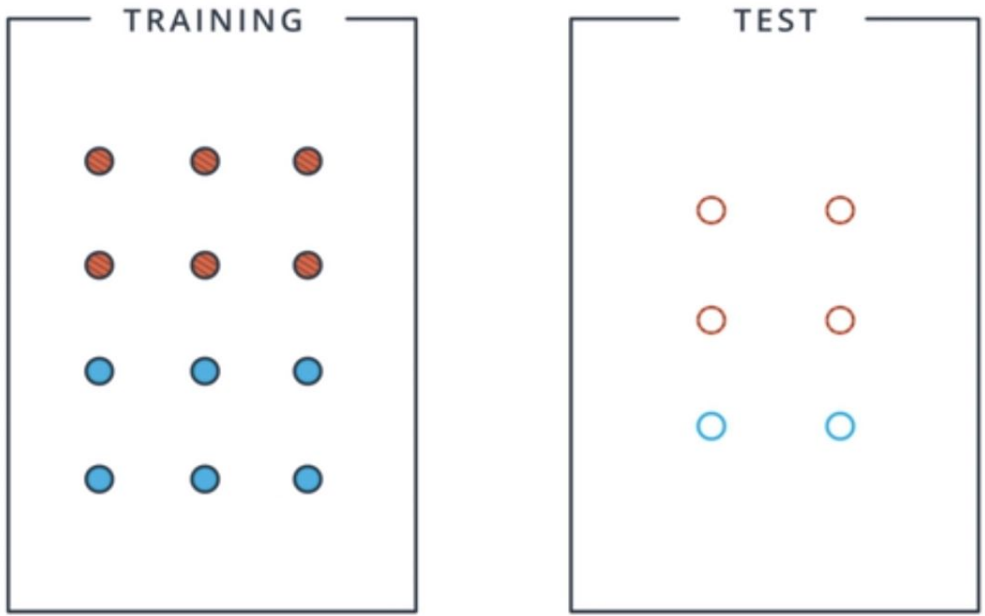

NOT DOGS
who wag their tail
DOGS
who wag their tail
GREAT ON TRAINING SET
BAD ON TESTING SET

MODEL COMPLEXITY GRAPH

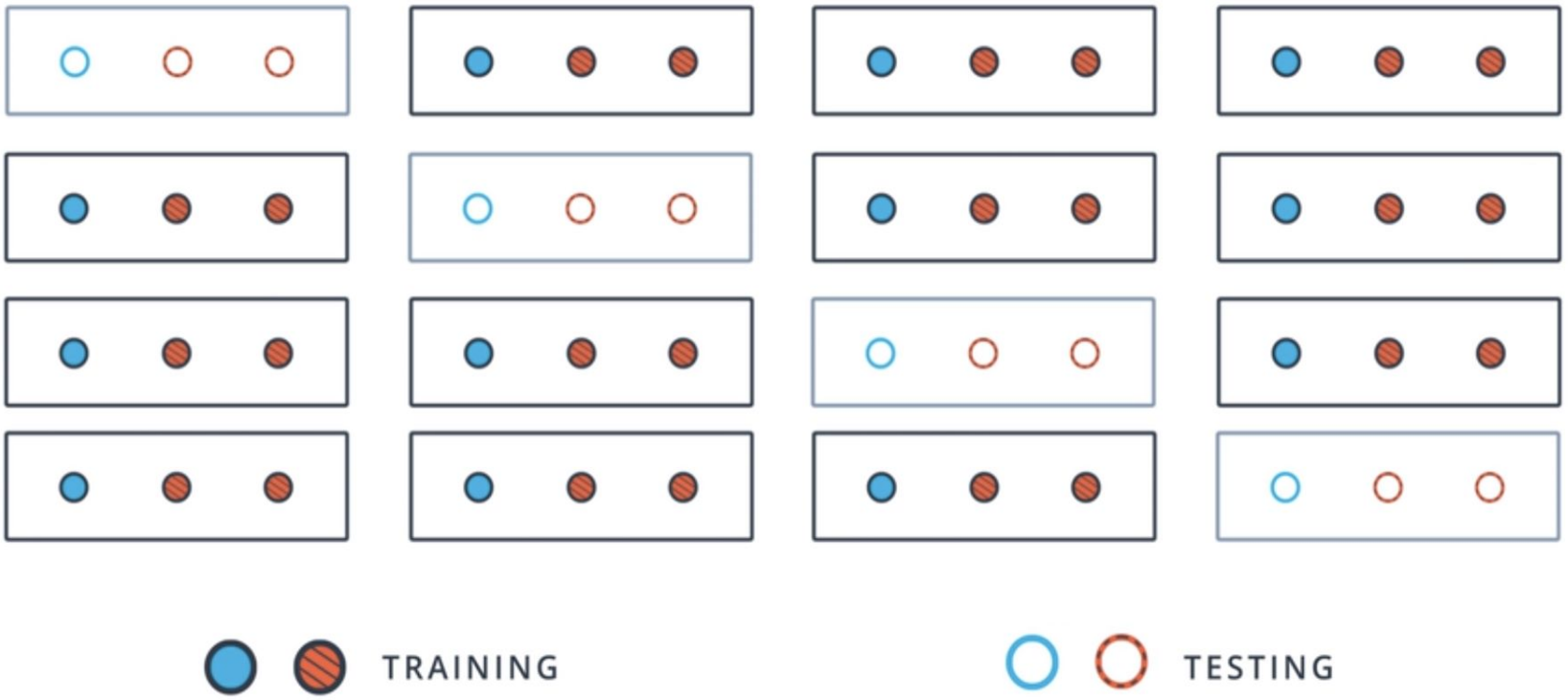


K-FOLD CROSS VALIDATION

K-FOLD CROSS VALIDATION



K-FOLD CROSS VALIDATION

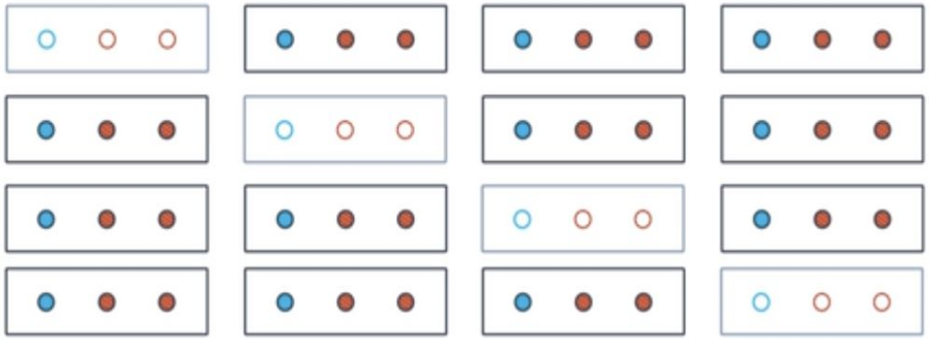


CROSS VALIDATION IN SKLEARN

```
from sklearn.model_selection import KFold
kf = KFold(12, 3)

for train_indices, test_indices in kf:
    print train_indices, test_indices
```

```
[ 3 4 4 6 7 8 9 10 11 ] [ 0 1 2 ]
[ 0 1 2 6 7 8 9 10 11 ] [ 3 4 5 ]
[ 0 1 2 3 4 5 9 10 11 ] [ 6 7 8 ]
[ 0 1 2 3 4 5 6 4  8 ] [ 9 10 11 ]
```

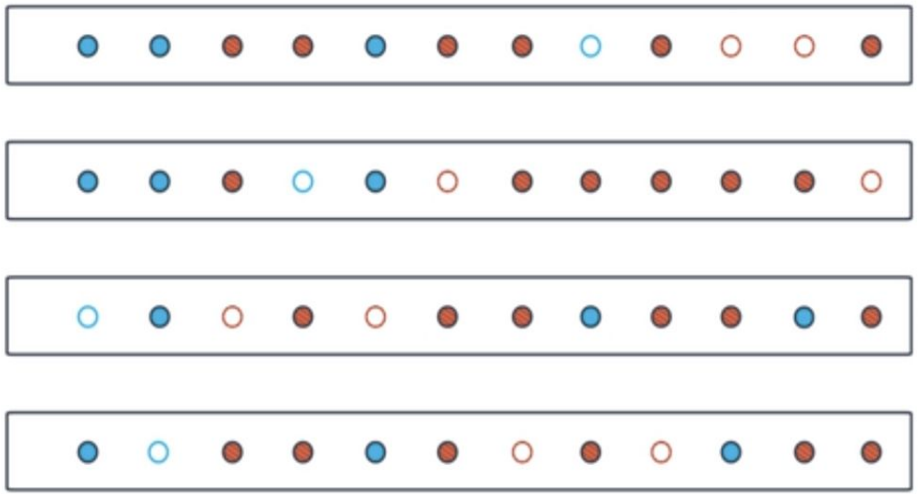


CROSS VALIDATION IN SKLEARN

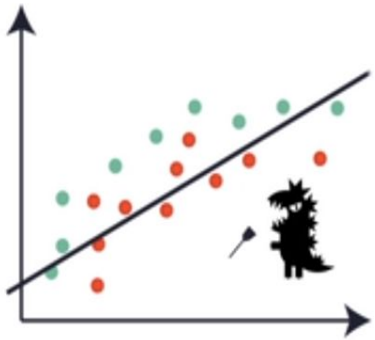
```
from sklearn.model_selection import KFold
kf = KFold(12, 3, shuffle = True)

for train_indices, test_indices in kf:
    print train_indices, test_indices
```

```
[ 0  1  2  3  4  5  6  8  11 ]   [7  9 10]
[ 0  1  2  4  6  7  8  9  10 ]   [3  5 11]
[ 1  3  5  6  7  8  9 10 11]     [0  2  4]
[ 0  2  3  4  5  7  9 10 11]     [1  6  8]
```

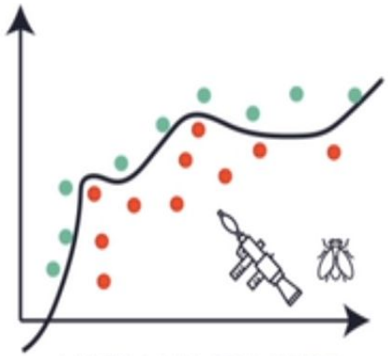


◦ LEARNING CURVES



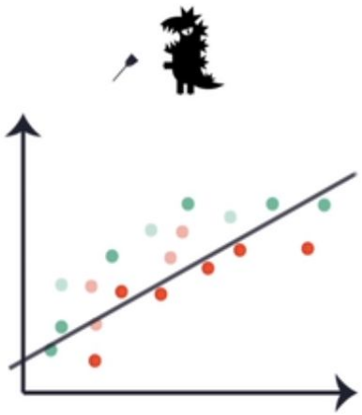
HIGH BIAS
DEGREE = 1

JUST RIGHT
DEGREE = 2



HIGH VARIANCE
DEGREE = 6

◦ LEARNING CURVES

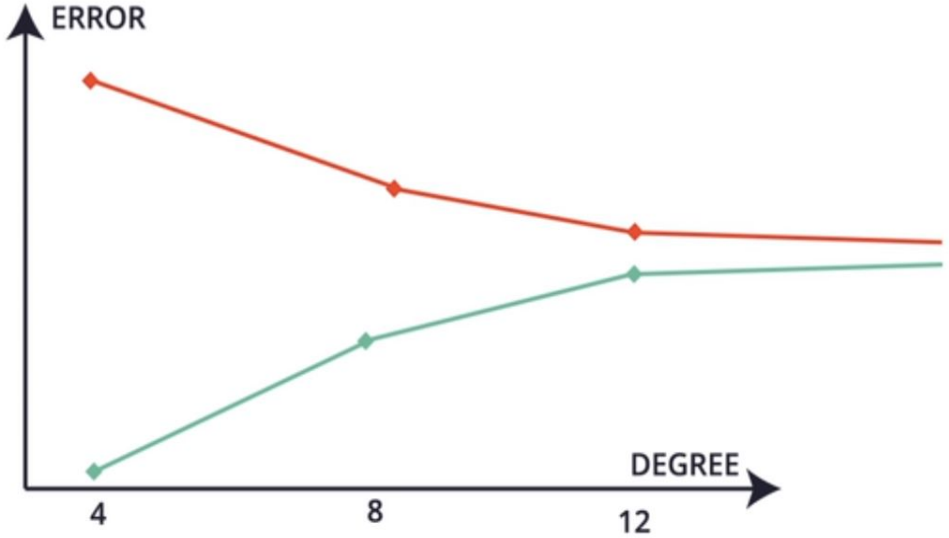


HIGH BIAS
DEGREE = 1

With 4 Training Points
Training Error: Tiny
CV Error: Large

With 8 Training Points
Training Error: Small
CV Error: Medium

With 12 Training Points
Training Error: Small
CV Error: Medium



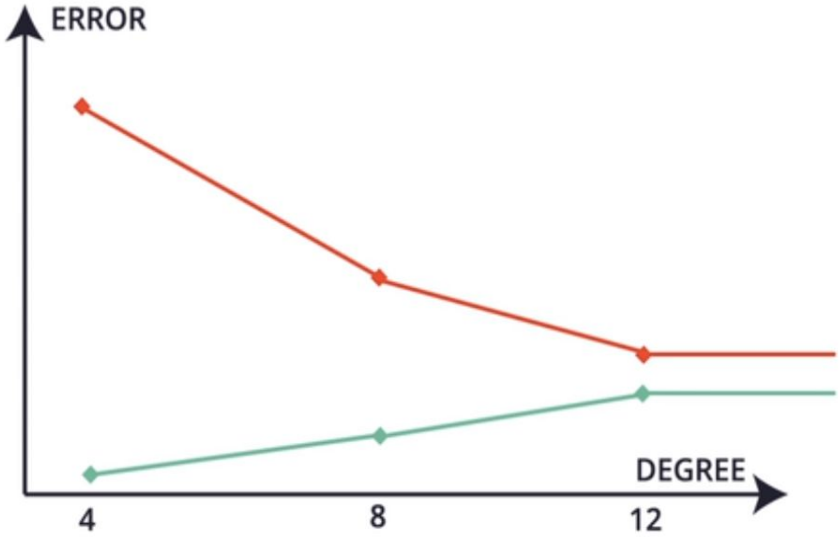
◦ LEARNING CURVES

JUST RIGHT
DEGREE = 2
(parabolas)

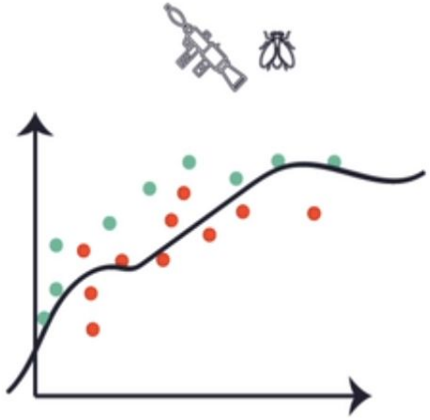
With 4 Training Points
Training Error: Tiny
CV Error: Large

With 8 Training Points
Training Error: Small
CV Error: Medium

With 12 Training Points
Training Error: Small
CV Error: Small



◦ LEARNING CURVES

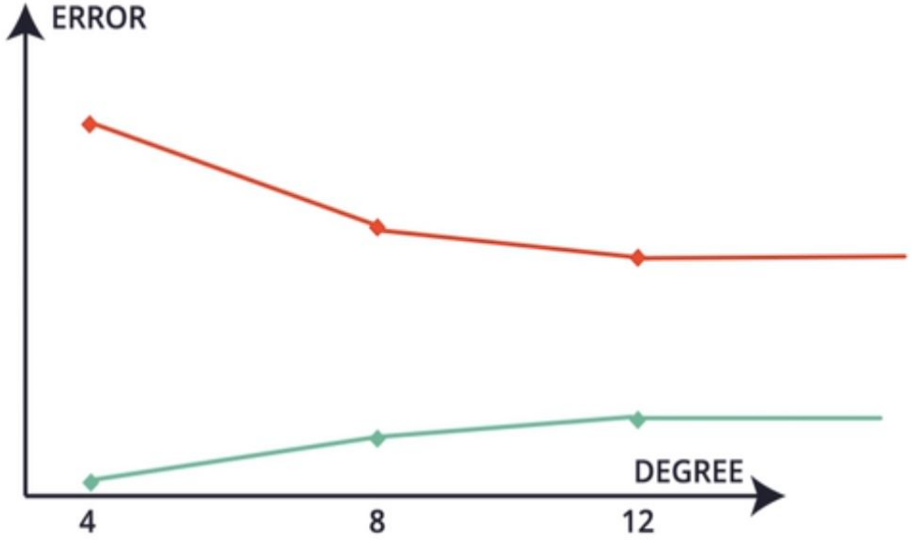


HIGH BIAS
DEGREE = 12 (CURVES)

With 4 Training Points
Training Error: Tiny
CV Error: Large

With 8 Training Points
Training Error: Small
CV Error: Large

With 12 Training Points
Training Error: Tiny
CV Error: Large



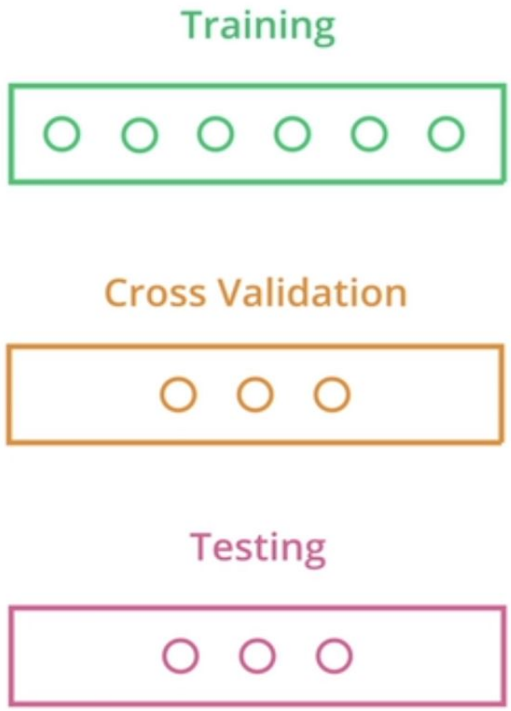
◦ LEARNING CURVES

HIGH BIAS

JUST RIGHT

HIGH VARIANCE

◦ TRAINING A LOGISTIC REGRESSION MODEL

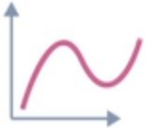


◦ TRAINING A LOGISTIC REGRESSION MODEL

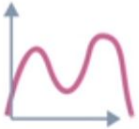
Degree = 1

Degree = 2

Degree = 3



Degree = 4

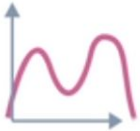


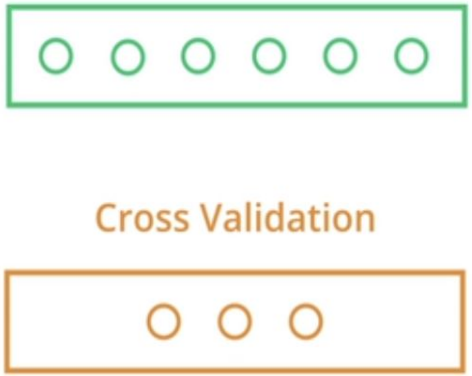
Training

Cross Validation

Testing

◦ TRAINING A LOGISTIC REGRESSION MODEL

	Parameters	F_1 Score
Degree = 1		0.5
Degree = 2		0.8
Degree = 3		0.4
Degree = 4		0.2



TRAINING A DECISION TREE

Hyper-parameters

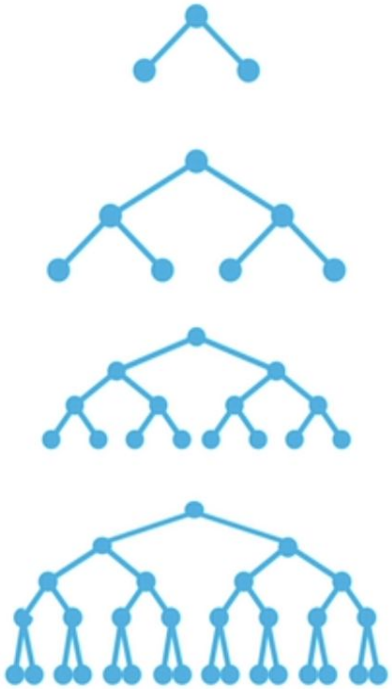
Depth =1

Depth = 2

Depth = 3

Depth = 4

Parameters



Training

Cross Validation

Testing

TRAINING A DECISION TREE

Hyper-parameters	Parameters	F_1 Score
Depth =1		0.4
Depth = 2		0.5
Depth = 3		0.9
Depth = 4		0.2

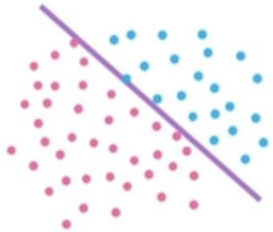
Training

Cross Validation

Testing

TRAINING A SUPPORT VECTOR MACHINE

Hyper-parameters

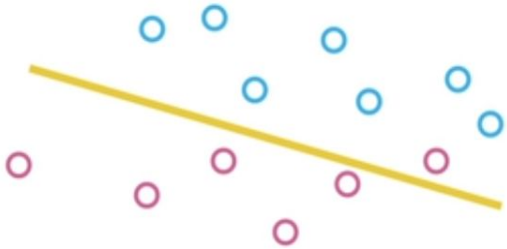


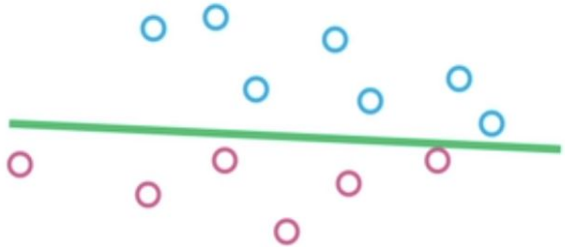
Kernel

Linear

Polynomial

C





Small

Large

◦ GRID SEARCH CROSS VALIDATION

Kernel / C	Linear	Polynomial
0.1	 F1 SCORE = 0.5	 F1 SCORE = 0.2
1	 F1 SCORE = 0.8	 F1 SCORE = 0.4
10	 F1 SCORE = 0.6	 F1 SCORE = 0.6