Model Evaluation (II): Model Selection, CV

Data Mining: Seminar 10

Dr. Thomas Price

AI Academy

NC STATE UNIVERSITY

Overfitting and Underfitting

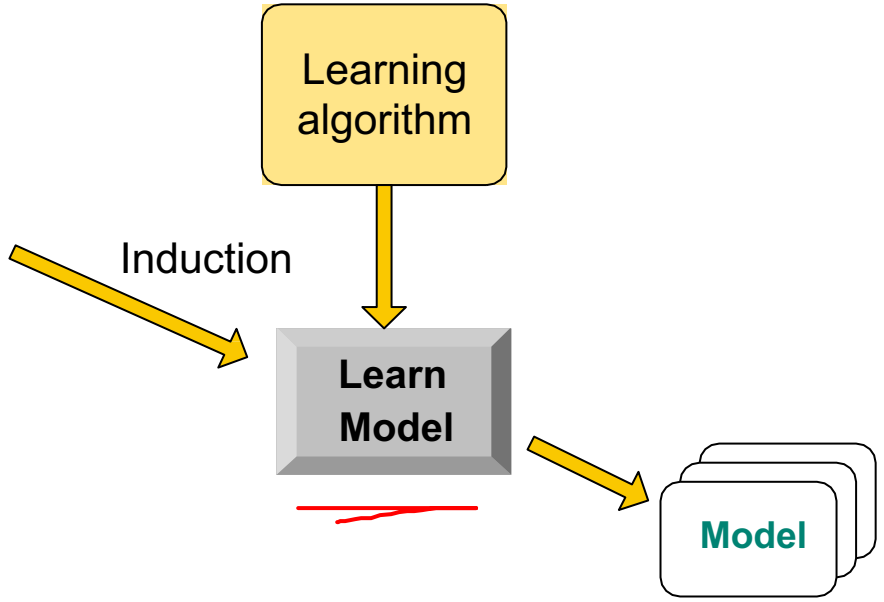
Lesson

AI Academy

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Challenges of Model Training

Training Goal: Learn the model with the lowest training error.

Problem: Low *training* error may not mean low *generalization* error.

- **Training data:** Data used to fit the model
- **Training error:** classification error on training data.

$$\frac{\text{\# incorrectly classified (in training)}}{\text{\# total (in training)}}$$

- **Generalization (true) error**: error over the whole population.

Challenges of Model Training

Solution: Test the model on unseen data.

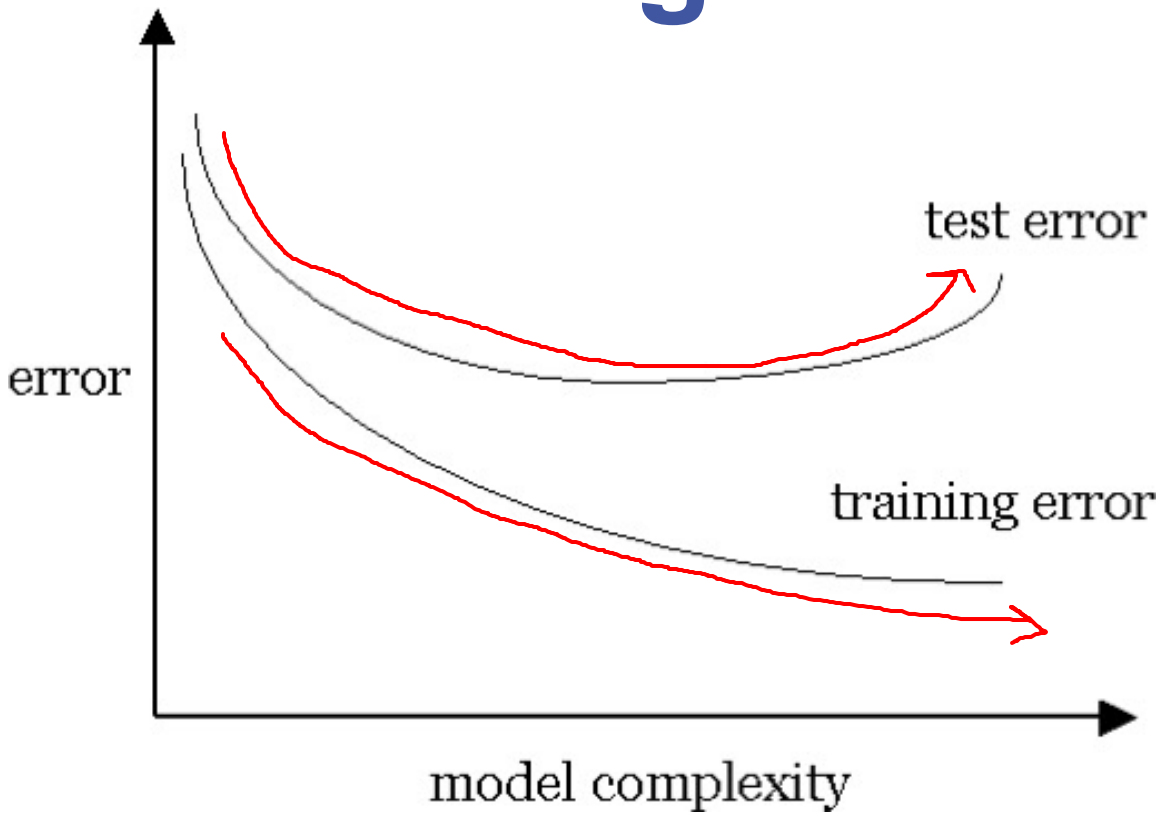
Limitation: Test error is still an imperfect estimate of generalization error.

- **Test data:** Labeled data used to evaluate (but not fit) the model.
- **Test error:** classification error on test data.

$$\frac{\text{\# incorrectly classified (in test)}}{\text{\# total (in test)}}$$

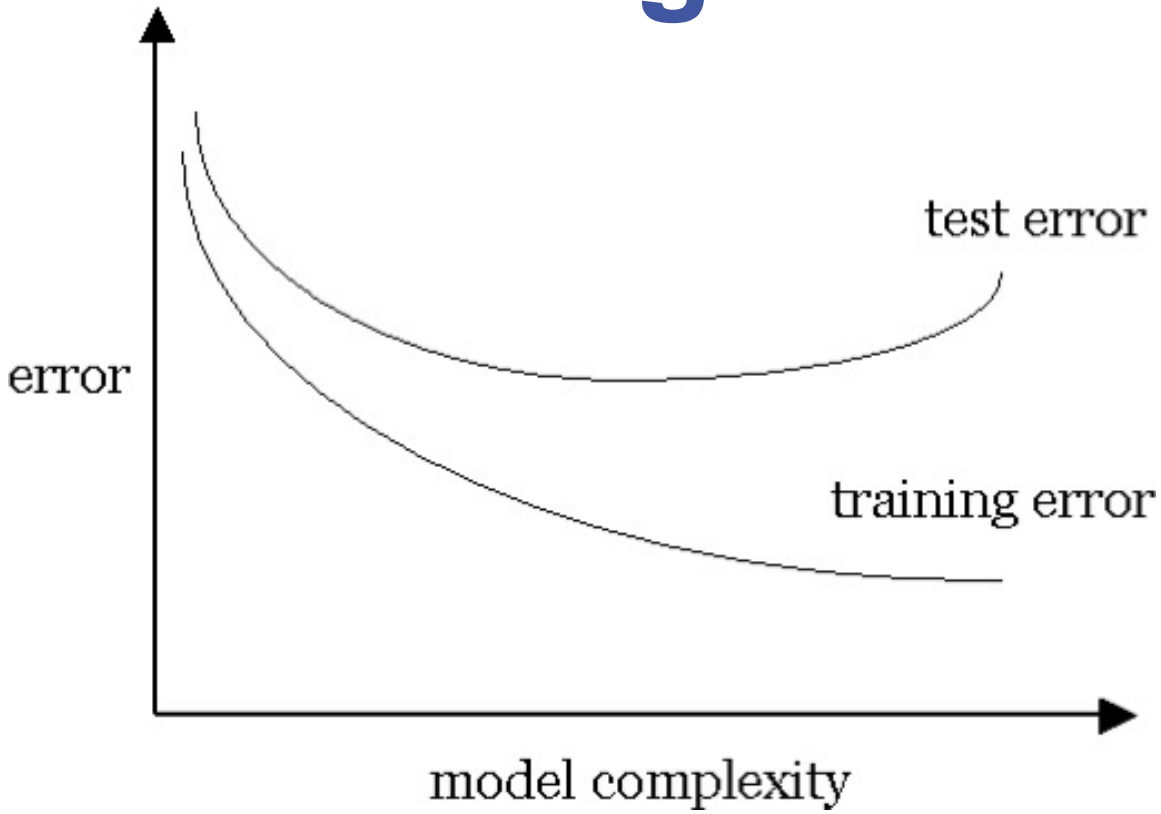
- **Generalization (true) error**: error over the whole population.

Training and Testing Error



Our goal: to find the model ***M*** which minimize the test error. This is called ***model selection***.

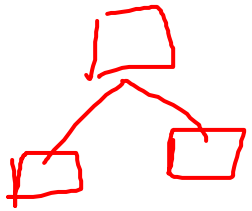
Training and Testing Error

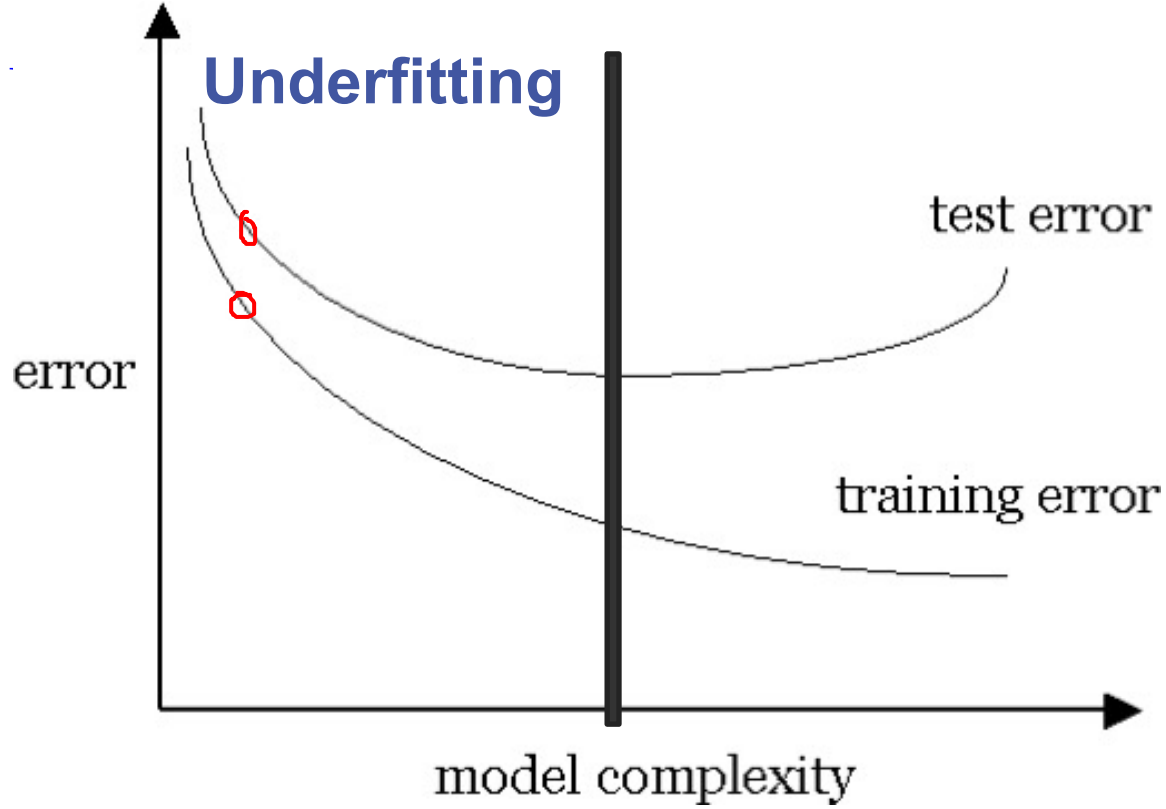


We call choosing a suboptimal model:

Underfitting ***or*** ***Overfitting***.

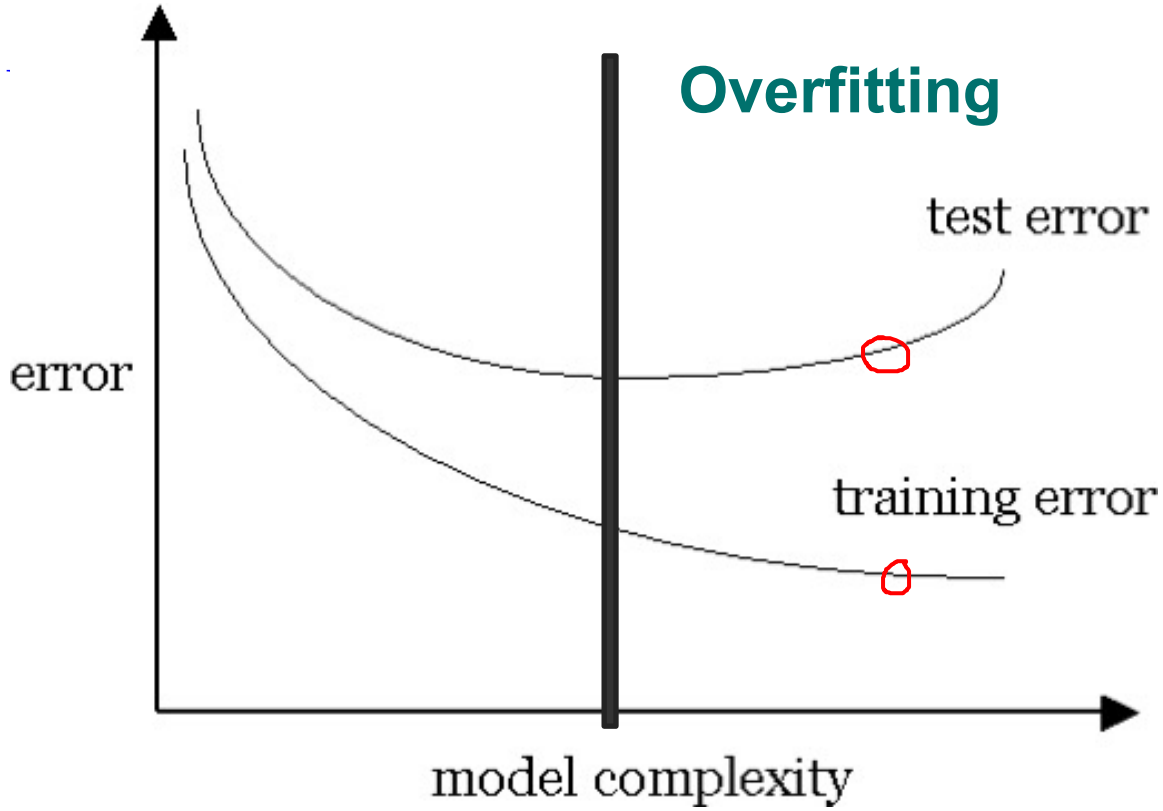
Underfitting and Overfitting





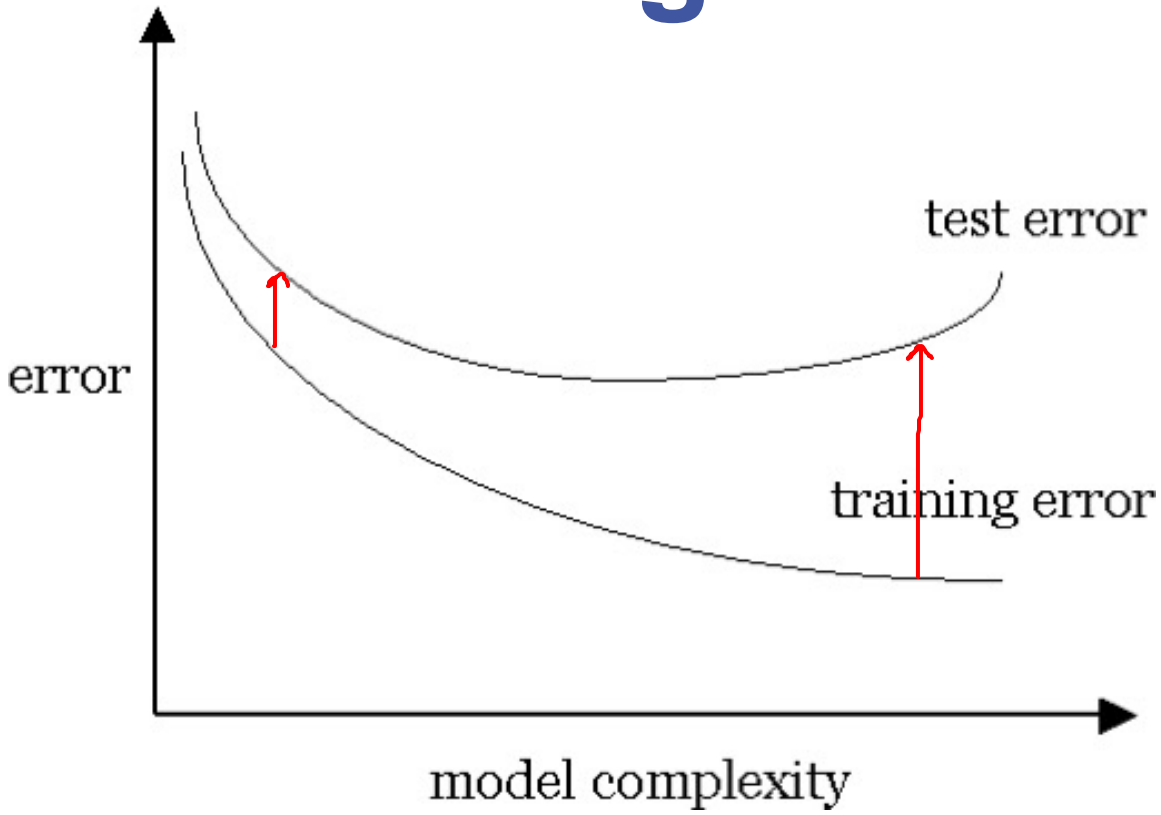
Underfitting: when model is too simple, both training and test errors are large.

Underfitting and *Overfitting*



Overfitting: when model is complex where training error is small but test error is large.

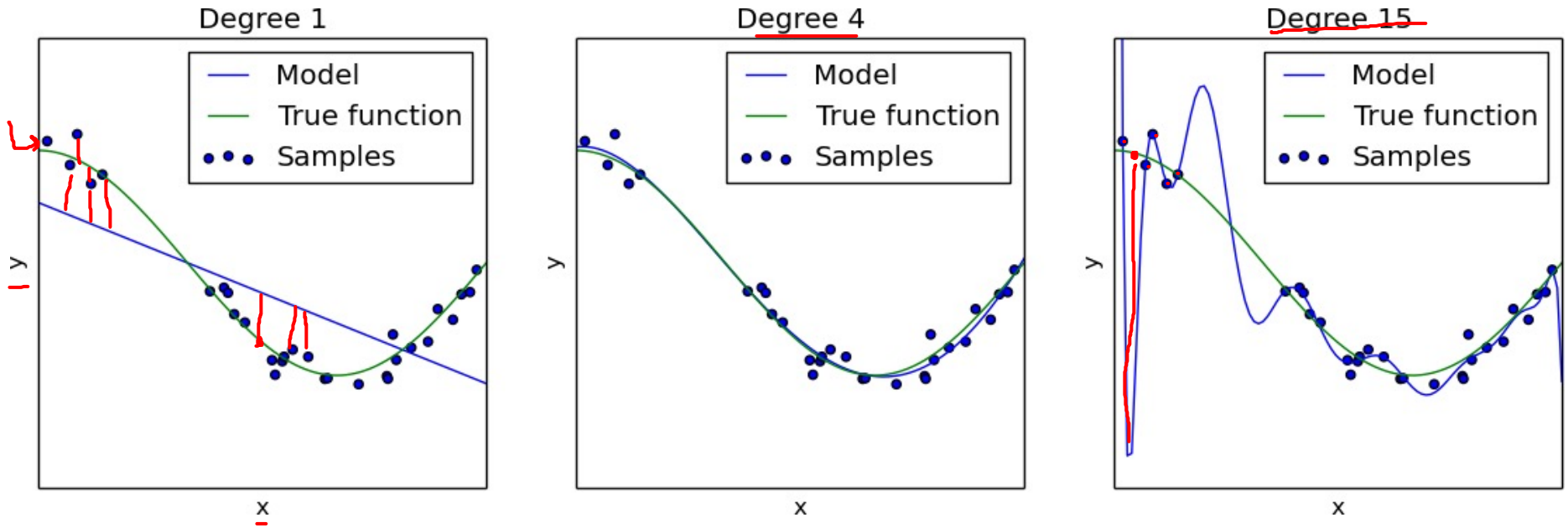
Training and Testing Error



We call choosing a suboptimal model:

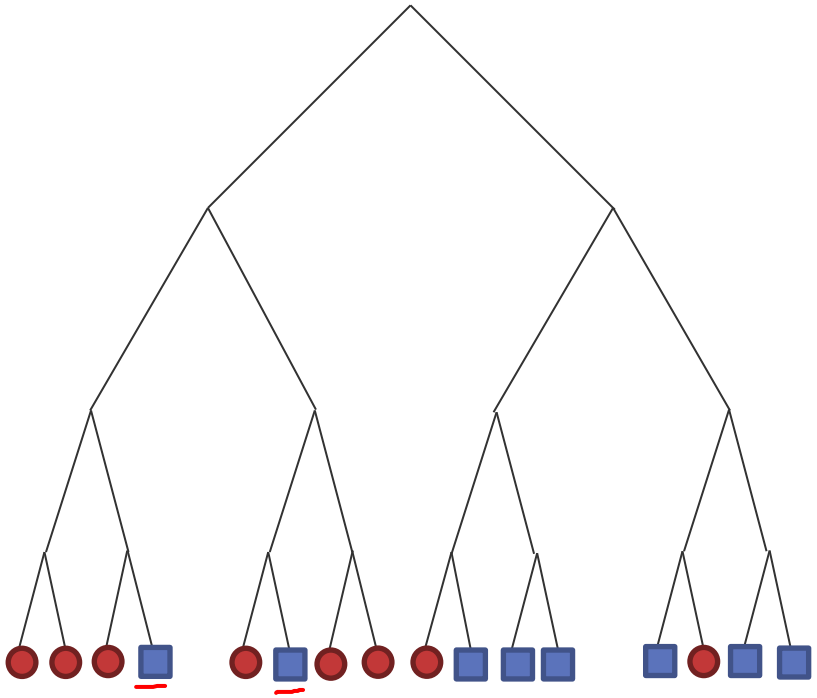
Underfitting **or** ***Overfitting***.

Polynomial Regression - Overfitting

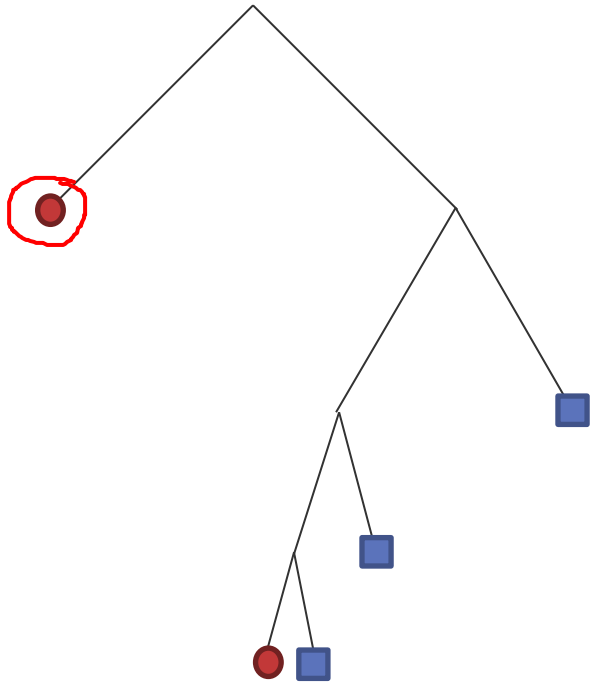


Source: https://scikit-learn.org/0.15/auto_examples/plot_underfitting_overfitting.html

Decision Trees – Overfitting



An overfit decision tree, with one leaf per data object



A pruned tree with higher training error, but maybe better generalization

Decision Trees – Overfitting

- Trivially, there is a consistent decision tree for any training dataset*.
  - One leaf node per data object.
- However, this likely won't generalize to new examples.
- It's better to find **more compact** decision trees.

**Unless the dataset has identical objects with different labels.*

Two Major Reasons for Overfitting

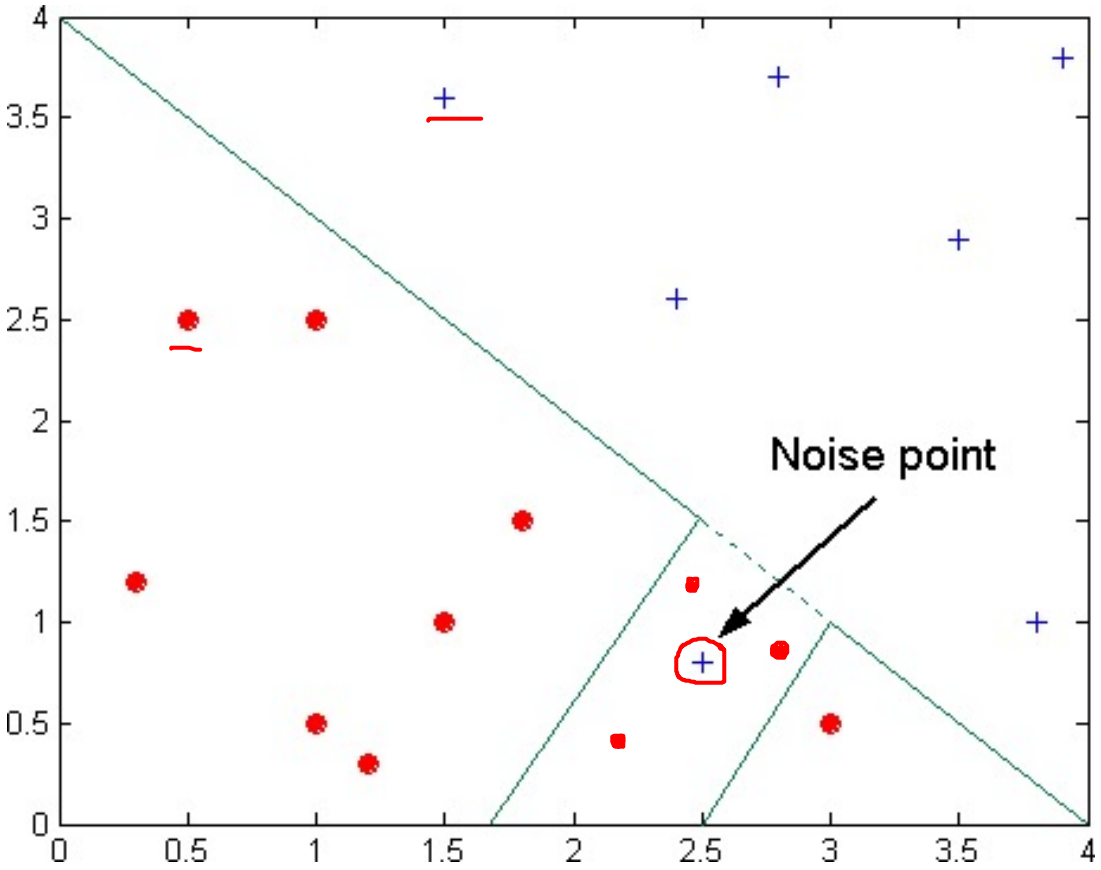
Noise:

- Complex models overreact to noise points.
- To get 1 more training instance correct…
- You may miss many test instances.

Insufficient Data:

- Model complexity should match data size.
- Complex rules based on small samples don't make sense!

Overfitting due to Noise

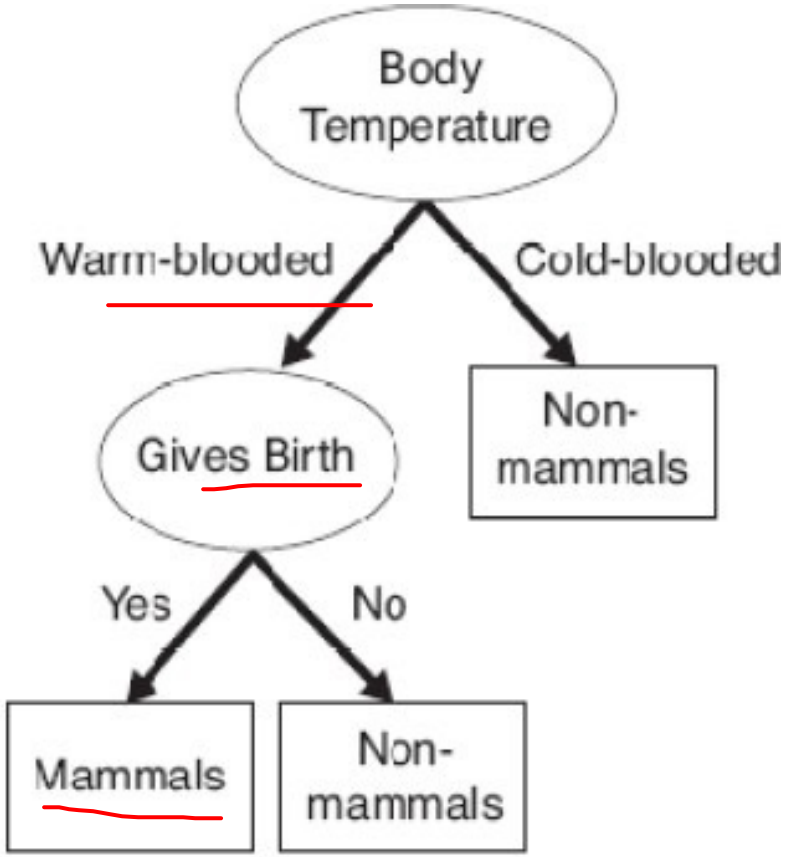


Decision boundary is distorted by noise point

Noise Example: Decision Tree

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	yes
whale	warm-blooded	yes	no	no	yes
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no
python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

Noise Example: Decision Tree

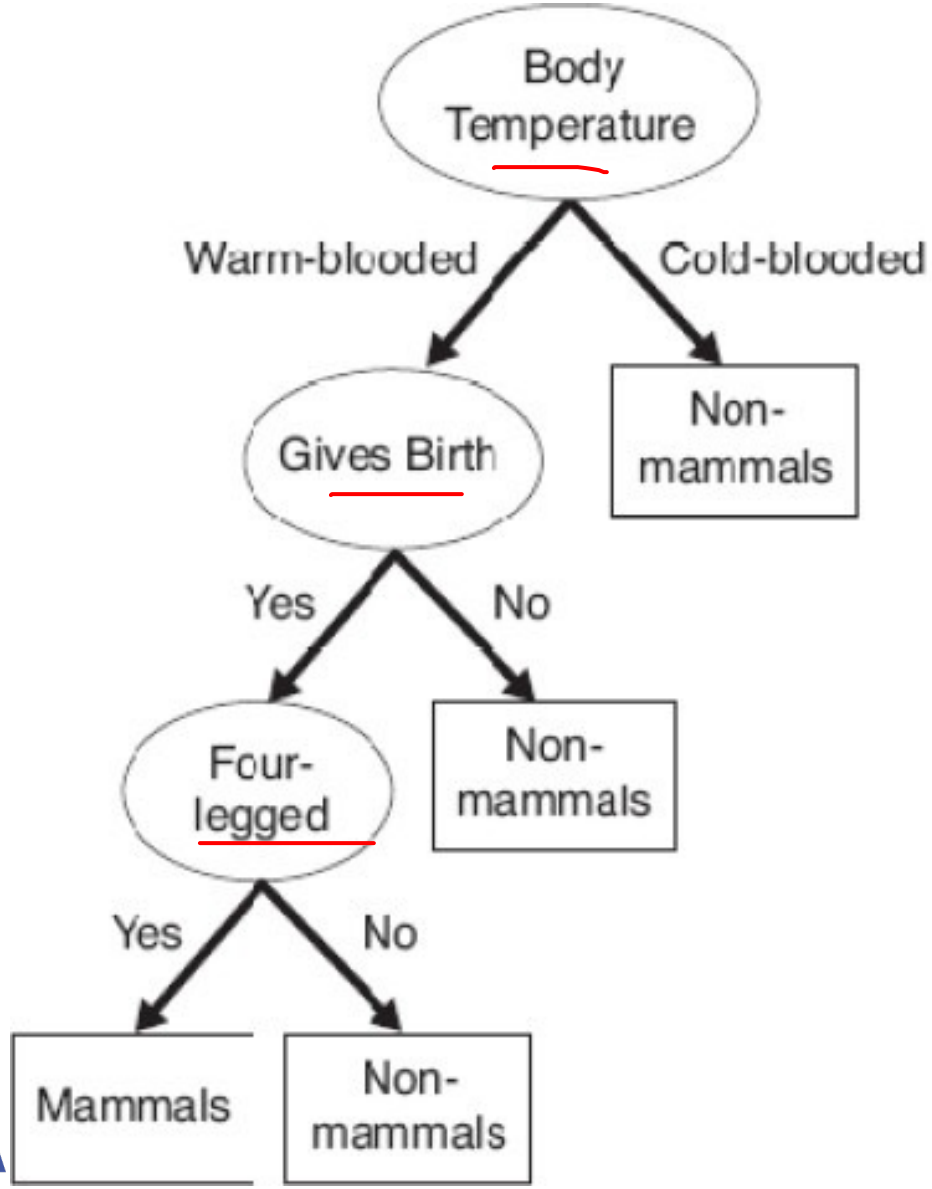


Noise Example: Decision Tree

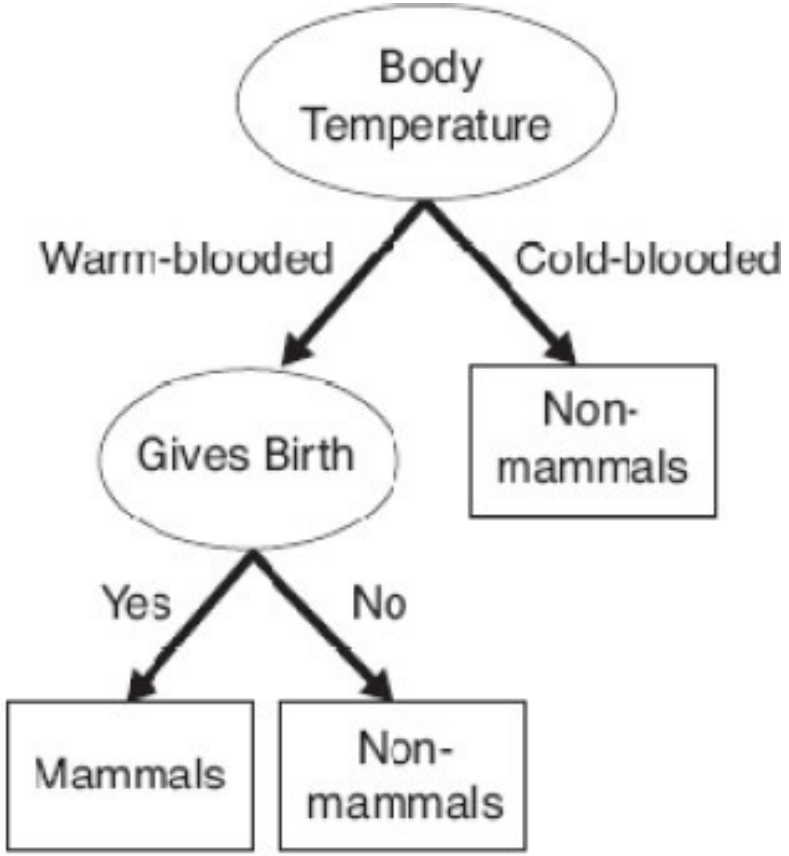
Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	no*
whale	warm-blooded	yes	no	no	no*
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no
python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

What if we add incorrectly labeled data?

A perfect tree for the imperfect data

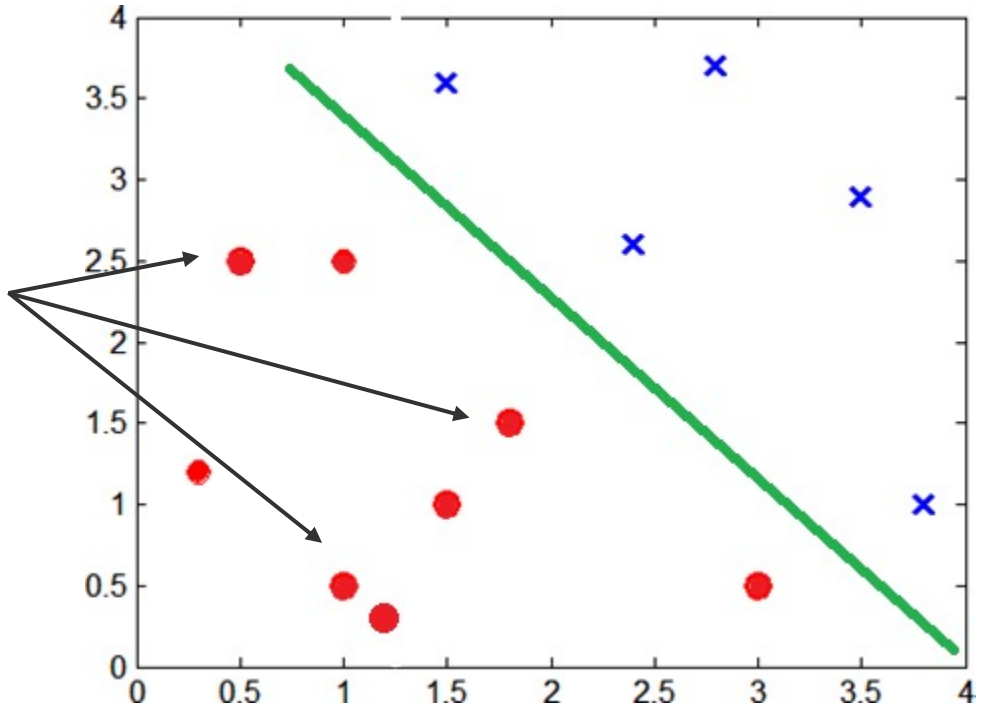


Original Tree

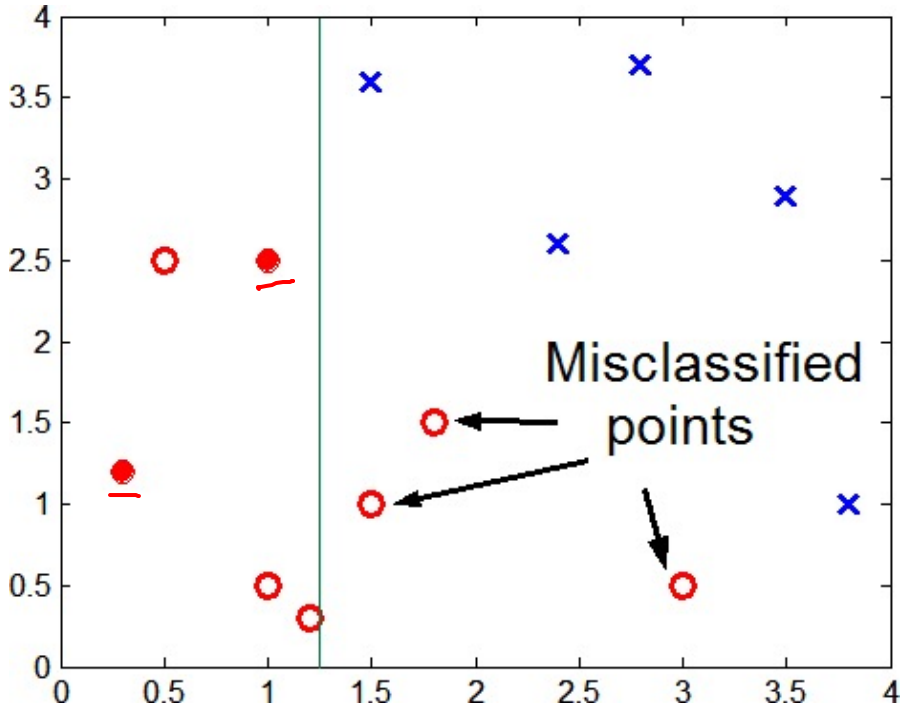


Overfitting: Insufficient Examples

What if we take away some points?



Overfitting: Insufficient Examples



- If we take away key examples, the decision boundary changes.
- The model ignores the region without data and overfits to the other regions.

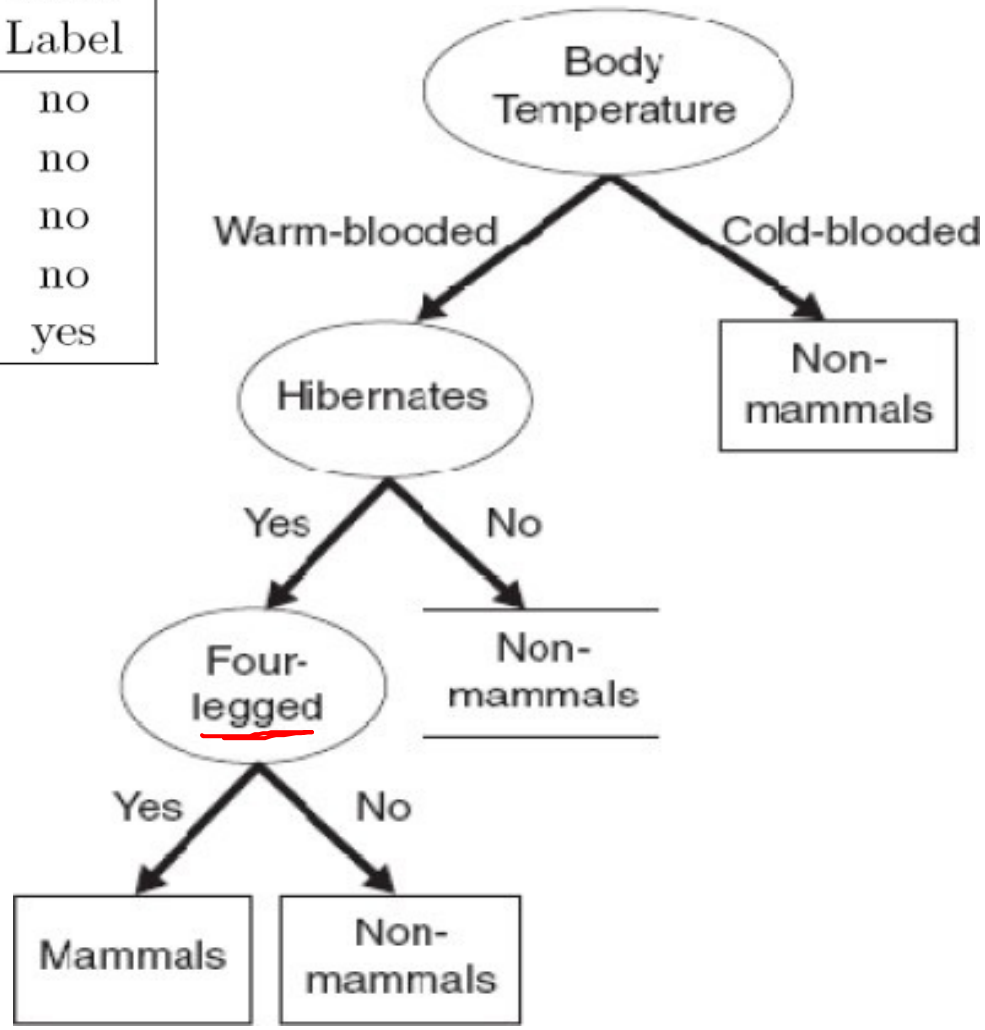
Insufficient Data: Decision Tree

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
salamander	cold-blooded	no	yes	yes	no
guppy	cold-blooded	yes	no	no	no
eagle	warm-blooded	no	no	no	no
poorwill	warm-blooded	no	no	yes	no
platypus	warm-blooded	no	yes	yes	yes

What if my only example of a mammal is the platypus?

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
salamander	cold-blooded	no	yes	yes	no
guppy	cold-blooded	yes	no	no	no
eagle	warm-blooded	no	no	no	no
poorwill	warm-blooded	no	no	yes	no
platypus	warm-blooded	no	yes	yes	yes

A perfect Decision Tree for the data above:



Notes on Overfitting

- Overfitting results in decision trees that are more complex than necessary.
- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records.
- Strategies like pruning help to prevent overfitting by limiting model complexity.

Learning Objectives: Overfitting & Underfitting

You now should be able to:

- Explain the problems caused by underfitting and overfitting data and identify when each is likely occurring.
- Differentiate types of error and what they tell us.

AI Academy
NC STATE

Overfitting and Underfitting

Exercises

AI Academy