Johns Hopkins Engineering

Applied Machine Learning for Mechanical Engineers

Machine Learning Fundamentals, Part 1, B

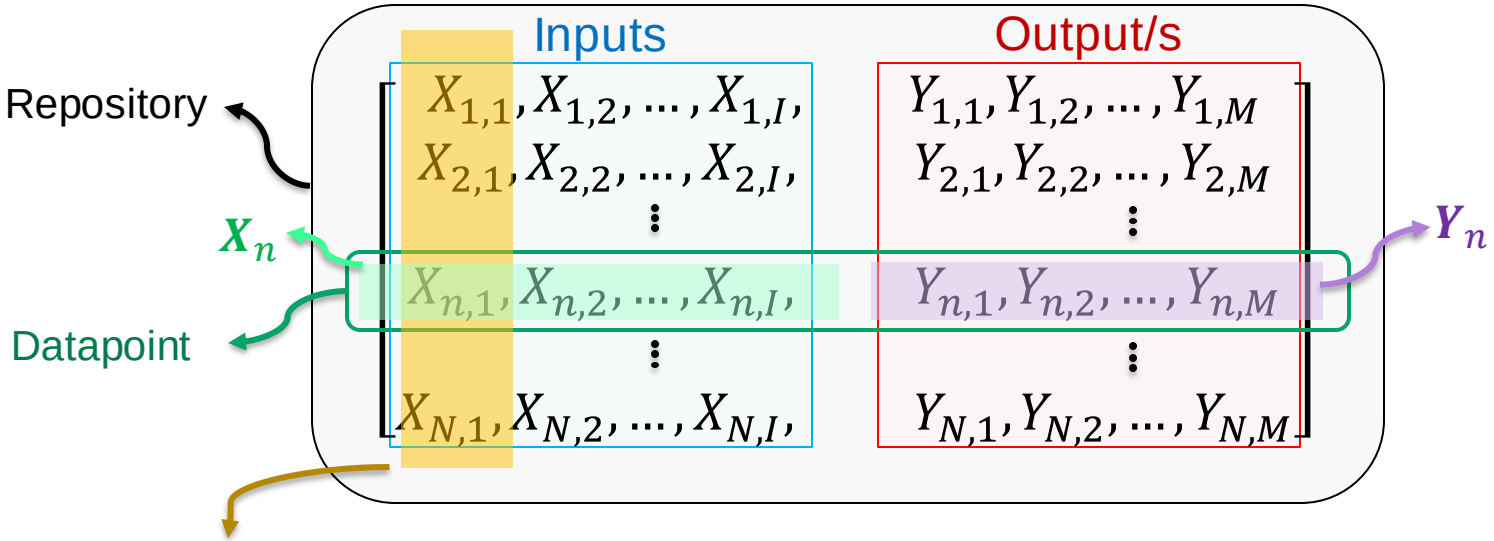
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Machine Learning Data Repository

- By the end of this lecture you will be able to:
  - Describe training-testing division for machine learning
  - Describe training-testing division statistical bias
  - Describe repository distribution bias
  - Describe overfitting in machine learning

Machine Learning Data Repository

- Jargons
  - Whether training or testing, a “datapoint” includes “inputs” and “output/s”



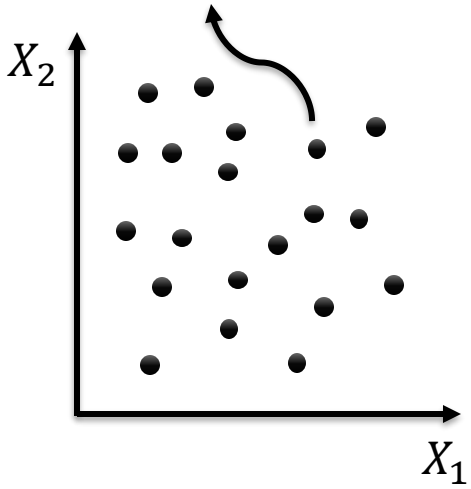
Machine Learning Data Repository

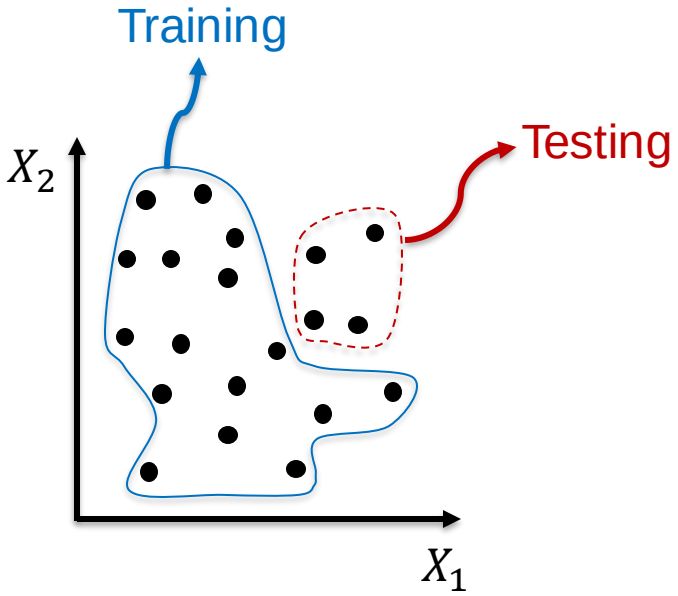
- Jargons
  - Whether training or testing, a “datapoint” include “inputs” and “output/s”
  - Repository or datapoints (training and testing all together)
  - Training repository/testing datapoints - training repository/training datapoints
  - Training inputs – training outputs
  - Testing inputs – testing outputs

Machine Learning Data Repository

- Training-testing division for machine learning

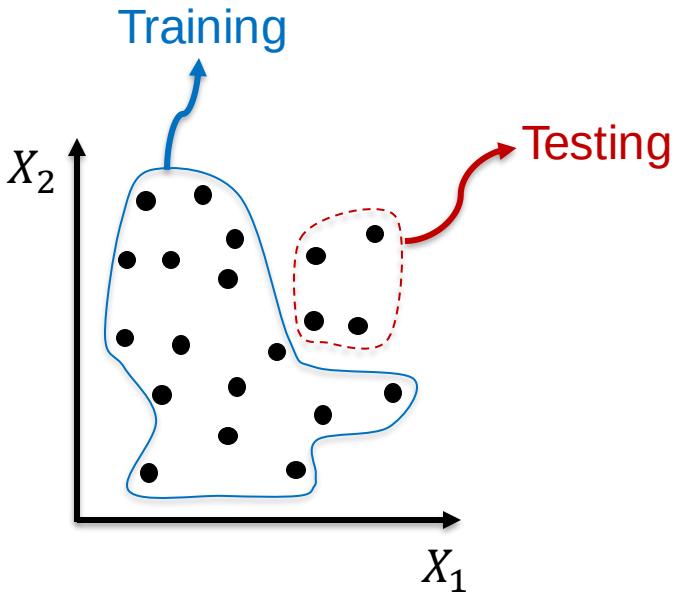
$$\boldsymbol{X}_i = [X_{i,1}, X_{i,2}] \quad \boldsymbol{Y}_i = [Y_{i,1}]$$





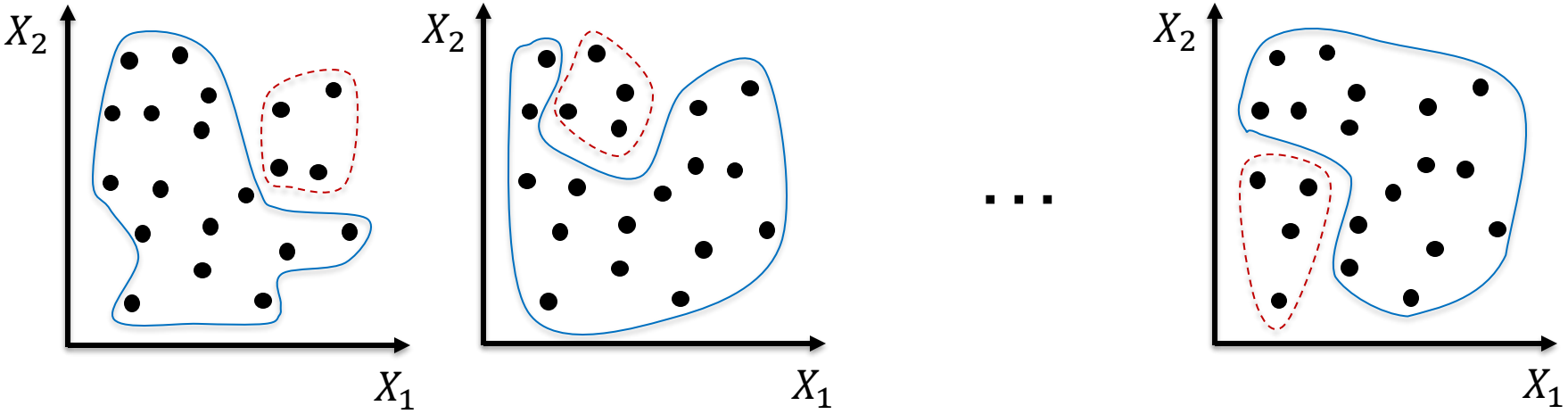
Machine Learning Data Repository

- Training-testing division for machine learning
  - Ratio of testing (RTT = 20%)
  - Measurement
    - Accuracy percentage
    - Mean-Squared-Error (MSE)
    - Mean-Absolute-Error (MAE)



Machine Learning Data Repository

- Training-testing division statistical bias
  - Repeated random sampling (RRS)



Training-testing 1 Training-testing 2 Training-testing 25

Machine Learning Data Repository

- Training-testing division statistical bias
  - Repeated random sampling (RRS)

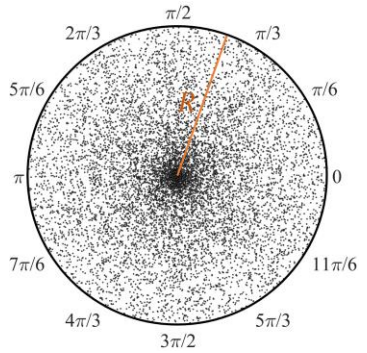
RTT	RRS			
	1	2	…	25
10%	92.1%	90.9%	…	93.7%
20%	90.3%	90.1%	…	91.2%
30%	89.9%	89.6%	…	90.0%
40%	88.7%	89.1%	…	88.7%
50%	86.1%	86.0%	…	85.9%

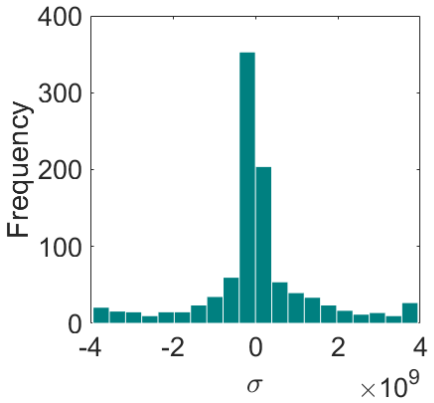
Machine Learning Data Repository

- Repository distribution bias
  - Poor distribution of attributes
  - Example: an image repository contaminated with societal bias could trick a machine learning model to conclude a higher false-negative rate of smile for a particular gender

Machine Learning Data Repository

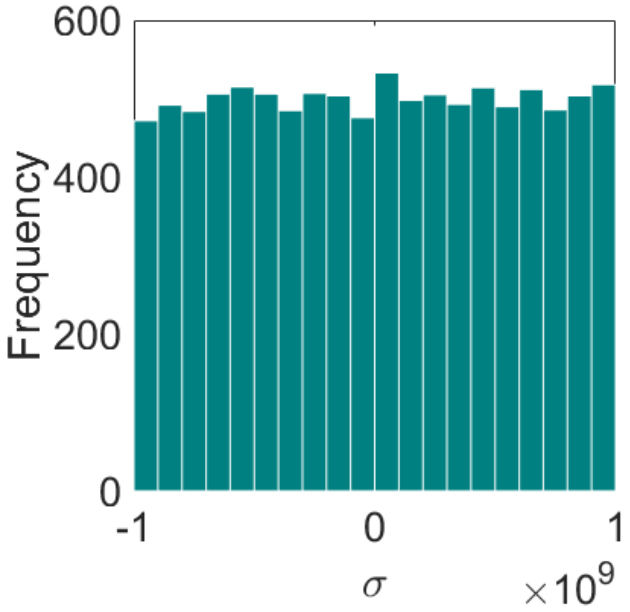
- Repository distribution bias
  - Poor distribution of attributes
  - Example: machine learning will often observe and learn the patterns of near-zero stress magnitudes compared with large magnitudes





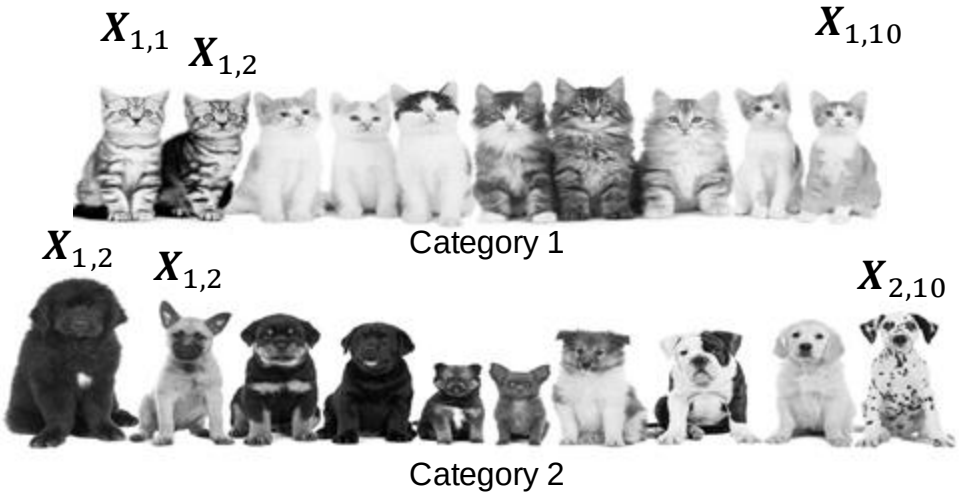
Machine Learning Data Repository

- Repository distribution bias
  - Poor distribution of attributes
  - Example: ideal distribution of attributes is a near-uniform distribution



Machine Learning Data Repository

- Overfitting
  - Fitting the machine learning to only recognize and fit the training data

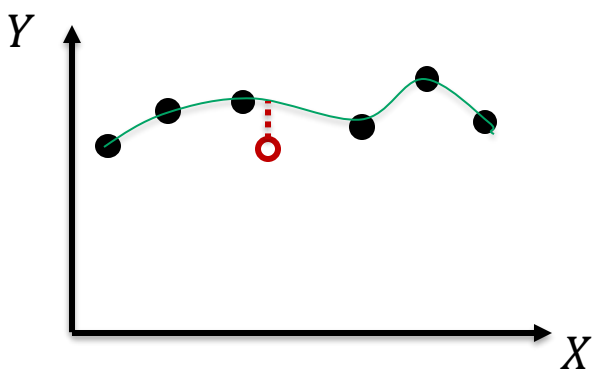


Training data

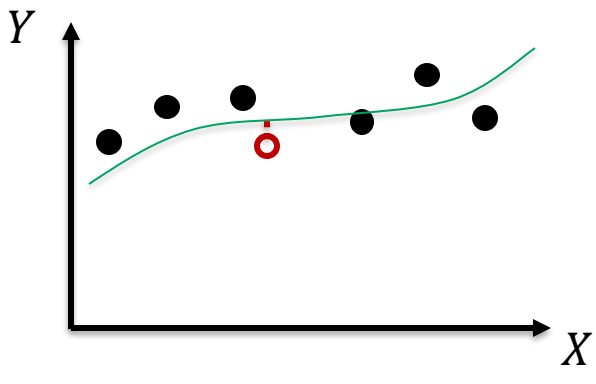
Testing data
Category 1 or 2?

Machine Learning Data Repository

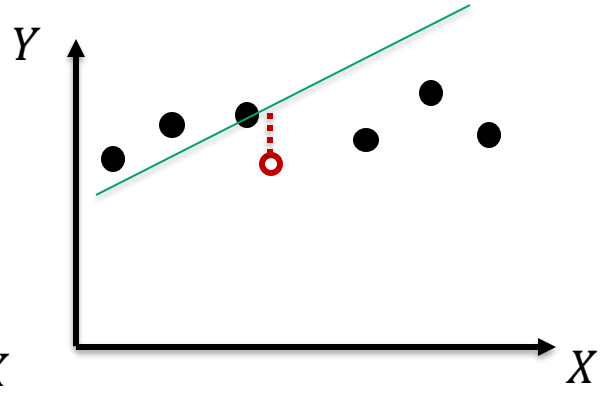
- Overfitting
  - Fitting the machine learning to only recognize and fit the training data



Overfitted



Fitted



Underfitted

Machine Learning Data Repository

- In this lecture, you learned about:
  - Training-testing division for machine learning
  - Training-testing division statistical bias
  - Repository distribution bias
  - Overfitting in machine learning
- In the next lecture, we will talk about perceptron and neural networks