4. Machine Learning Overview

4.1 Core Machine Learning Concepts - Part I

Cristiano Bacelar de Oliveira

Agosto, 2020

Universidade Federal do Ceará

Index

- Machine Learning Overview
- Application Scenarios
- Common Tasks
- Classification of Learning Algorithms
- Basic Concepts

Disclaimer

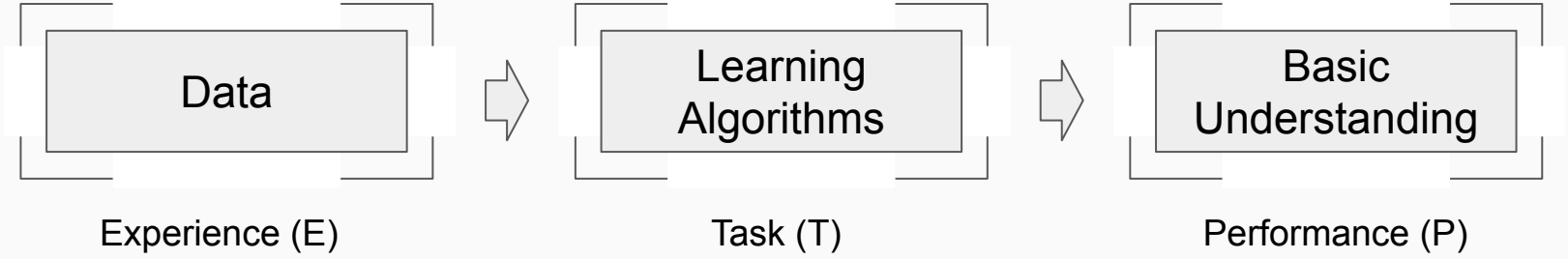
The following content is heavily based on HCIA-AI Course material by Huawei Technologies Co., Ltd., authored by Zhang Luxiao. Distribution is not allowed.

Machine Learning Overview

Machine Learning Overview

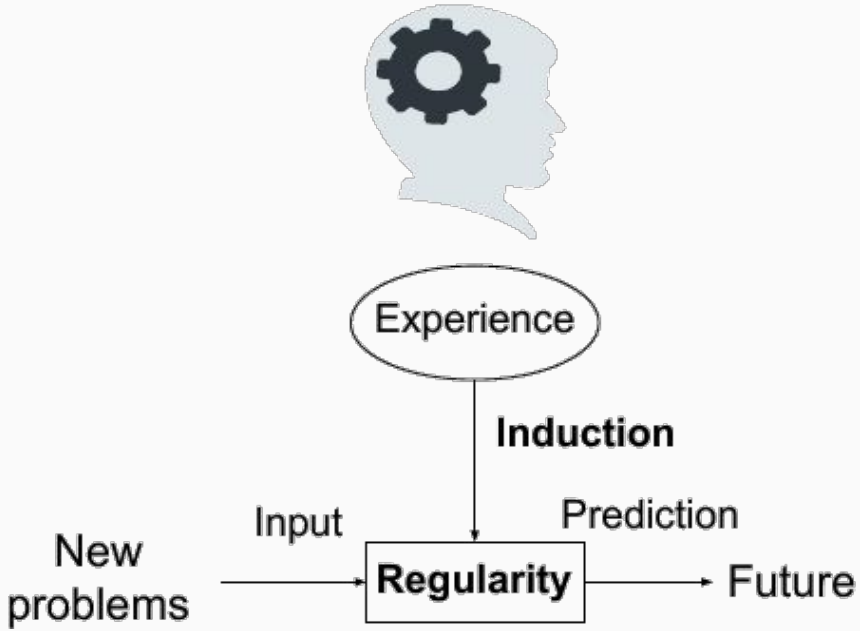
Machine learning (including deep learning) is a study of learning algorithms.

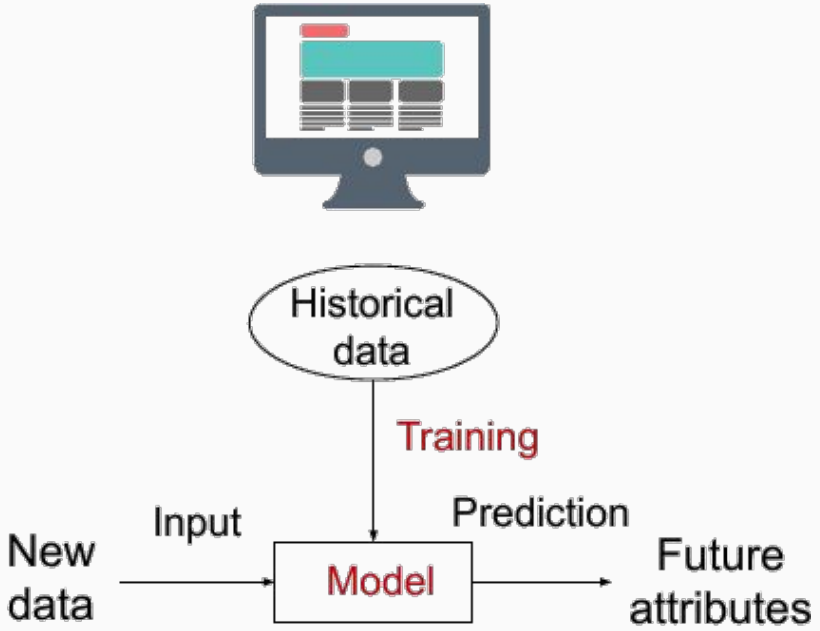
“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”[1]



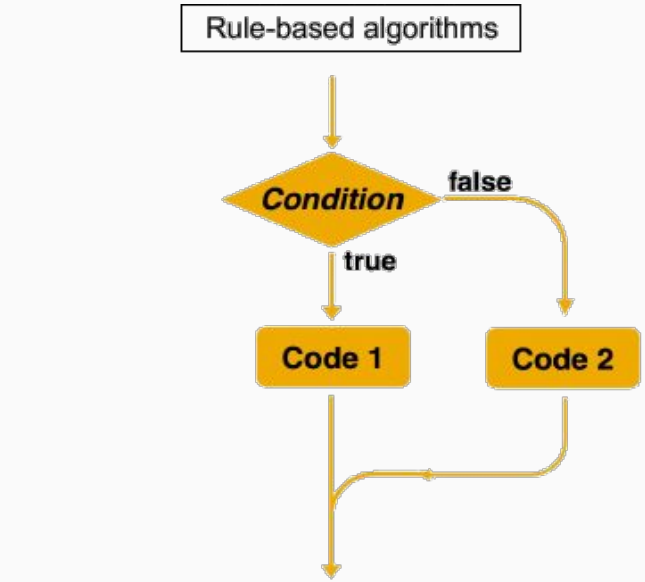
[1]From the book: Machine Learning, by Tom Mitchel, IMcGraw-Hill Science, 1997.

Machine Learning Overview (cont.)

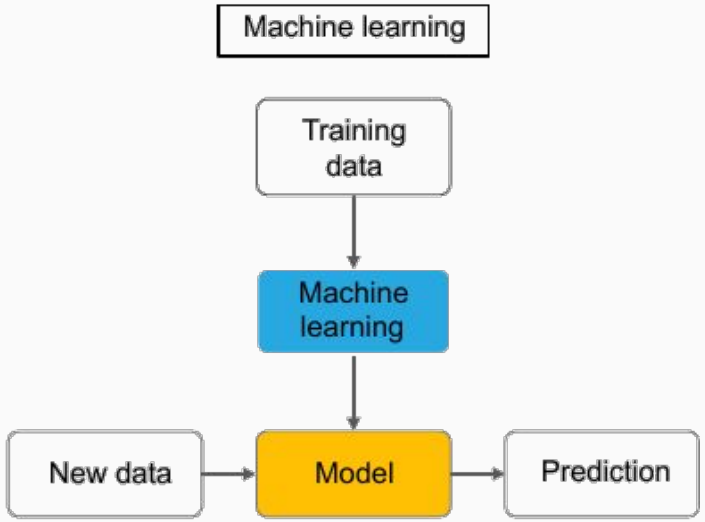




Machine Learning vs. Traditional Rule-Based Algorithms



- Explicit programming is used to solve problems.
- Rules can be manually specified.



- Samples are used for training.
- The decision-making rules are complex or difficult to describe.
- Rules are automatically learned by machines.

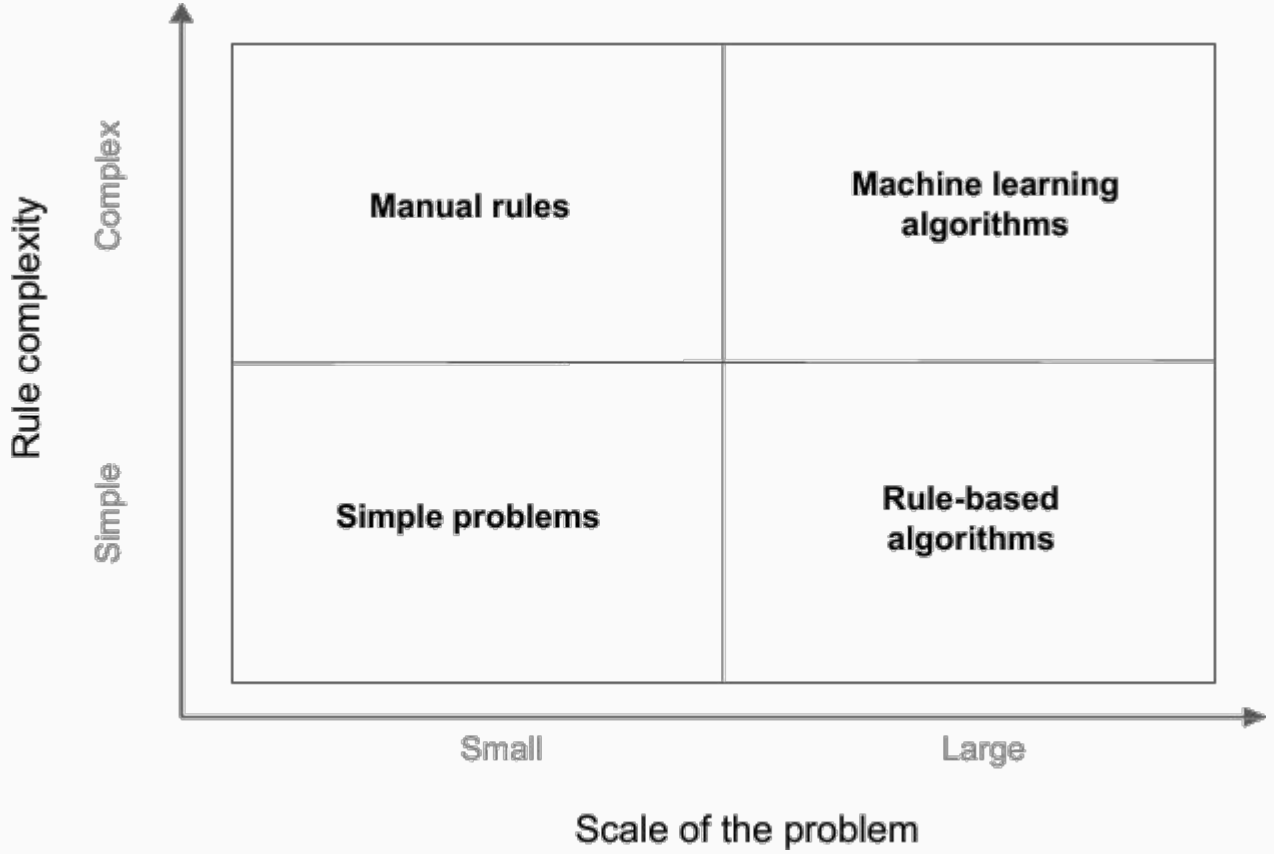
Application Scenarios of ML

The solution to a problem is complex, or the problem may involve a large amount of data without a clear data distribution function.

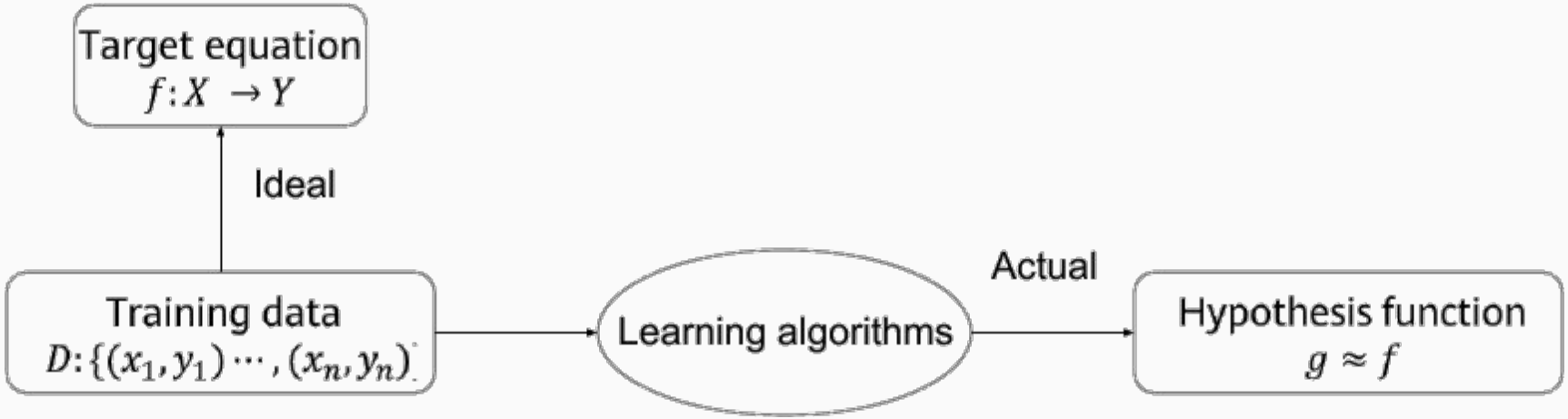
Machine learning can be used in the following scenarios:

- Rules are complex or cannot be described
  - Example: Facial recognition and voice recognition.
- Task rules change over time.
  - Example: in the part-of-speech tagging task, new words or meanings are generated at any time.
- Data distribution changes over time, requiring constant readaptation
  - Example: Predicting the trend of commodity sales.

Application Scenarios of ML



Rational Understanding of Machine Learning Algorithms



Common types of tasks for ML

Machine learning algorithms can deal with many type of tasks. The most common are:

- Classification
  - The program needs to assign labels (categories or classes) to the inputs.
  - $f : R^n \rightarrow (1,2,...,k)$
- Regression
  - The program needs to output a prediction for a given input.
  - $f : R^n \rightarrow R$
- Clustering
  - The program needs group unlabeled inputs into categories according to the internal similarity of the data.

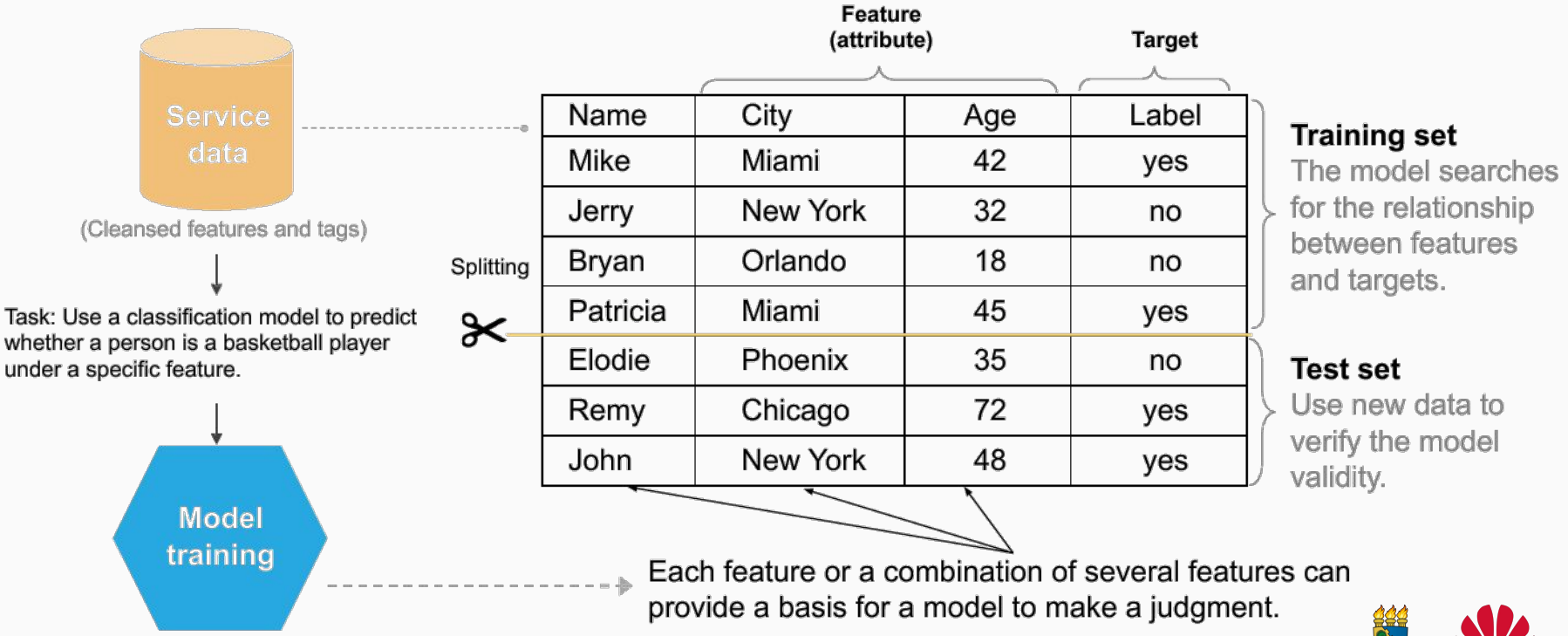
Classification of Learning Algorithms

- Supervised learning
  - Obtain an optimal model with required performance through training and learning based on the samples of known categories. Then, use the model to map all inputs to outputs and check the output for the purpose of classifying unknown data.
- Unsupervised learning
  - For unlabeled samples, the learning algorithms directly model the input datasets. Clustering is a common form of unsupervised learning. We only need to put highly similar samples together, calculate the similarity between new samples and existing ones, and classify them by similarity.

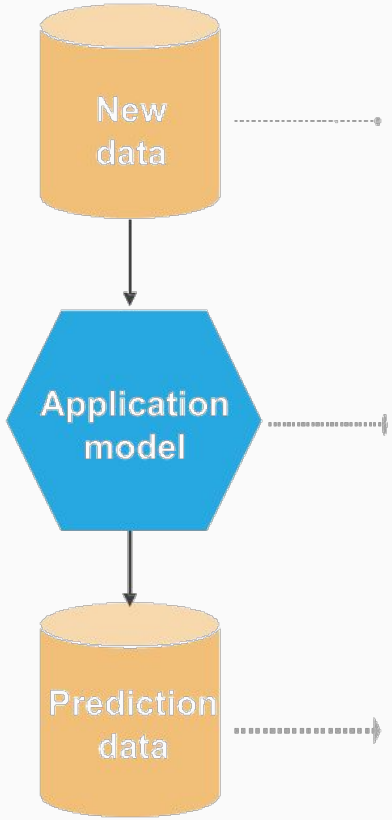
Classification of Learning Algorithms

- Semi-supervised learning
  - In one task, a machine learning model that automatically uses a large amount of unlabeled data to assist learning directly of a small amount of labeled data.
- Reinforcement learning
  - It is an area of machine learning concerned with how agents ought to take actions in an environment to maximize some notion of cumulative reward. The difference between reinforcement learning and supervised learning is the teacher signal. The reinforcement signal provided by the environment in reinforcement learning is used to evaluate the action (scalar signal) rather than telling the learning system how to perform correct actions.

Supervised Learning Example



Supervised Learning Example (cont.)



Name	City	Age	Label
Marine	Miami	45	?
Julien	Miami	52	?
Fred	Orlando	20	?
Michelle	Boston	34	?
Nicolas	Phoenix	90	?

Unknown data
Recent data, it is not known whether the people are basketball players.

IF city = Miami → Probability = +0.7
IF city= Orlando → Probability = +0.2
IF age > 42 → Probability = +0.05*age + 0.06
IF age ≤ 42 → Probability = +0.01*age + 0.02

Name	City	Age	Prediction
Marine	Miami	45	**0.3**
Julien	Miami	52	**0.9**
Fred	Orlando	20	**0.6**
Michelle	Boston	34	**0.5**
Nicolas	Phoenix	90	**0.4**

Possibility prediction
Apply the model to the new data to predict whether the people are basketball players.

Basic Concepts

Basic Concepts - Datasets

- Dataset
  - A collection of data used in machine learning tasks. Each data record is called a sample. Events or attributes that reflect the performance or nature of a sample in a particular aspect are called features.
- Training set:
  - A dataset used in the training process, where each sample is referred to as a training sample. The process of creating a model from data is called learning (training).
- Test (or Validation) set
  - Testing refers to the process of using the model obtained after learning for prediction. The dataset used is called a test set, and each sample is called a test sample.

Basic Concepts - Datasets

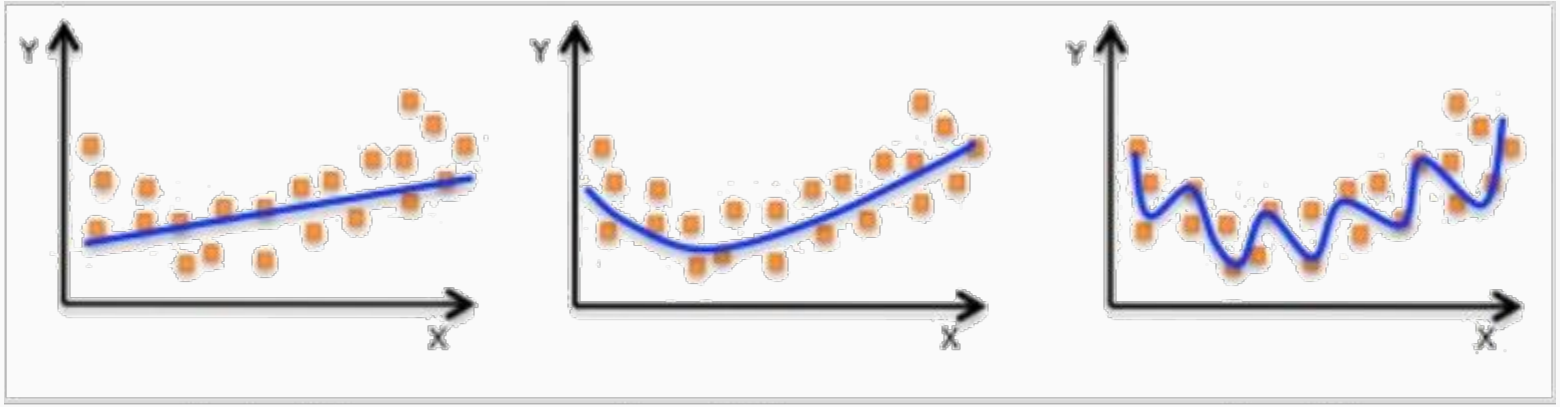
- Typical dataset form

		Feature 1	Feature 2	Feature 3	Label
	No.	Area	School Districts	Direction	House Price
Training set	1	100	8	South	1000
	2	120	9	Southwest	1300
	3	60	6	North	700
	4	80	9	Southeast	1100
Test set	5	95	3	South	850

Basic Concepts - Model Validity

- Generalization capability
  - The goal of machine learning is that the model obtained after learning should perform well on new samples, not just on samples used for training. The capability of applying a model to new samples is called generalization or robustness.
- Underfitting
  - Occurs when the model or the algorithm does not fit the data well enough.
- Overfitting
  - Occurs when the training error of the model obtained after learning is small but the generalization error is large (poor generalization capability).

Basic Concepts - Model Validity



Underfitting
Not all features are learned.

Good fitting

Overfitting
Noises are learned.

Basic Concepts - Error

- Error
  - Difference between the sample result predicted by the model obtained after learning and the actual sample result.
- Training error
  - Error that you get when you run the model on the training data.
- Generalization (Testing) error
  - error that you get when you run the model on new samples. Obviously, we prefer a model with a smaller generalization error.

Thank You!

Next: 4.2 - Core Machine Learning Concepts - Part II