The Data Science Process, Part 1

Neba Nfonsang

The Data Science Process Outline: Part 1

- Introduction to data science
- Types of learning task
- Data preprocessing
  - Data cleaning
  - Data exploration
  - Data transformation
  - Feature selection/extraction

Introduction to Data Science Tools

What Is Data Science?

- Data science is an interdisciplinary field that uses a collection of tools, methods, and algorithms from other fields such as machine learning, data mining, statistics, and computing.
- Data science is a systematic process of extracting meaning from data.
- The explosive growth of data and increased computing power have promoted the use of data science tools.

Components of a Dataset

An observation, instance, example, a tuple

Variables

A dataset

	Height	Weight	City
1	65	112	Denver
2	71	136	Aurora
3	69	153	Boulder
4	68	142	Denver
5	67	144	Denver
6	68	123	Boulder

Data-Related Terminology

- **A variable** is a characteristic or property of an entity or object that takes on different values. Each column in a dataset is a variable; for example, height, weight, and city.
- **Data** are values or facts associated with variables.
- **An observation** is a single instance in the dataset. It is a row of data values in the dataset. Alternative names include **instance, example, tuple, data point, or record.**
- **A dataset** is a collection of observations.

Types of Variables

Categorical (nominal) variable

- This is a variable whose values fall into categories. This is a variable whose unique values form a finite set.
- For example, the categorical variable, gender, could assume values (male or female).

Numerical variable

- This is a variable whose values are numbers with consistent intervals; for example, age.
- Numerical variables take on numbers that are real-valued (continuous) or integer-valued (discrete).

Input and Output Variables

- **Input variable:** this is an independent variable, also called:
  - Input
  - Feature (input feature)
  - Attribute
  - Predictor
  - Covariate
  - X-variable
- **Output variable**: this is the dependent variable, also called:
  - Output
  - Target
  - Response
  - Outcome
  - Class or label (when categorical)
  - Y-variable

Input and Output Variables

Input variables

Output variable

	Credit Score	Missed Payments	Home-owner	Credit Age	Time on Job	Loan Status
1	600	2	0	5.6	7	1
2	649	0	1	5.0	7	1
3	485	4	0	7.4	6	0
4	624	2	1	7.4	6	0

Training Examples and Set

- In a dataset, a pair of input and output $(x^{(i)}, y^{(i)})$ in supervised learning is called a **training example.**
- A training set, D, is a collection of training examples:
  - D = $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ or
  - D = $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, \mathrm{N}\}$
- i indicates the ith training example and N is the total number of training examples.
- In unsupervised learning, the training example is $(x^{(i)})$ and the training set is:
  - D = $\{(x^{(i)})\}_{i=1}^{N}$

Algorithms

- An algorithm is a step-by-step procedure for solving a problem.
- Algorithms used in data science originated from different fields, including data mining and machine learning.
- Algorithms used in data science are iterative.
- These iterative algorithms automate the process of searching for an optimal solution for a given data problem.
- Algorithms are used to learn (or search for) optimal solutions.

The End

Types of Learning Tasks

Types of Learning Tasks

What types of tasks can be achieved in data science?

- **Description:** involve the description of patterns and trends in the dataset
- **Estimation:** to approximate the numeric value of a target or output variable; for example, estimate GPA given SAT scores
- **Classification:** to approximate the categorical value of a target variable
- **Prediction:** to approximate the numeric value of target variable that lie in the future; for example, predict stock prices for the next week

Types of Learning Tasks (cont.)

- **Clustering:** grouping observations into classes or clusters of similar objects. This is similar to classification, but no target variable is used
- **Anomaly detection:** involves predicting if a data point is an outlier compared to other data points in the dataset
- **Association:** involves finding the relationship between attributes by establishing an if-then rule
- Association can be used to find out which items in a supermarket are purchased together

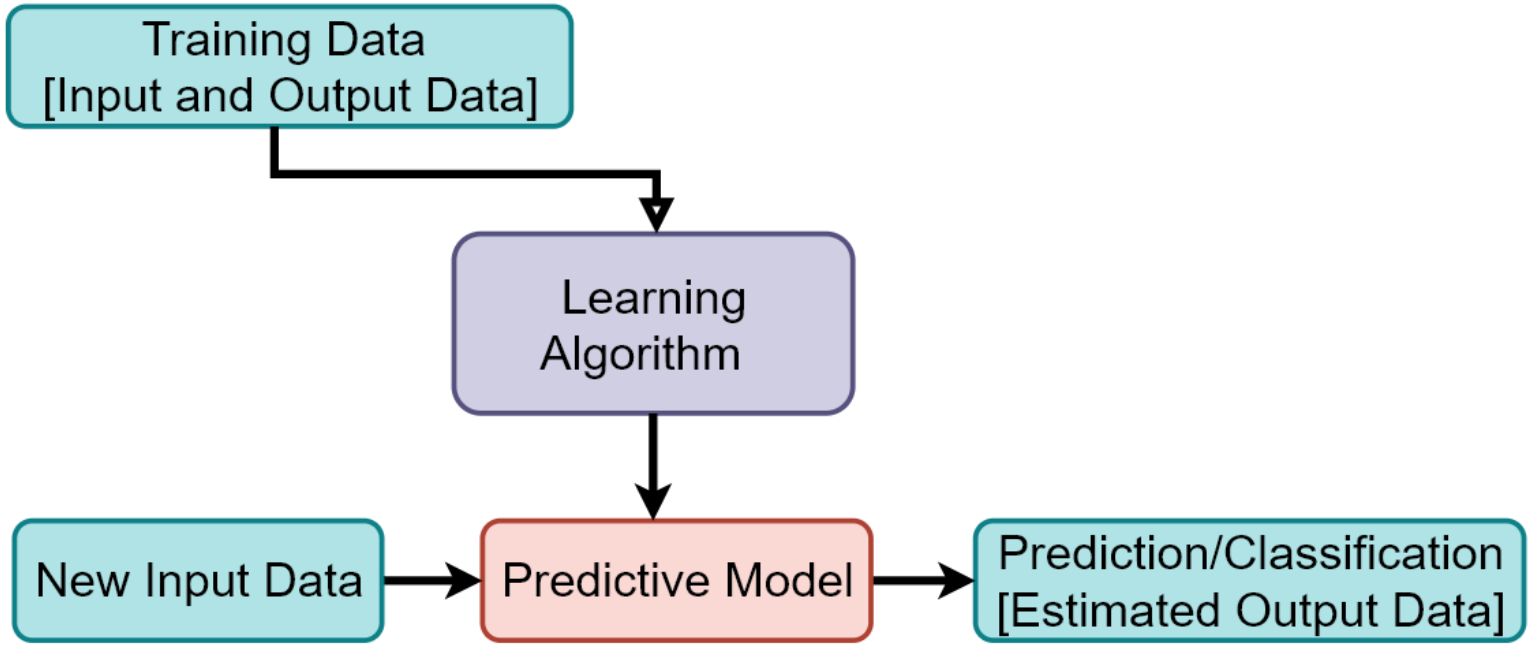
Types of Learning Problems

- Data science actually starts with asking a question or defining a problem that can be addressed using data science tools.
- There are different types of problems that can be solved with data science tools.
- This course focuses on supervised and unsupervised learning problems.
  - The goal of supervised learning is to build predictive models.
  - Unsupervised learning focuses on learning the structure in the dataset.

Supervised Learning

- Supervised learning uses a training dataset containing input and output values to build a model that is later used for predicting the output values of new input data
- Supervised learning can be further divided into regression and classification supervised learning
  - Supervised regression models predict continuous outcomes
  - Supervised classification models predict categorical outcomes

Supervised Learning (cont.)

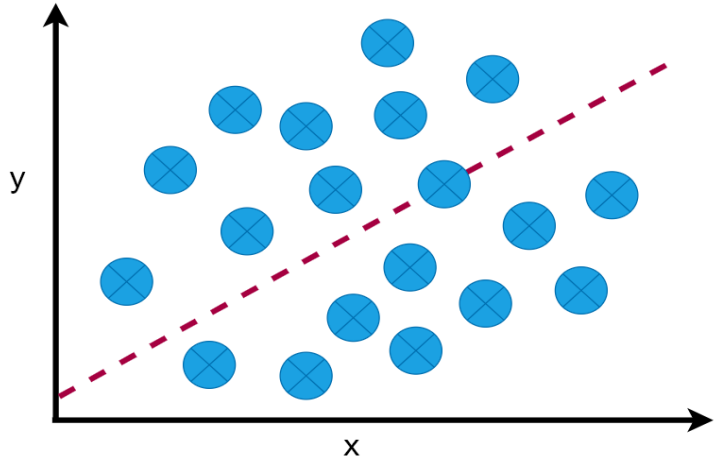


Supervised Learning (cont.)

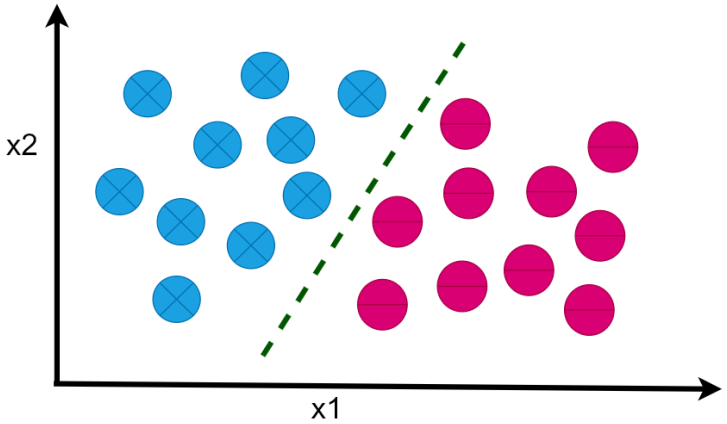
- Commonly used supervised learning approaches for solving supervised learning problems include:
  - Decision trees
  - Random forest
  - K-nearest neighbor
  - Support vector machines
  - Naïve Bayes
  - Linear regression
  - Polynomial regression
  - Logistic regression
  - Neural networks
  - And others

Supervised Learning (cont.)

An example of supervised regression

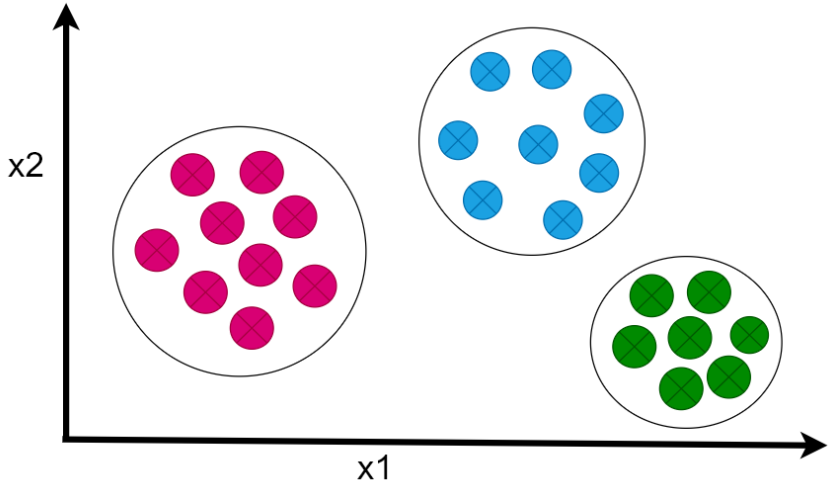


An example of a supervised classification



Unsupervised Learning

- Unsupervised learning involves finding the hidden structure in unlabeled data.
- An example of unsupervised learning is clustering. This approach groups objects based on their similarity on the input data.



The End