Introduction to Data Mining

Instructor: Junghye Lee

Department of Industrial Engineering

Ulsan National Institute of Science and Technology, Korea

junghyelee@unist.ac.kr

What is Data Mining?

- The **efficient discovery** of previously unknown, valid, potentially useful, understandable **patterns** in **large datasets**

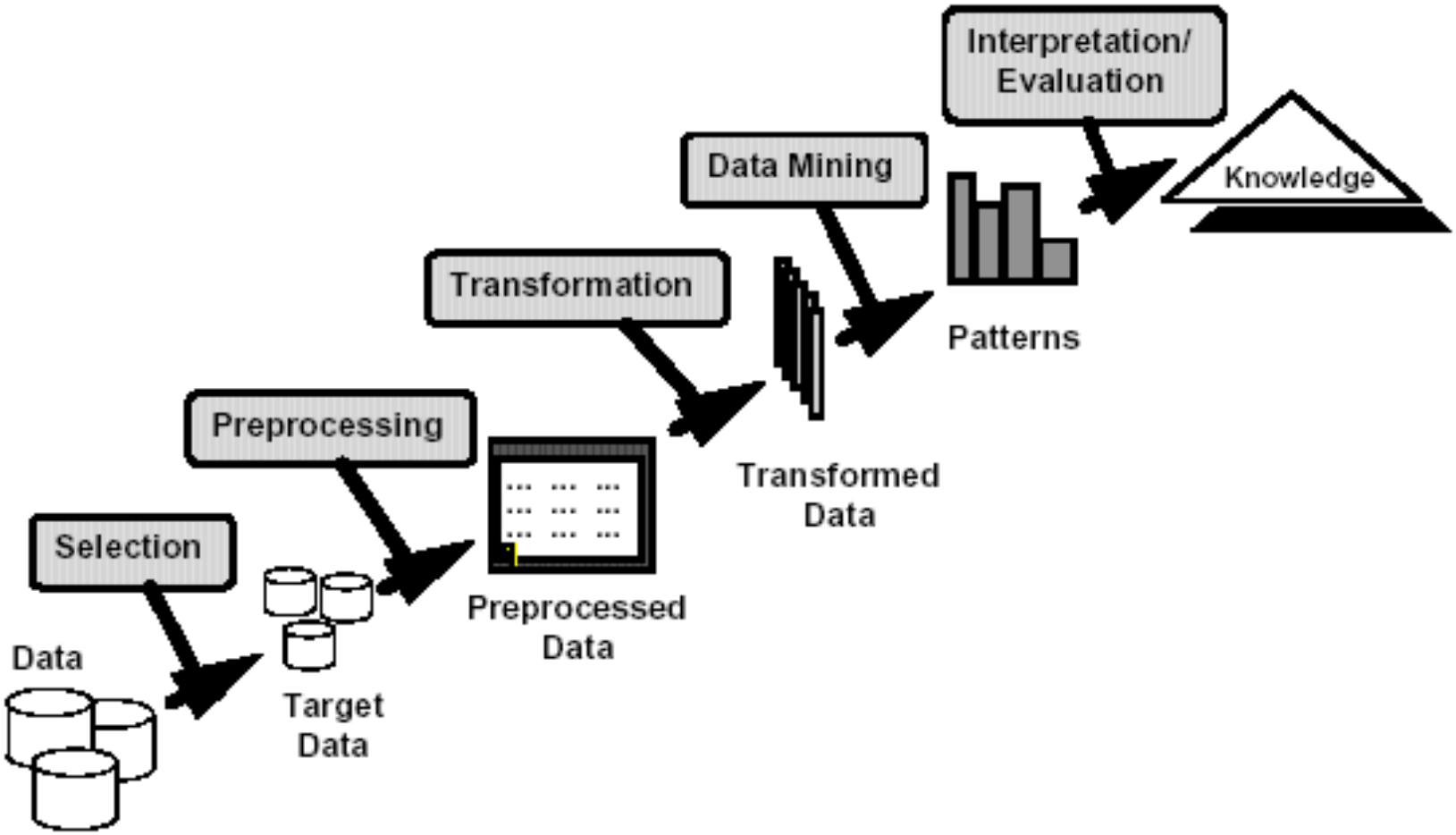
- The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner

Data Mining

- Process of semi-automatically analyzing large databases to find patterns that are:
  - **valid**: hold on new data with some certainty
  - **novel**: non-obvious to the system
  - **useful**: should be possible to act on the item
  - **understandable**: humans should be able to interpret the pattern

- Also known as Knowledge Discovery in Databases (KDD)

KDD


Interpretation/
Evaluation
Knowledge
Data Mining
Patterns
Transformation
Transformed
Data
Preprocessing
Preprocessed
Data
Selection
Target
Data
Data

KDD

- **Problem formulation**
- **Data collection**
  - subset data: sampling might hurt if highly skewed data
  - dimensionality reduction: principal component analysis (feature extraction), heuristic search (feature selection)
- **Pre-processing: cleaning**
  - name/address cleaning, different meanings (annual, yearly), duplicate removal, supplying missing values
- **Transformation:**
  - map complex objects e.g. time series data to features e.g. frequency
- **Choosing mining task and mining method**
- **Result evaluation and Visualization**

Applications

- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

How Data Mining is Used

- Identify the problem
- Use data mining techniques to transform the data into information
- Act on the information
- Measure the results

The Data Mining Process

1. Understand the domain
2. Create a dataset:
   - Select the interesting attributes
   - Data cleaning and pre-processing
3. Choose the data mining task and the specific algorithm
4. Interpret the results, and possibly return to 2

Why Data Pre-Processing?

- Data in the real world is dirty
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - ✓ Fill in missing values (e.g., average, min, max)
  - **noisy**: containing errors or outliers
    - ✓ Identify outliers and smooth out noisy data
  - **inconsistent**: containing discrepancies in codes or names
    - ✓ Correct inconsistent data

- No quality data, no quality mining results
  - **Garbage in garbage out**
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

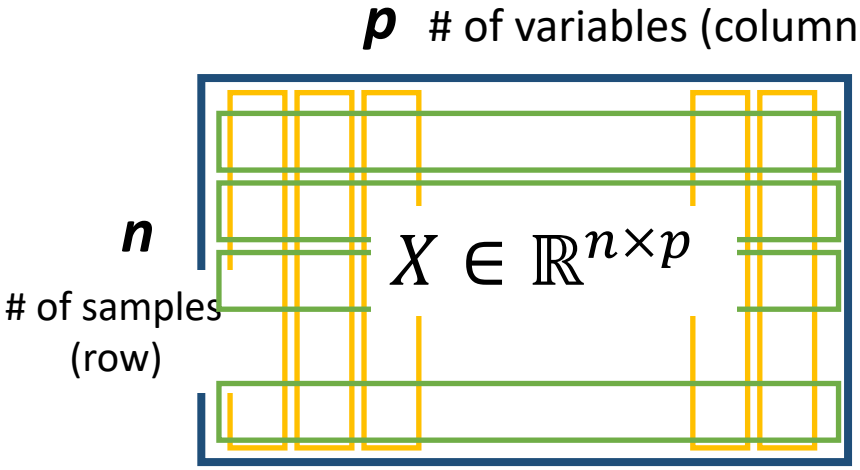
Data Mining Tasks

- Prediction Tasks (Supervised Learning)
  - Use **some variables (independent variables)** to predict unknown or future values of **other variables (dependent variable)**
- Description Tasks (Unsupervised Learning)
  - Find human-interpretable patterns that describe **the data (independent variables)**.

- Common data mining tasks
  - **Regression [Predictive]**
  - **Classification [Predictive]**
  - Clustering [Descriptive]
  - Association Rule Discovery [Descriptive]

How Data Looks Like?

- Data matrix ⟶ two modes; set of vectors



$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

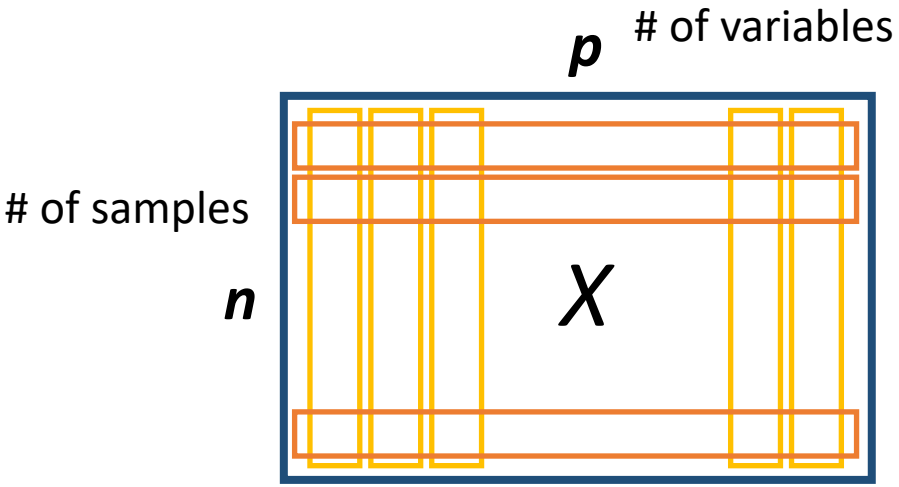
No.	Age	Weight	Exercise	Gender	Diseased?
1	67	60	Frequent	M	Yes
2	32	44	Seldom	F	Yes
3	14	50	Seldom	F	Yes
4	39	54	Seldom	F	Yes
5	24	48	Everyday	F	No
6	22	37	Everyday	M	No
7	46	68	Frequent	M	Yes
8	66	98	Everyday	M	No
9	19	42	Seldom	M	No

9 samples,
5 variables

9 x 5 matrix

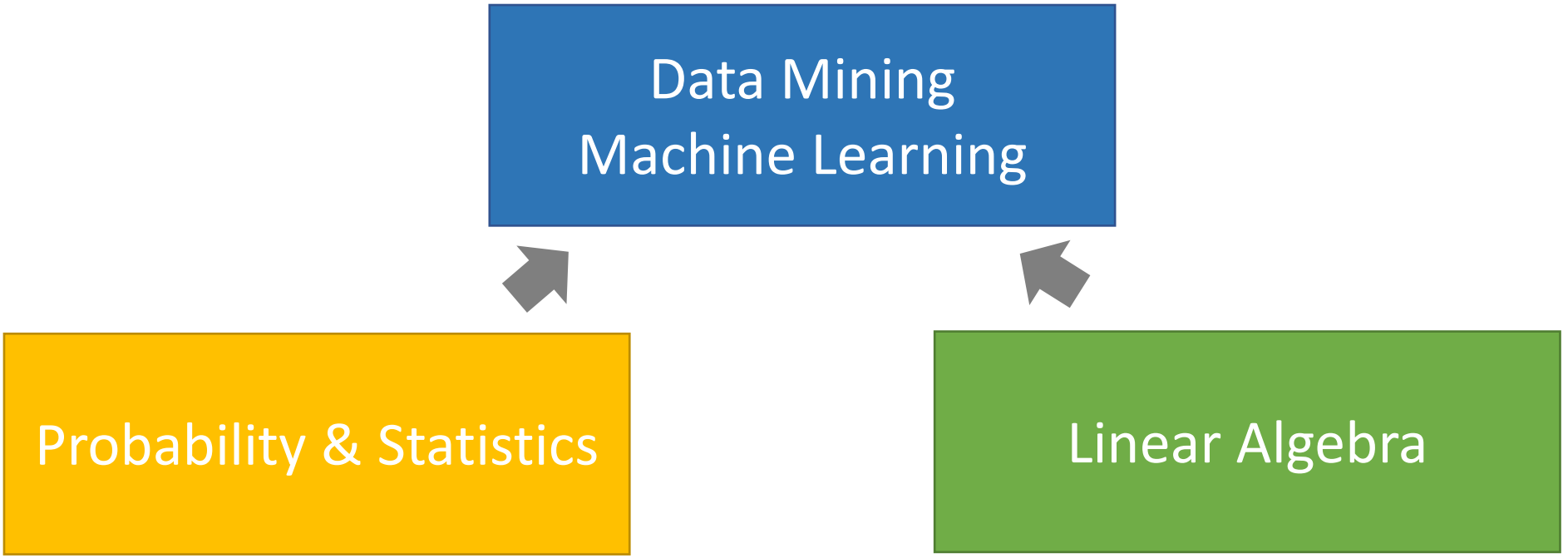
Data

- Data matrix = set of vectors
  - set of row vectors = set of column vectors



  - In terms of row vector: what is the dimensionality?
  - In terms of column vector: what is the dimensionality?
- In Data Science, we are mostly based on row vector.
  - When we say “the dimensionality of data” is “the length of each row vector”

How They are Related Each Other



Why We Learn about These?

- Learning the theoretical background for data science or machine learning can be a *daunting* (벅찬, 주눅이 들게 하는) experience, as it involves multiple fields of mathematics.
- In this lecture, my goal is to let you know the **most basic knowledge** to build the **mathematical background necessary to get up** and running in data science practical/research work.
- These suggestions are derived from my own experience in the data science field, and following up with the latest resources suggested by the community.

Reference: https://towardsdatascience.com/mathematics-for-data-science-e53939ee8306

Why Linear Algebra in Data Mining?

- Its concepts are a crucial prerequisite for understanding the theory behind Data Mining.
- You don't need to understand *Linear Algebra* before getting started in *Data Mining*, but at some point, you may want to gain a better understanding of how the *different algorithms* really work under the hood.

Reference: https://medium.com/@rathi.ankit/linear-algebra-for-data-science-a9648b9daee0

Linear Algebra

- **Linear algebra** is the branch of **mathematics** concerning **1) linear equations** such as

$$a_1x_1 + a_2x_2 + \cdots + a_px_p = b, \text{ and}$$

2) linear functions such as

$$(x_1, \ldots, x_p) \mapsto a_1x_1 + a_2x_2 + \cdots + a_px_p,$$

and **3) their representations through matrices and vector spaces**.

- Linear algebra is central to almost all areas of mathematics and engineering areas, because it allows **modeling many natural phenomena**, and **efficiently computing with such models**.
- For nonlinear systems, which cannot be modeled with linear algebra, linear algebra is often used as a first-order approximation.

Vector Arithmetic

- Data usually comes with a matrix form.
- Data is a set of vectors.
- It is basically based on vector arithmetic.
  - Addition (subtraction)
  - Multiplication
  - Inner product
  - Norm
  - Normalization (standardization)
  - Distance
- Later, you can generalize this into matrix or tensor arithmetic.

Data Mining Tasks

- Prediction Tasks (Supervised Learning)
  - Use **some variables (independent variables)** to predict unknown or future values of **other variables (dependent variable)**
- Description Tasks (Unsupervised Learning)
  - Find human-interpretable patterns that describe **the data (independent variables)**.

- Common data mining tasks
  - **Regression [Predictive]**
  - **Classification [Predictive]**
  - **Clustering [Descriptive]**
  - **Association Rule Discovery [Descriptive]**

Notation

- **Independent variable**
  - Predictor
  - Explanatory variable
  - Factor
- **Dependent variable**
  - Response
  - Target
  - Outcome

- **Matrix:** bold and capital letter
  - e.g.) **X**
- **Vector:** bold and small letter
  - e.g.) **y**
- **Value:** not bold and small letter
  - e.g.) (x, y) = (2, 3)
  - i.e., scalar

c.f.) **A random variable:** not bold and capital letter e.g.) $X_1, X_2, \ldots X_p$

Notation

No.	Age	Weight	Exercise	Gender	Diseased?
1	67	60	Frequent	M	Yes
2	32	44	Seldom	F	Yes
3	14	50	Seldom	F	Yes
4	39	54	Seldom	F	Yes
5	24	48	Everyday	F	No
6	22	37	Everyday	M	No
7	46	68	Frequent	M	Yes
8	66	98	Everyday	M	No
9	19	42	Seldom	M	No

- This is for prediction analysis, especially, classification problem (why?).
  - Independent variables ($\boldsymbol{X} \in \mathbb{R}^{n \times p}$): Age, Weight, Exercise, Gender
  - Dependent variable ($\boldsymbol{y} \in \mathbb{R}^{n \times 1}$): Diseased? (class variable)
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$
  - where $n = 9, p = 4$.
- $\boldsymbol{X}$ is a matrix and $\boldsymbol{y}$ is a vector.
  - $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}]$ and $\boldsymbol{x_j}$ $(j = 1, \dots, 4)$ is a vector (i.e., $\boldsymbol{x_j} \in \mathbb{R}^{nx1}$)
  - x_{ij} is a scalar, for example, $x_{31} = 14, x_{42} = 54$.
    - *i* represents a sample (row), *j* represents a variable (column).

Variable Type

- Continuous/ordinal/binary/nominal/numeric/discrete
- **Category 1:** Numerical vs. nominal
- **Category 2:** Continuous vs. discrete

What type of each feature?

	continuous and numeric	continuous and numeric	ordinal and numeric, discrete	binary and nominal, discrete	binary and nominal, discrete
No.	**Age**	**Weight**	**Exercise**	**Gender**	**Diseased?**
1	67	60	Frequent	M	Yes
2	32	44	Seldom	F	Yes
3	14	50	Seldom	F	Yes
4	39	54	Seldom	F	Yes
5	24	48	Everyday	F	No
6	22	37	Everyday	M	No
7	46	68	Frequent	M	Yes
8	66	98	Everyday	M	No
9	19	42	Seldom	M	No

Predictive Analysis (Supervised Learning)

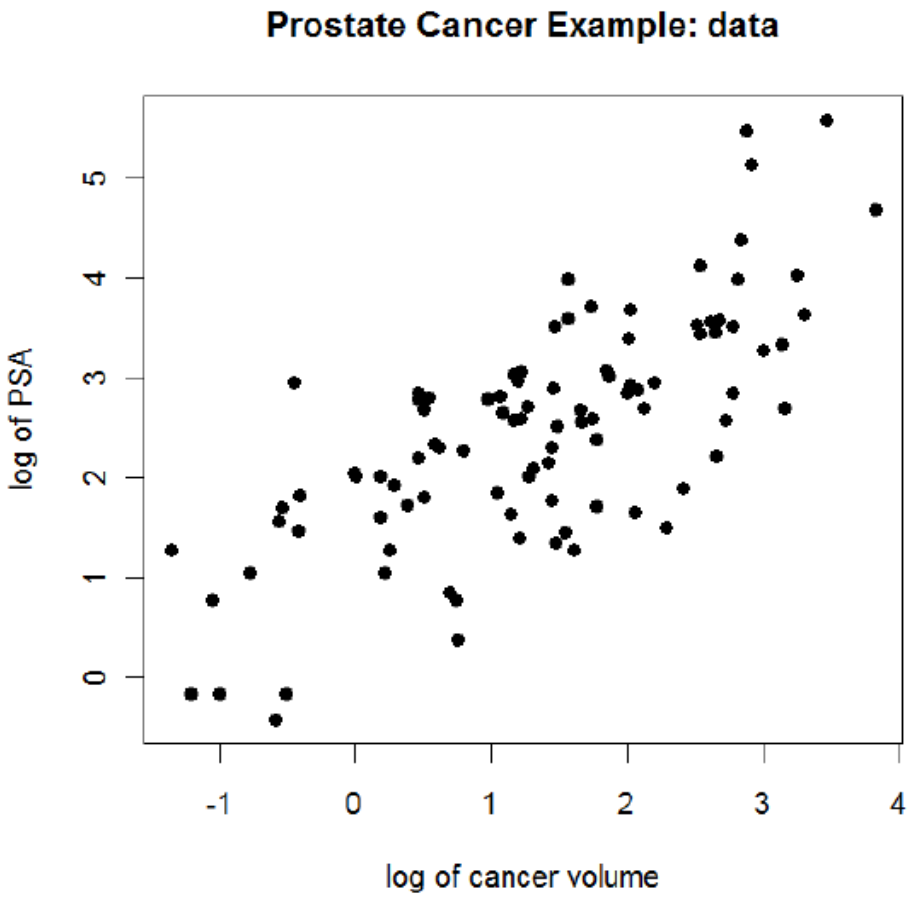
What is Regression?

- Given data on predictor variables (inputs, $\boldsymbol{X}$) and a **continuous response variable** (output, Y) build a model for:
  - Predicting the value of the response from the predictors.
  - Understanding the relationship between the predictors and the response.

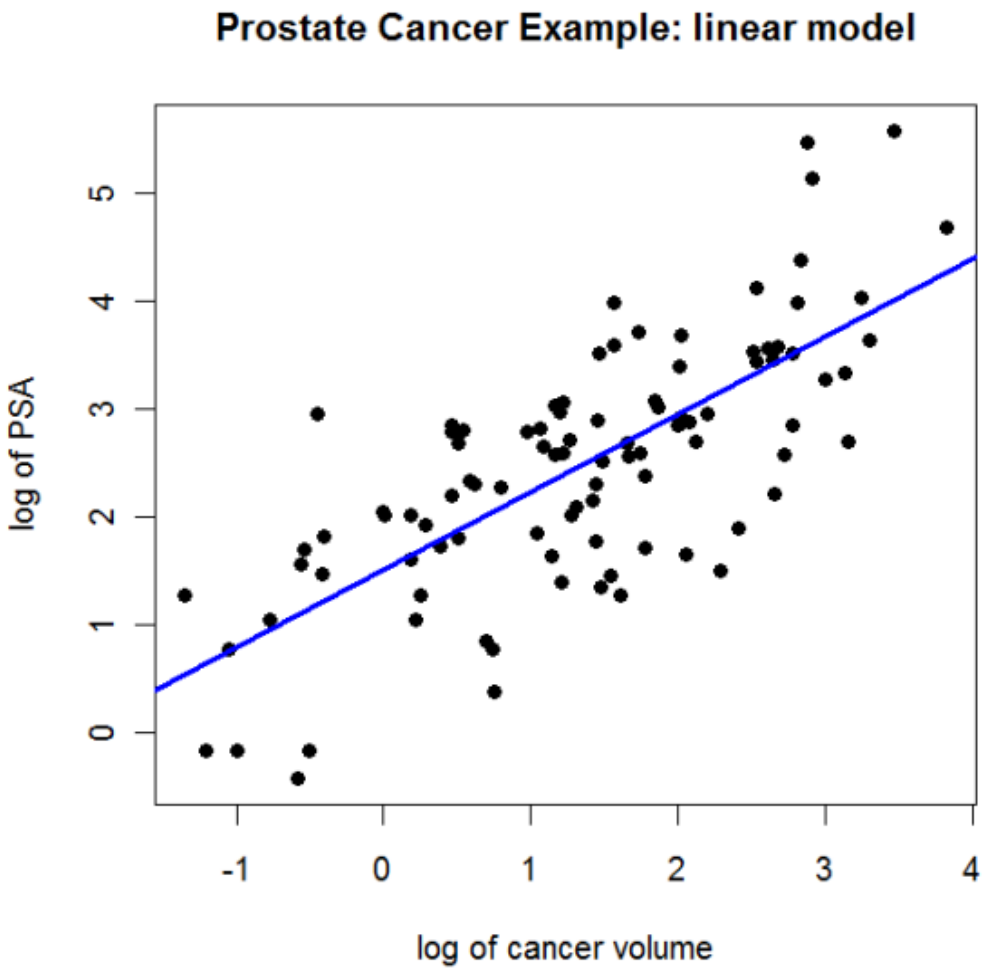
- E.g.) predict a person's **systolic blood pressure** based on their age, height, weight, etc.

Regression Picture

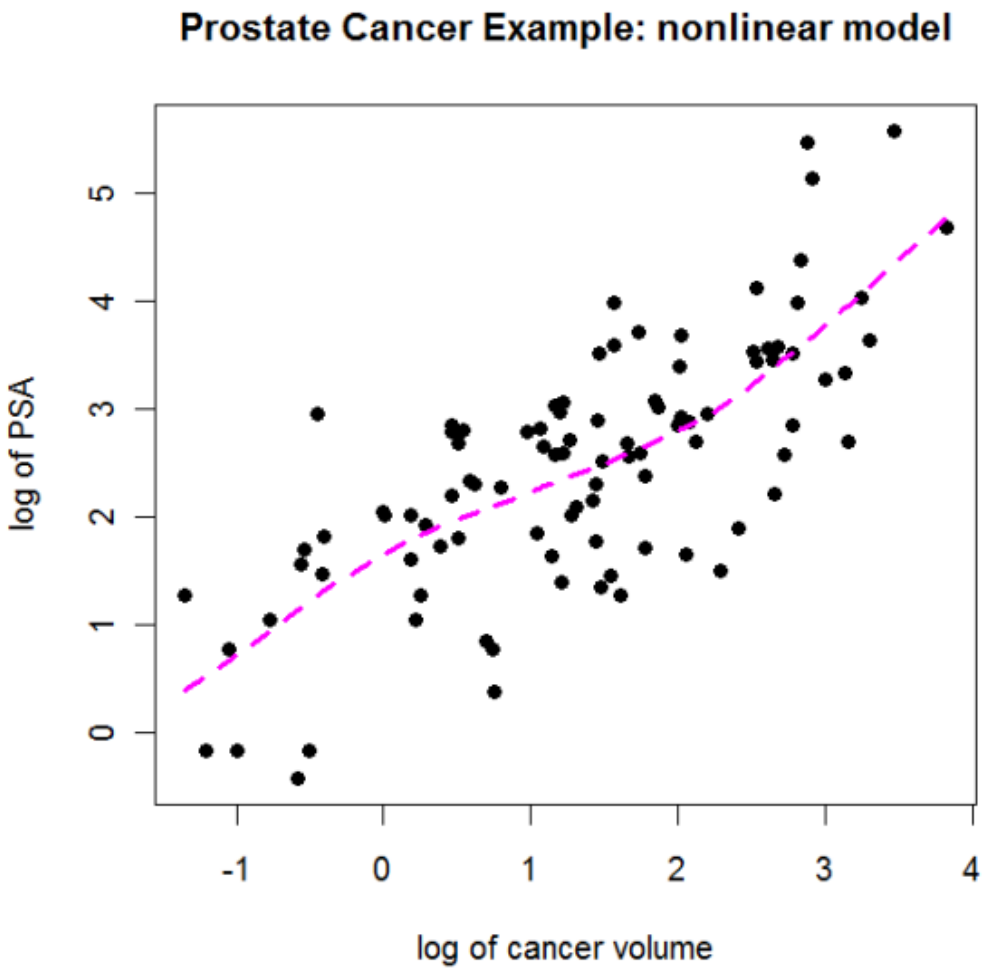
- Prostate-Specific Antigen (PSA)



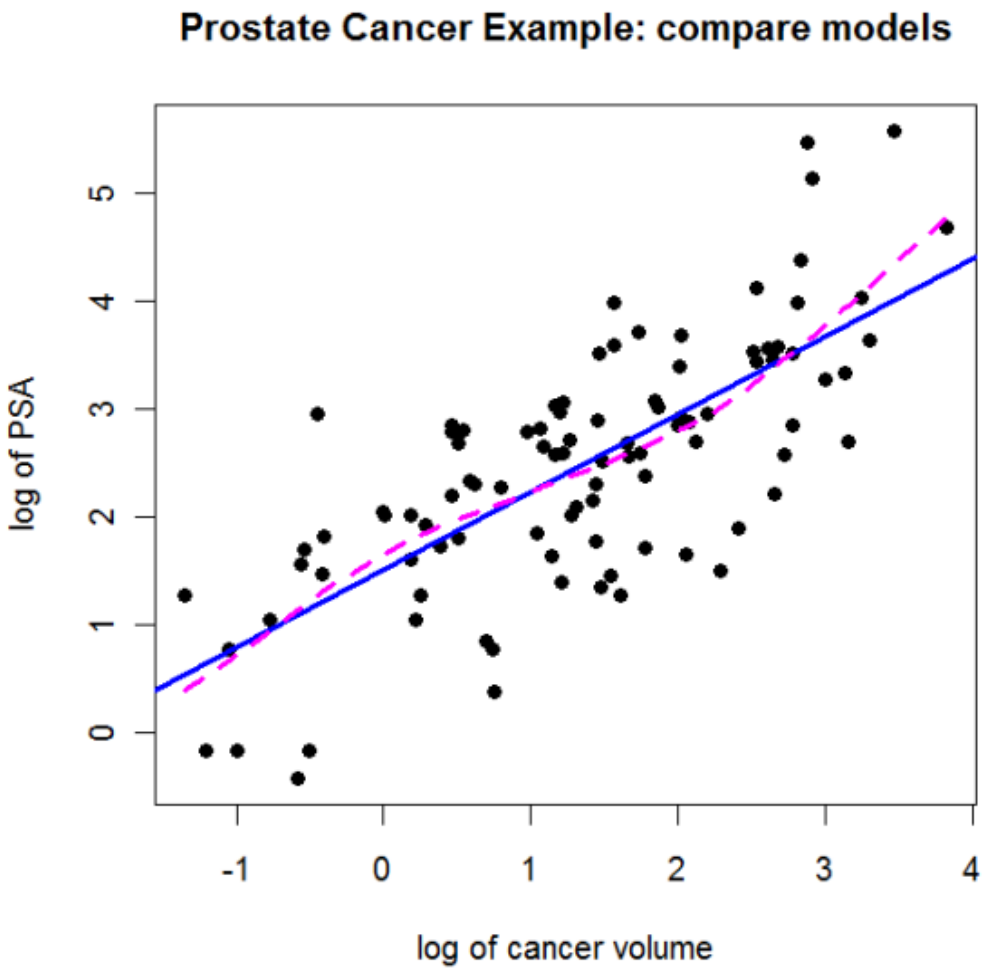
Regression Picture



Regression Picture



Regression Picture



What is Classification?

- Given data on predictor variables (inputs, $\boldsymbol{X}$) and a **categorical response variable** (output, Y) build a model for:
  - Predicting the value of the response from the predictors.
  - Understanding the relationship between the predictors and the response.

- E.g.) predict **a person's 5-year-survival (yes/no)** based on their age, height, weight, etc.

Classification Examples

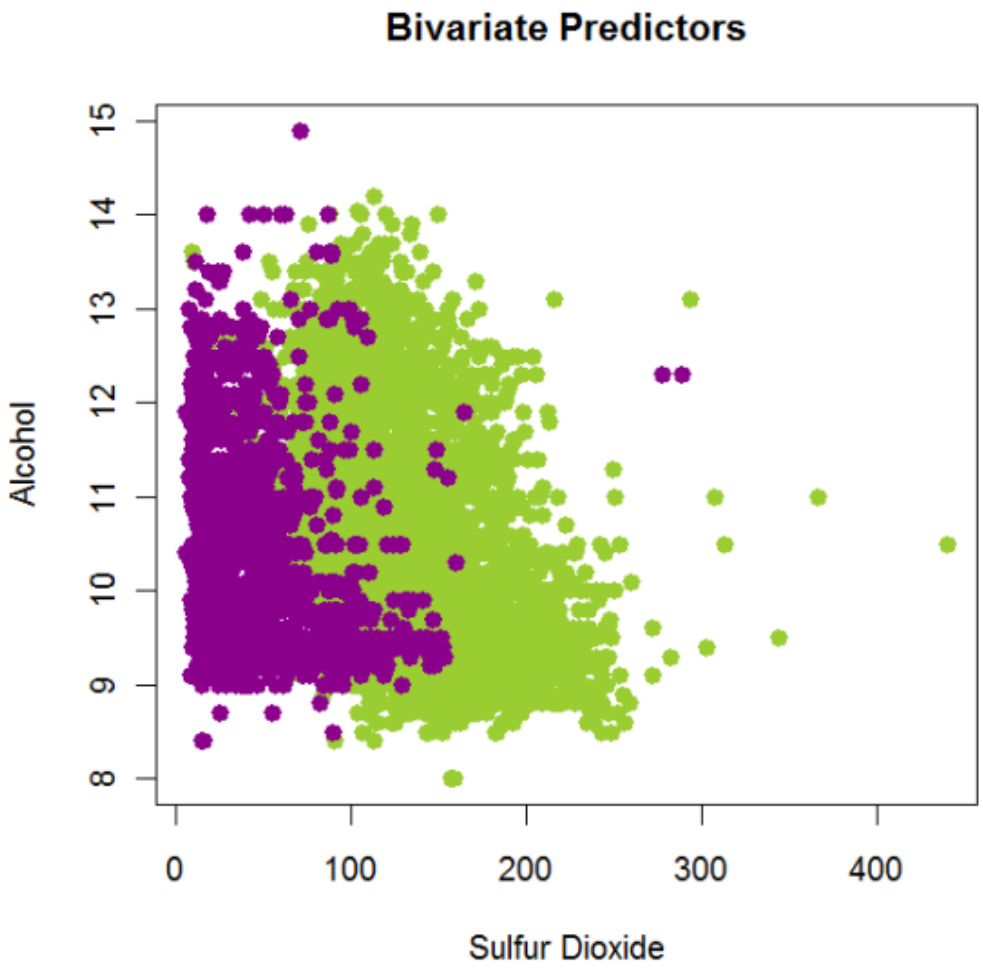
- Y: presence/absence of disease
- $\boldsymbol{X}$: diagnostic measurements

- Y: land cover (grass, trees, water, roads…)
- $\boldsymbol{X}$: satellite image data (frequency bands)

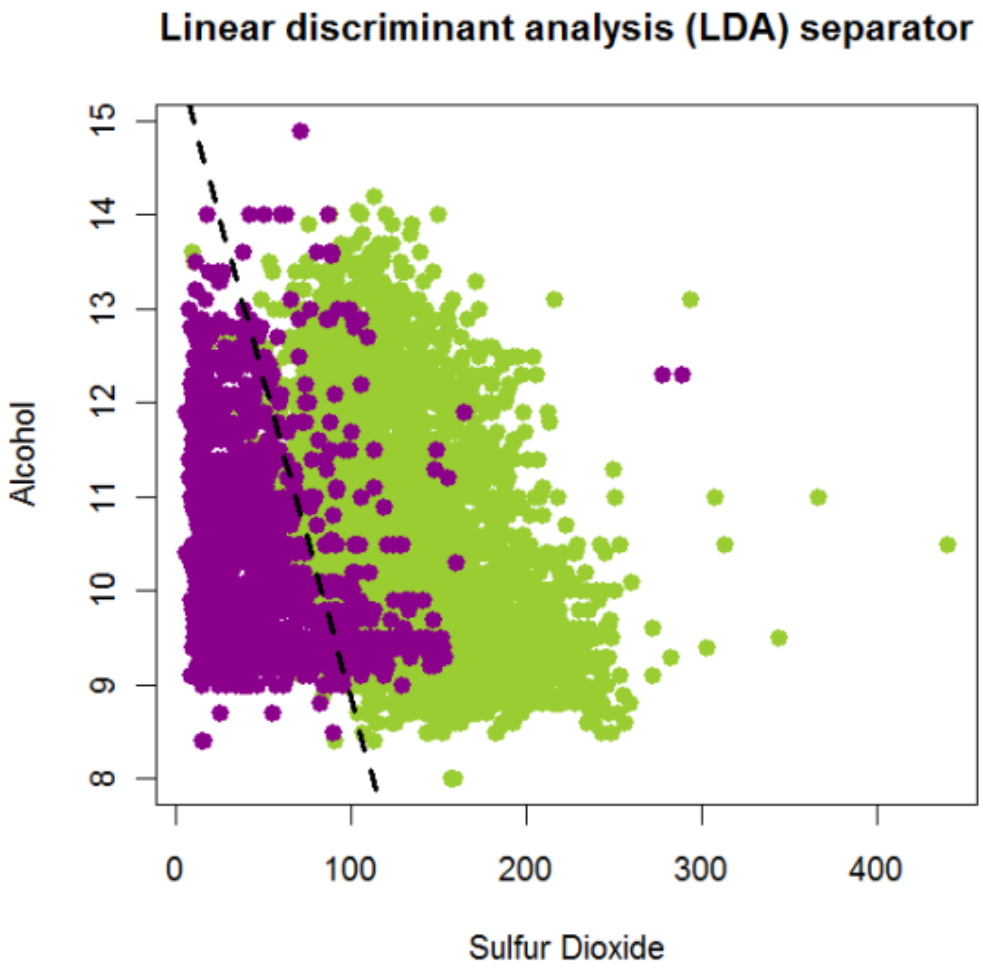
- Y: loan defaults (yes/no)
- $\boldsymbol{X}$: credit score, own or rent, age, marital status, …

- Y: dementia status
- $\boldsymbol{X}$: scores on a battery of psychological tests

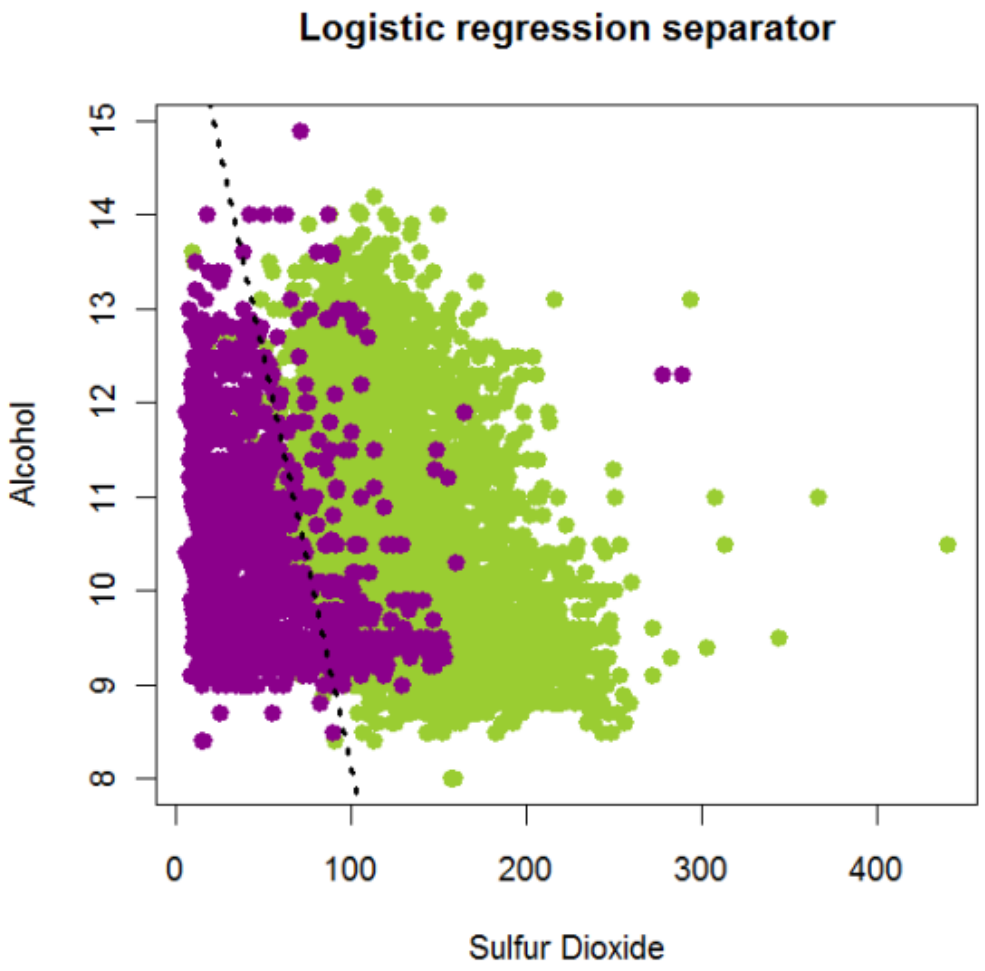
Classification Picture



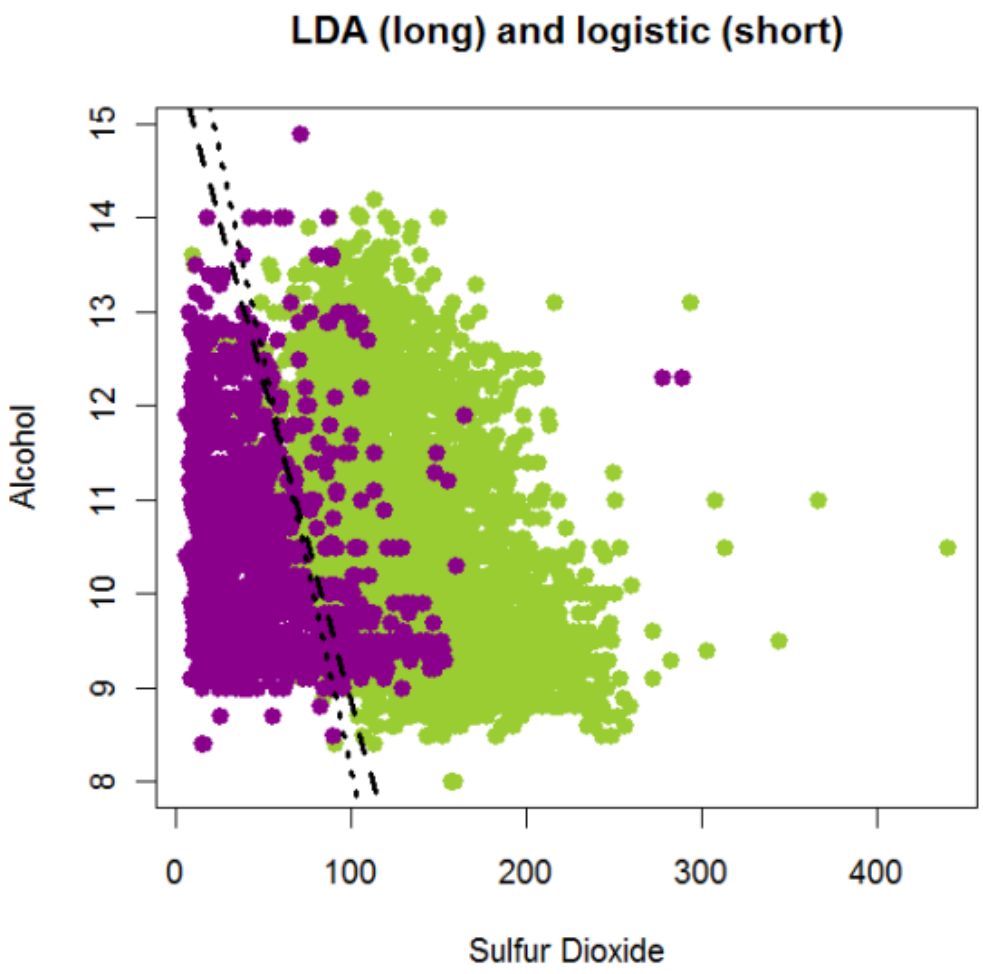
Classification Picture



Classification Picture



Classification Picture

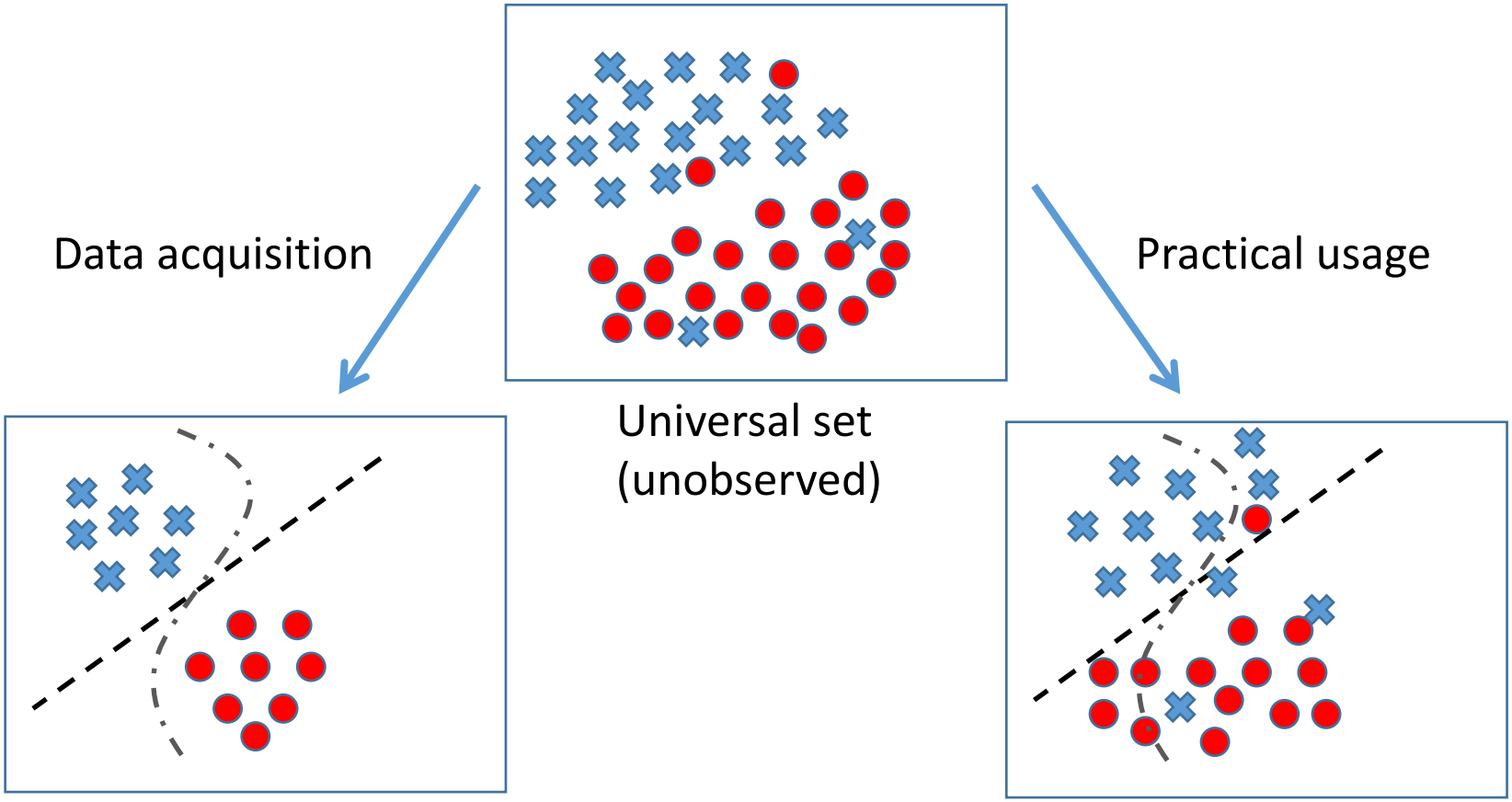


Regression and Classification

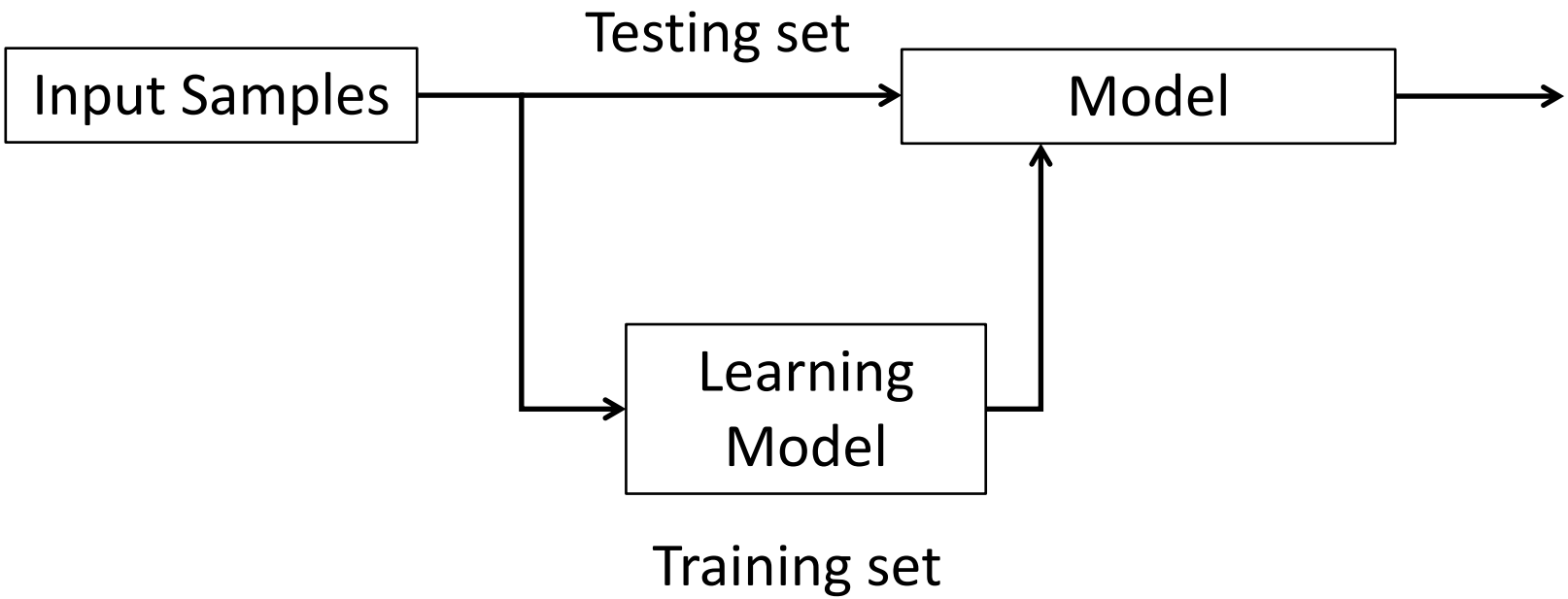
- Given data $D = \{(\boldsymbol{x}_i, y_i),\ \ i = 1, \dots, n\}$
  - where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$, build a model $\hat{f}$ so that
  - $\hat{y} = \hat{f}(\boldsymbol{X})$ for random variables $\boldsymbol{X} = (X_1, \dots, X_p)$ and Y.

- Then f-hat will be used for:
  - Predicting the value of the response from the predictors: $\hat{y} = \hat{f}(\boldsymbol{x}_0)$ where $\boldsymbol{x}_0 = (x_{01}, \dots, x_{0p})$.
  - Understanding the relationship between the predictors and the response.

Training and Testing



Learning System Model



Descriptive Analysis (Unsupervised Learning)

Clustering Definition

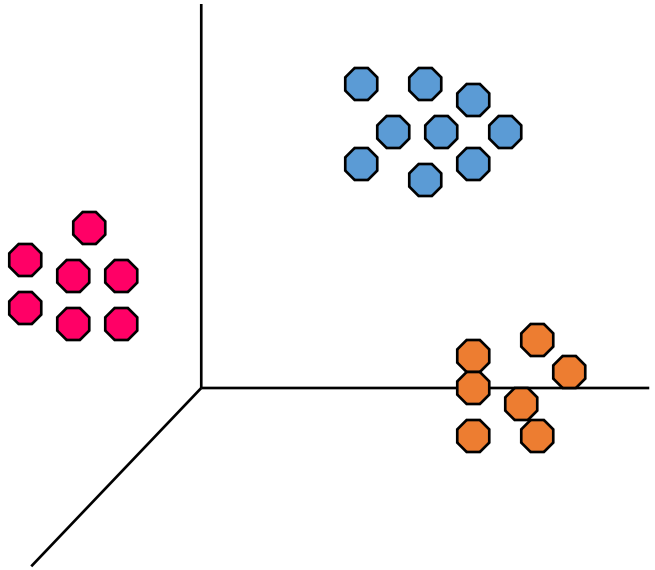
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.

- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

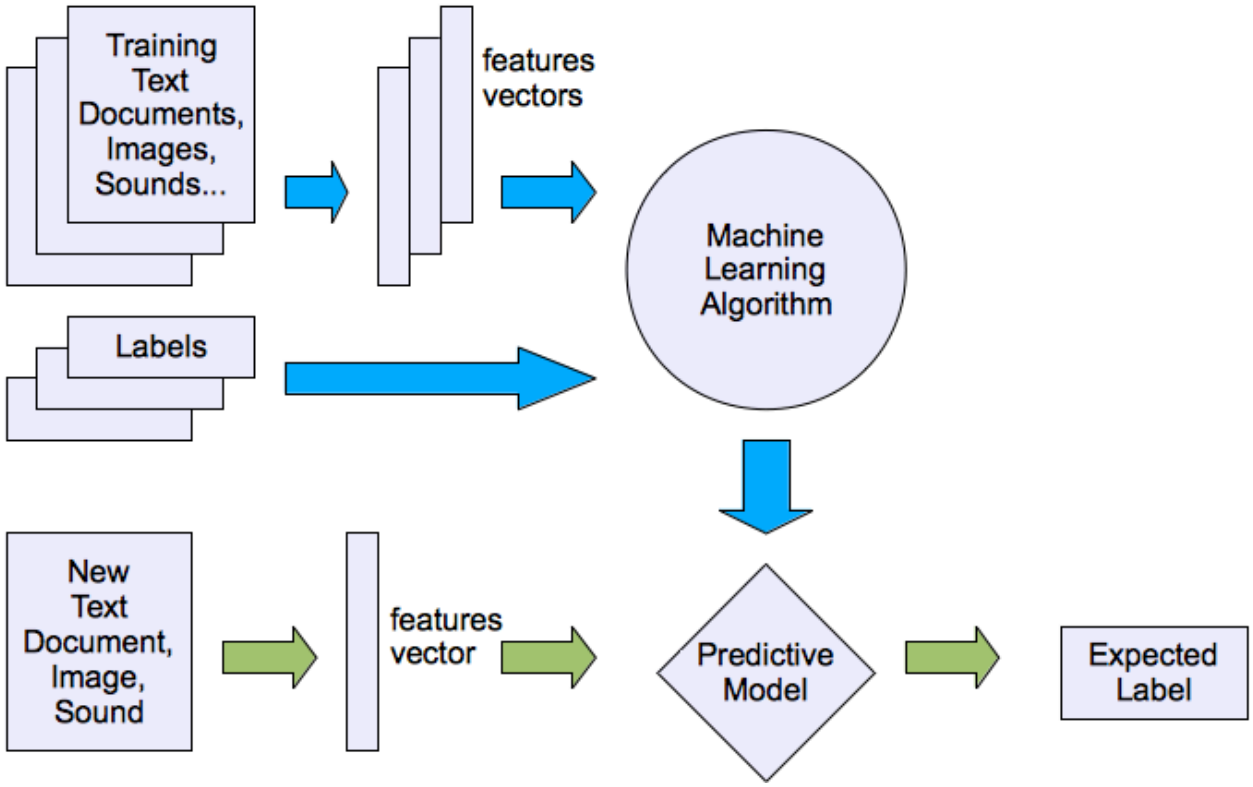
Intra-cluster distances are minimized

Inter-cluster distances are maximized

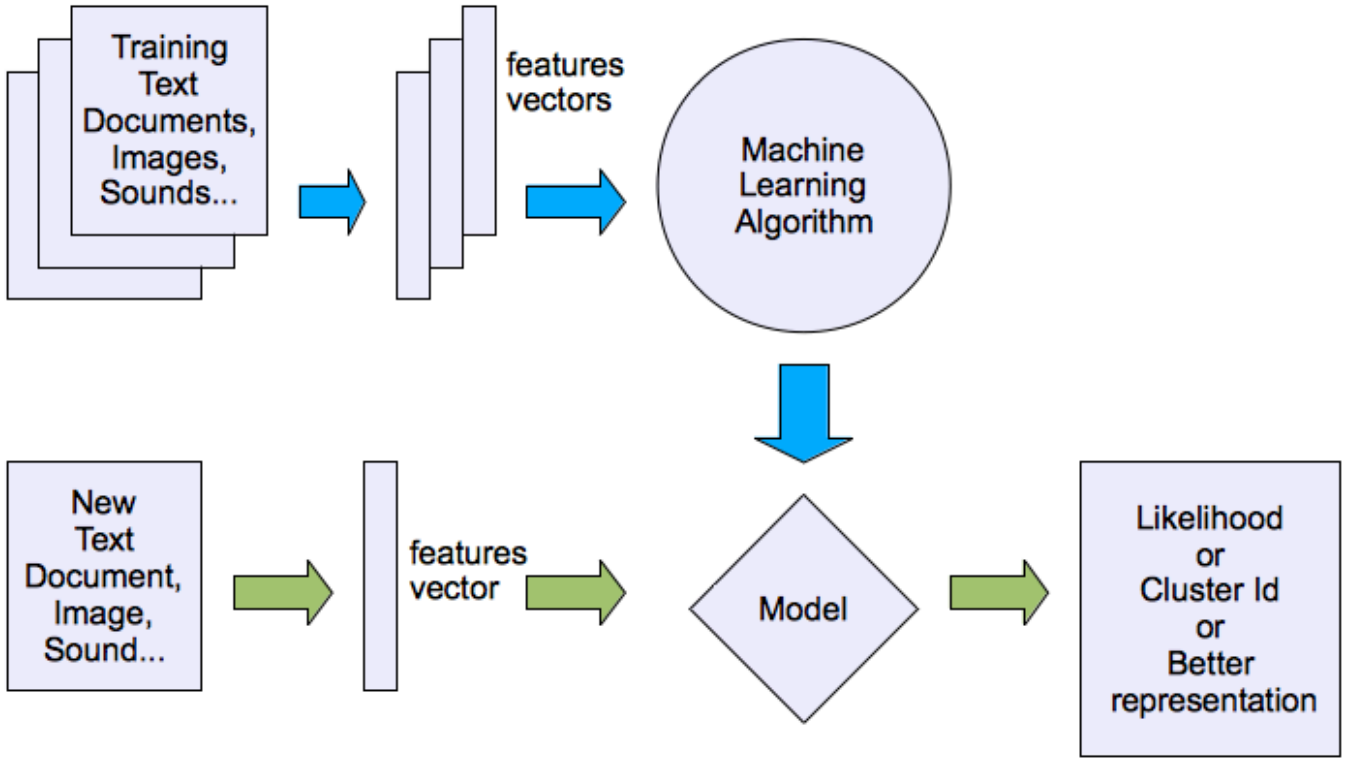


Euclidean distance based clustering in 3D space

Supervised Learning



Unsupervised Learning



Q&A