POLITECNICO DI MILANO

Pattern Analysis and Machine Intelligence

Course Introduction

Prof. Matteo Matteucci

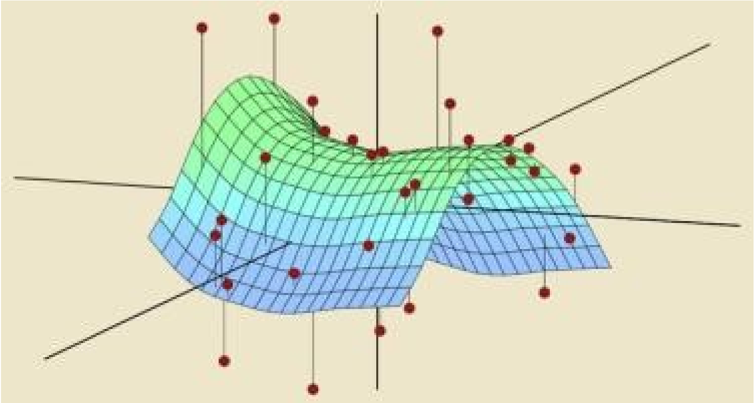
- Last year we had a novel course edition …
  - Same name & program
  - New teachers (more or less)
  - New book (somehow easier)
  - New approach (more practical)

- It worked out !!! At least in my opinion …

- Lectures given by:
  - Matteo Matteucci (Lecturer)
    matteo.matteucci@polimi.it
  - Davide Eynard (Teaching Assistant)
    davide.eynard@polimi.it

PAMI Course Program and Schedule

- Topics covered during the course
  - Statistical Learning Theory
  - Regression methods
  - Classification methods
  - Clustering methods

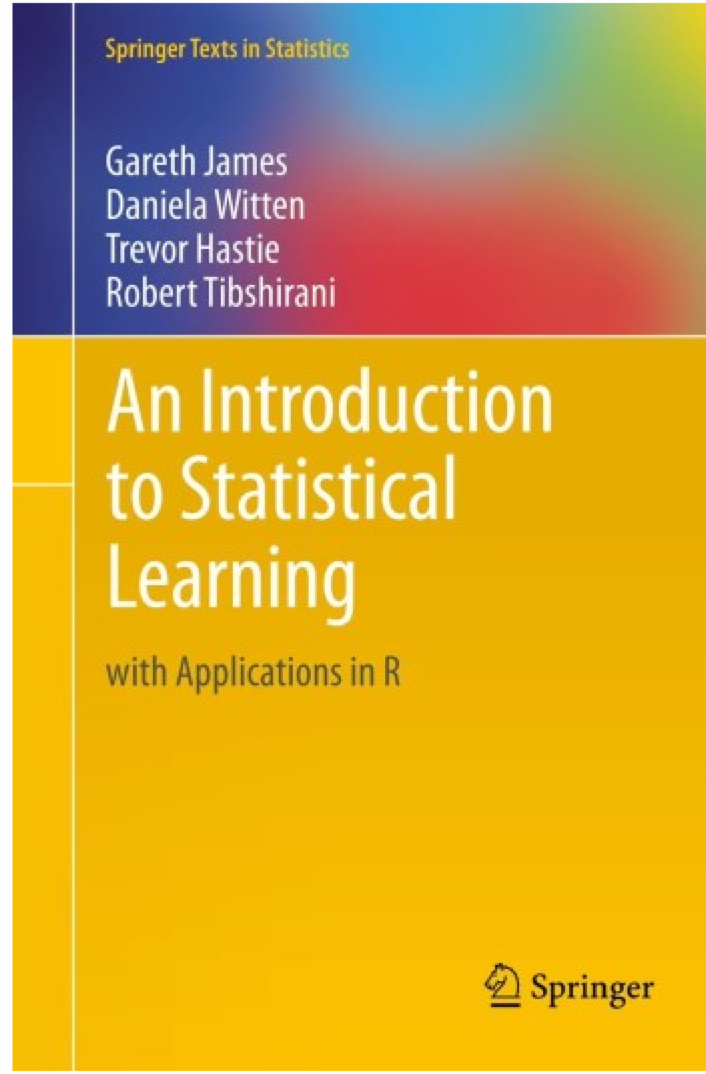


- Course Organization
  - For each topic, on average: 6hL + 4hE + 2hLab
  - Lab in class with R, you can use your laptop
- Detailed Schedule (Tentative) available online (**soon!**)

http://chrome.ws.dei.polimi.it/index.php/Pattern_Analysis_and_Machine_Intelligence

Course Material

- Book for this year edition
  - Same authors of ESL (the best!)
  - Easier than ESL
  - Practical perspective
  - Labs and Exercises using the R language
  - Available online as pdf
    www.statlearning.com
- Slides from the teachers
  - Taken from the book
  - Published after the lectures
- Contributed material (TBD)
  - Notes, exercises and questions from students? [google docs]



Exam Rules for PAMI

- The exam grade is composed by a written exam & homework
  - WE - Written Exam is compulsory
    - 2 Theoretical questions + 2 Practical Exercises (no R code)
    - Covering the 4 topics previously introduced
    - Possibly sampled from the suggested ones (if more than 20 exist)
  - HW - Home Works are NOT compulsory, but …
    - … they help you in understanding the course and prepare WE
    - … they get you in the "nitty gritty" of algorithm implementations
    - … they can only increase the final grade!
  - Final GRADE combines the two
    - GRADE = MAX(WE, 0.7*WE + 0.3HW)
    - Grade get frozen for DAR students … but there should not be here!

DAR = IRDM + PAMI

- Data Analysis and Retrieval (**only for Management!**) [10 CFU]
  - Information Retrieval and Data Mining [5 CFU]
  - Pattern Analysis and Machine Intelligence [5 CFU]
- Teachers
  - DAR: M. Matteucci (L) (**only for Management!**)
  - IRDM: M. Matteucci (L) + L. Bondi (TA)
  - PAMI: M. Matteucci (L) + D. Eynard (TA)
- Exam grading in DAR (**only for Management!!!**)
  - DAR = (PAMI * 5 + IRDM * 5)/10 = (PAMI + IRDM) / 2
    - Both PAMI and IRDM have to be passed with grade ≥18
    - Both exams are on the same day one after the other
    - You can take them separately and register them later

Questions and Comments?

- …

Hypothesis behind Book & Course (1-2)

1. *Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.*
   - Rather than attempting to consider every possible approach (an impossible task), we concentrate on presenting the methods that we believe are most widely applicable.

2. *Statistical learning should not be viewed as a series of black boxes.*
   - No single approach will perform well in all possible applications. We attempt to carefully describe the model, intuition, assumptions, and trade-offs behind each of the methods that we consider.

Hypothesis behind Book & Course (3-4)

3. *While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine in the box!*
   - We minimize discussion of technical details related to fitting procedures and theoretical properties. We assume the reader is comfortable with basic mathematical concepts, but we do not assume a graduate degree in mathematical science.

4. *We presume that the reader is interested in applying statistical learning methods to real-world problems.*
   - We devote some time to R computer labs. In each lab, we walk through a realistic application of the methods considered during the lectures.

Three Reference Scenarios (Wage)

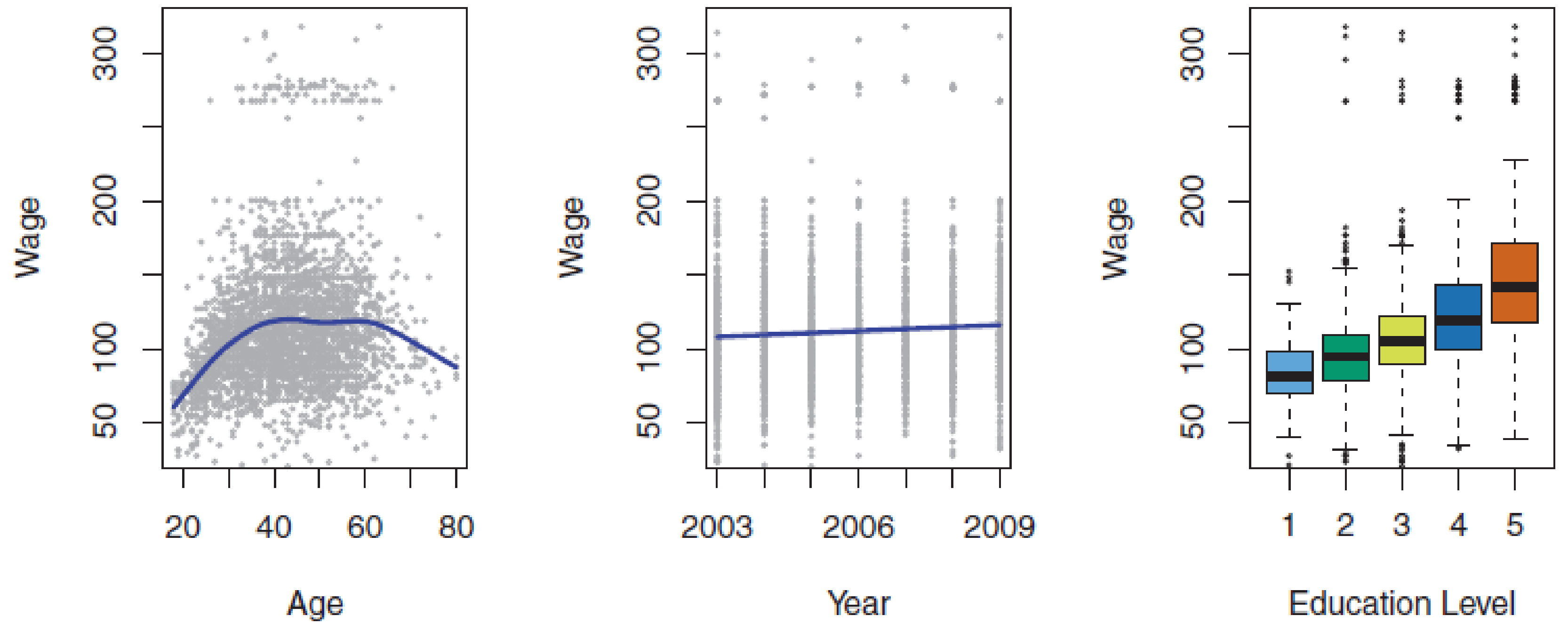
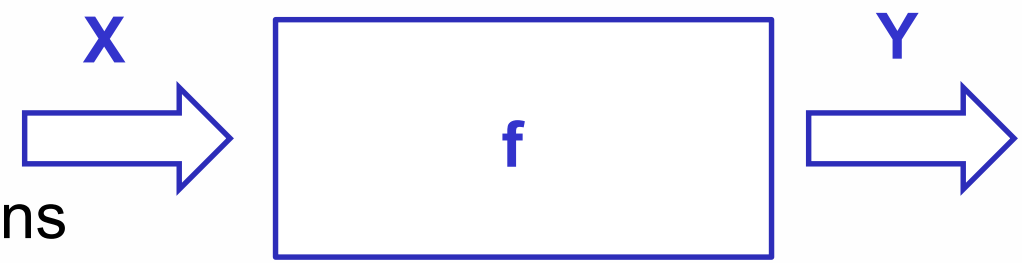


FIGURE 1.1. Wage *data, which contains income survey information for males from the central Atlantic region of the United States.* Left: wage *as a function of* age. *On average,* wage *increases with* age *until about 60 years of age, at which point it begins to decline.* Center: wage *as a function of* year. *There is a slow but steady increase of approximately* \$10,000 *in the average* wage *between* 2003 *and* 2009. Right: *Boxplots displaying* wage *as a function of* education, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

Terminology in Regression

- Input
- Predictor
- Observations
- Independent Variable



- Output
- Prediction
- Response
- Dependent Variable

- Model
- Function
- Inductive Hypothesis
- Learning Machine
- …

Three Reference Scenarios (Smarket)

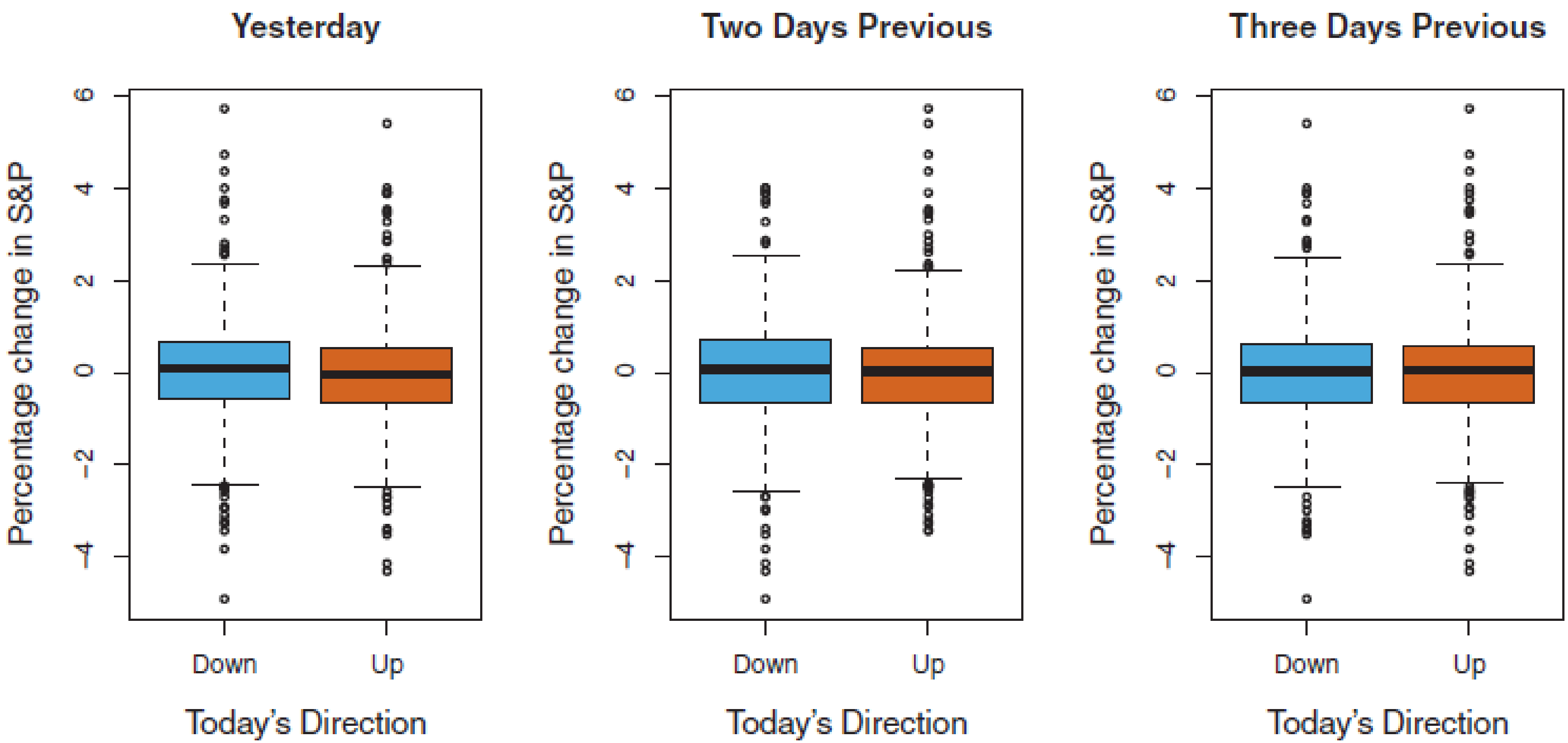


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* Smarket *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Three Reference Scenarios (Smarket)

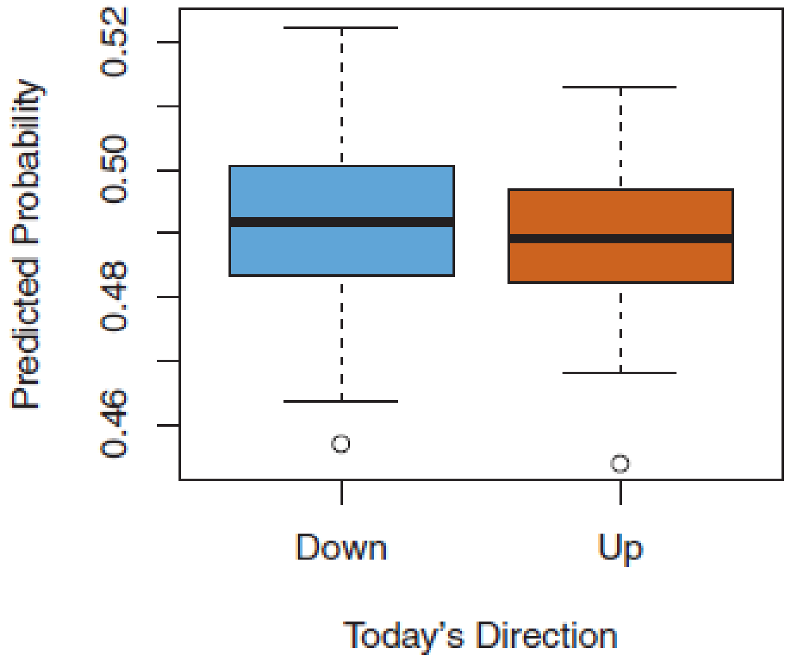
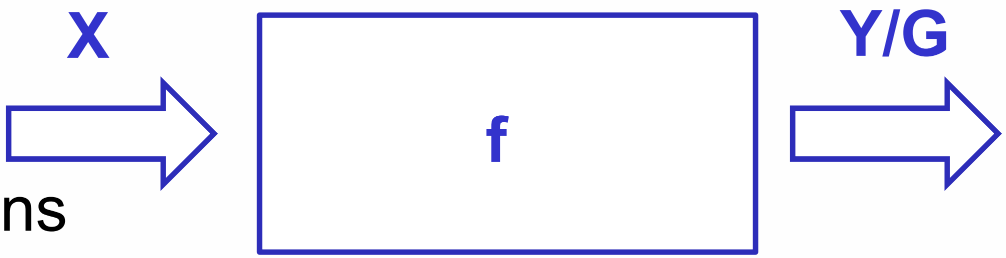


FIGURE 1.3. *We fit a quadratic discriminant analysis model to the subset of the* Smarket *data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.*

Terminology in Classification

- Input
- Features
- Observations
- Independent Variable



- Output
- Class
- Dependent Variable

- Classifier
- Inductive Hypothesis
- Learning Machine
- …

Three Reference Scenarios (NCI60)

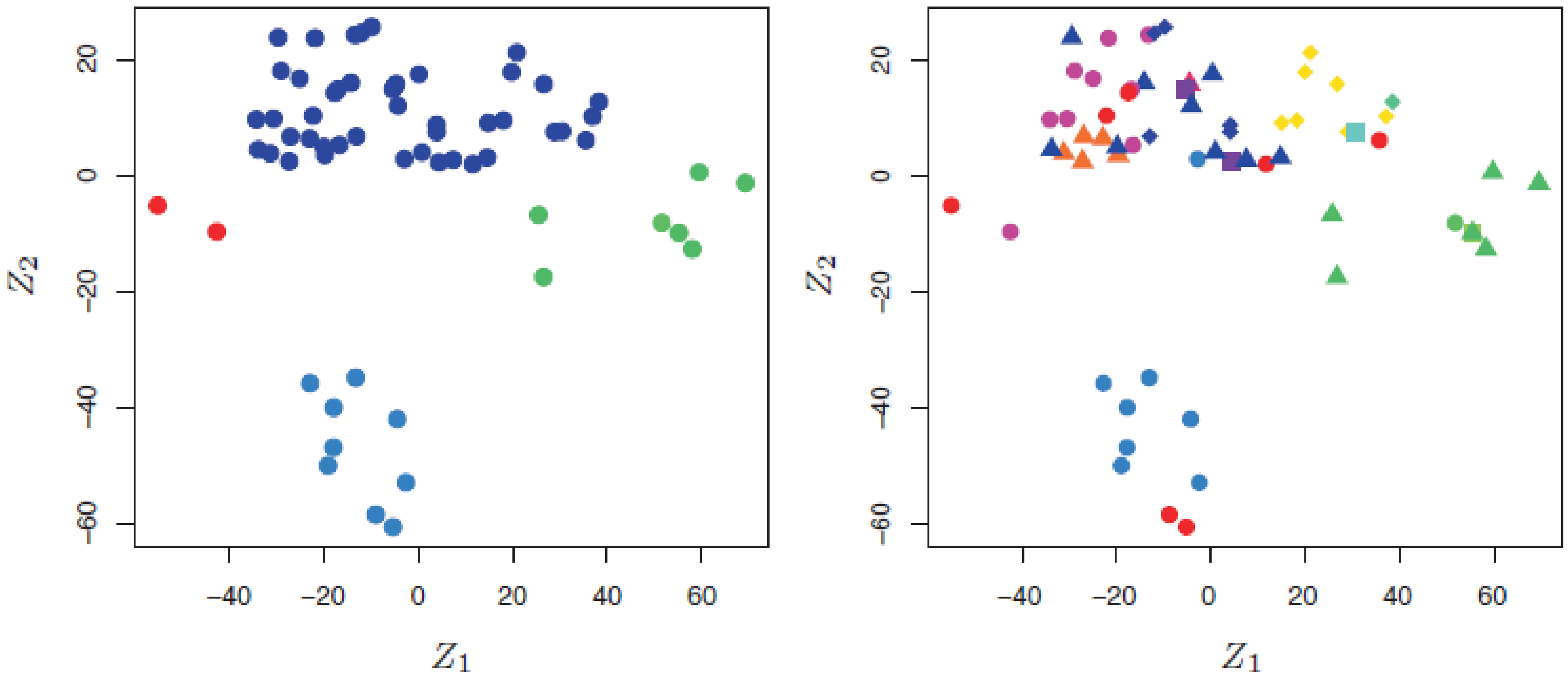


FIGURE 1.4. Left: *Representation of the* NCI60 *gene expression data set in a two-dimensional space,* Z_1 *and* Z_2*. Each point corresponds to one of the 64 cell lines. There appear to be four groups of cell lines, which we have represented using different colors.* Right: *Same as left panel except that we have represented each of the 14 different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.*

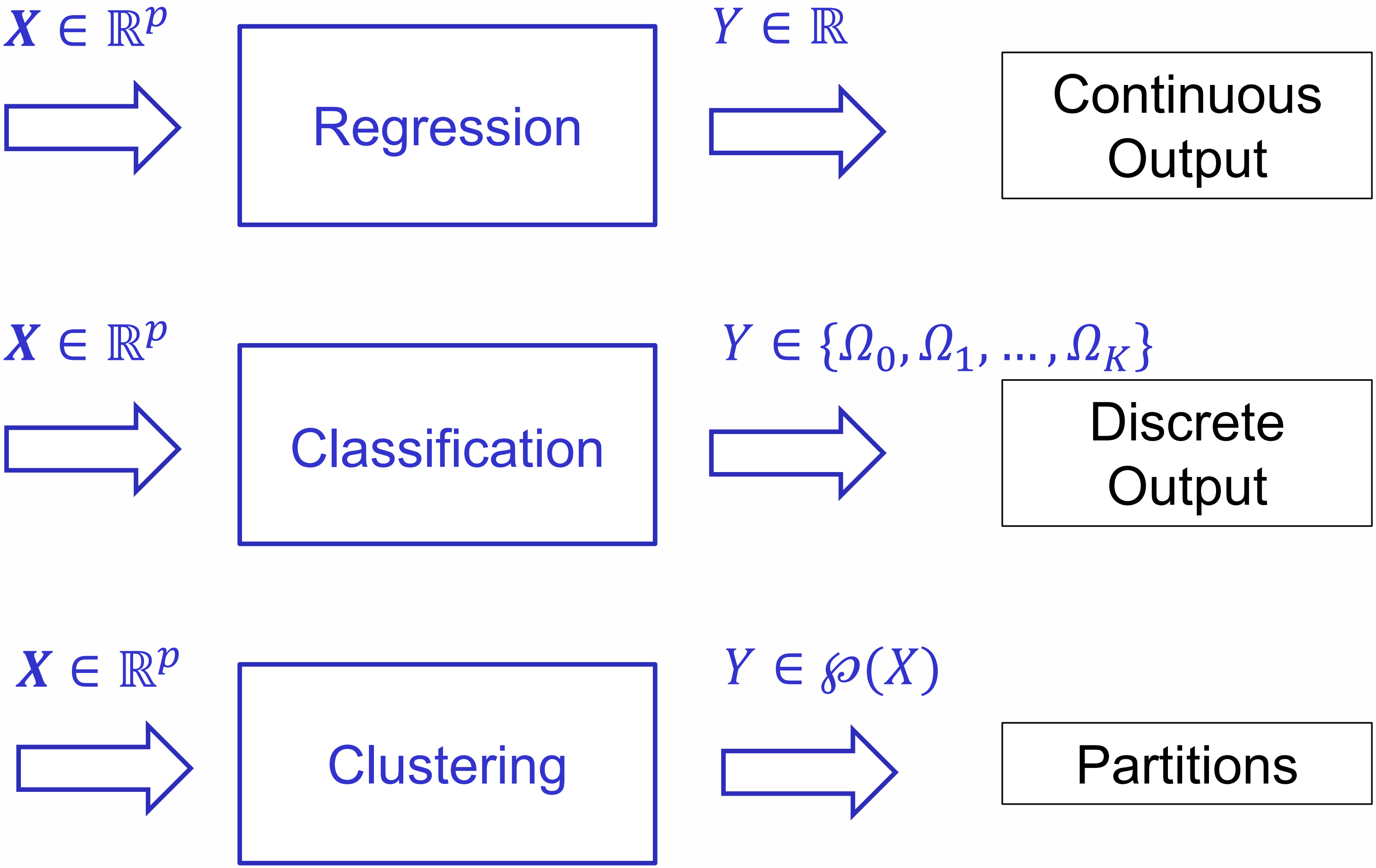
Terminology in Clustering

- Input
- Features
- Observations
- Data

- Clusters
- Groups
- Partitions
- Classes

- Clustering Algorithm
- …

Notation in Brief



Notation Expanded

- Input and Output Matrixes

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Data sample

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \Rightarrow \mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

- Dataset $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

Wrap up!

- Course bureaucracy
  - Material and Schedule
  - Exam and grading
  - DAR = IRDM + PAMI
- Course topics
  - Statistical learning
  - Regression
  - Classification
  - Clustering
- Notation
  - Input/Output, Feature/Class, Data/Clusters, …
  - Consistent matrix notation