PQHS 471
Machine Learning & Data Mining
Lecture 1: Introduction

Jan 17, 2023

Syllabus

- Instructor
  - Hao "Harry" Feng, Ph.D.
  - Assistant Professor, Population and Quantitative Health Sciences, Case School of Medicine
  - Email: hxf155@case.edu
  - Office hour: Tuesday 4-5pm over Zoom. Link on Canvas.
  - Office: SOM Wood Building WG-82T. In-person meeting by appointment.
- Research Interests
  - Biostatistics, Bioinformatics, High-throughput Data, *-omics*.
- Class Website
  - Canvas

Teaching Assistants

- Daoyu Duan
  - PhD student in Epidemiology and Biostatistics
  - Email: dxd429@case.edu
  - TA office hour: Monday 4PM-5PM
- Leslie Meng
  - PhD student in Epidemiology and Biostatistics
  - Email: gxm324@case.edu
  - TA office hour: Wednesday 4PM-5PM

Syllabus(cont'd)

- Class Time
  - Tuesday & Thursday, Jan/17/2023 - Apr/27/2023
  - 2:30PM - 3:40PM
- $6 \sim 7$ sessions will be hands-on programming lab sessions (laptop required).
- Key Dates
  - Jan 31, NO class
  - Feb 7, NO class
  - Mar 9, Mid-term exam
  - Mar 14, NO class (Spring break)
  - Mar 16, NO class (Spring break)
  - April 27 is the last day of class.

Syllabus(cont'd)

- Prerequisites:
  - PQHS 431: Statistical Methods I.
- Nice-to-haves:
  - Additional knowledge in Linear Algebra, Basic Probability and Statistical Inference are extremely helpful for a thorough understanding of algorithms introduced in this course.
- Evaluation:
  - Homework (65%): Six sets of unequally-weighted, applied, R programming-based assignments. Due dates and submission method will be announced with assignments. Late homework submission: 15% penalty upfront + additional 10% penalty each day.
  - Midterm exam (25%): In class, closed book.
  - Class participation (10%).

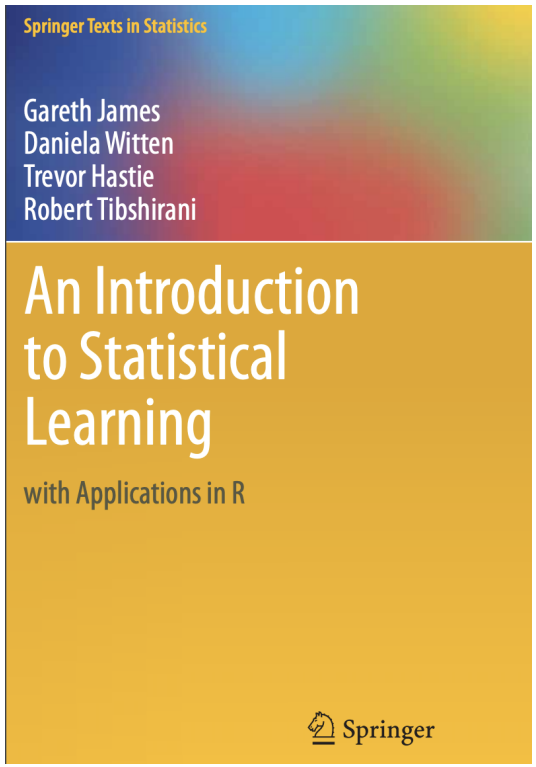
Academic Integrity

Students are expected to complete all homework assignments ALONE, without collaboration with others. However, consulting the TAs is allowed.
Students are expected to uphold standards of academic integrity.
Procedures will be taken for academic misconduct:

- https://case.edu/gradstudies/sites/case.edu.gradstudies/files/2018-04/SGS-Academic-Integrity-Policies-and-Rules.pdf
- https://case.edu/gradstudies/about-school/policies-procedures

Syllabus(cont'd)

- Textbook:
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
  - FREE! https://www.statlearning.com/s/ISLR-Seventh-Printing.pdf
  - Free PDF copy also available on Canvas.



Syllabus(cont'd)

Other references:

- The Elements of Statistical Learning (2nd edition) Hastie, Tibshirani and Friedman (2009). Springer-Verlag.
  FREE! https://web.stanford.edu/~hastie/ElemStatLearn/
- Computer Age Statistical Inference: Algorithms, Evidence and Data Science
  https://web.stanford.edu/~hastie/CASI/
- Deep Learning
  https://www.deeplearningbook.org/

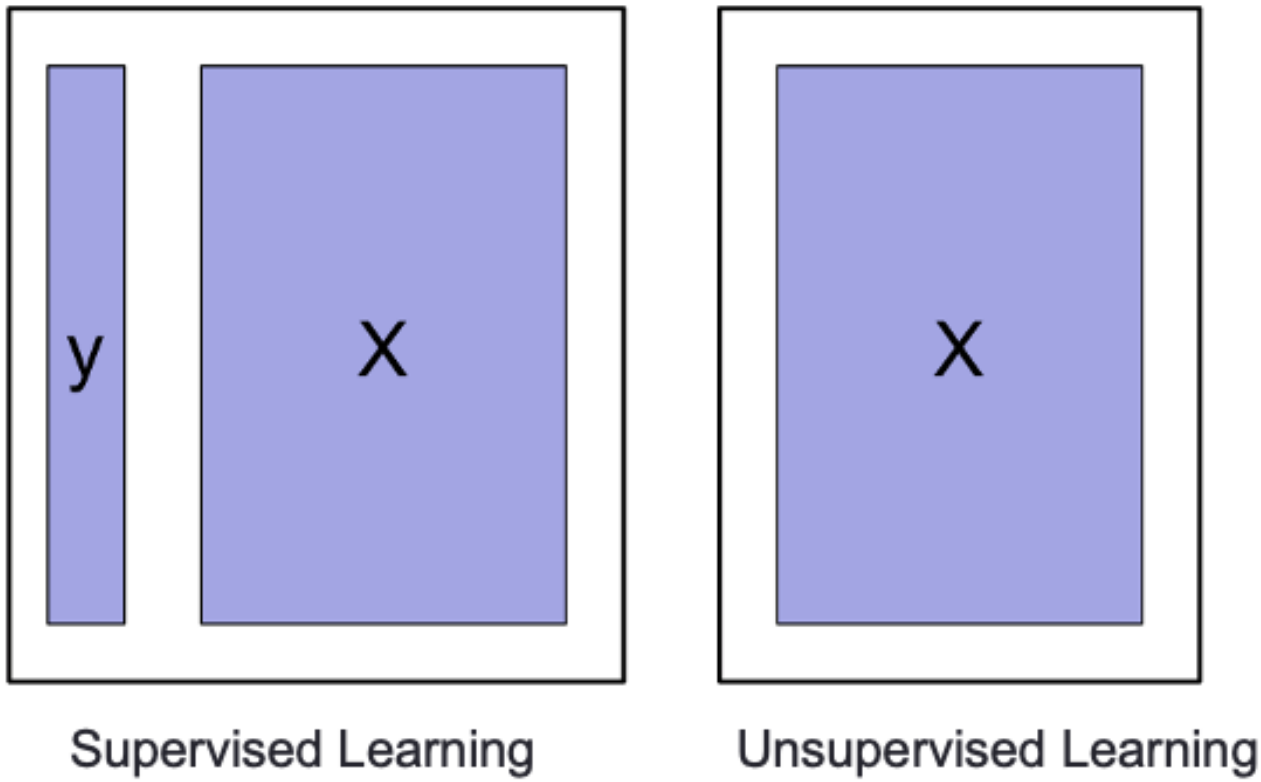
Syllabus(cont'd)

- Topics:
  - Dimension Reduction
  - Similarity measures, k-means, Hierarchical Clustering
  - Frequent Pattern Mining
  - Fundamentals in Supervised Learning
  - Decision Tree, Bayes Classifier, KNN
  - GLM, LDA, QDA
  - Cross-Validation and Bootstrap
  - Tree and Forest, Tree based method
  - Bagging, Boosting
  - Support Vector Machines
  - Neural Network
  - (intro) Deep Learning

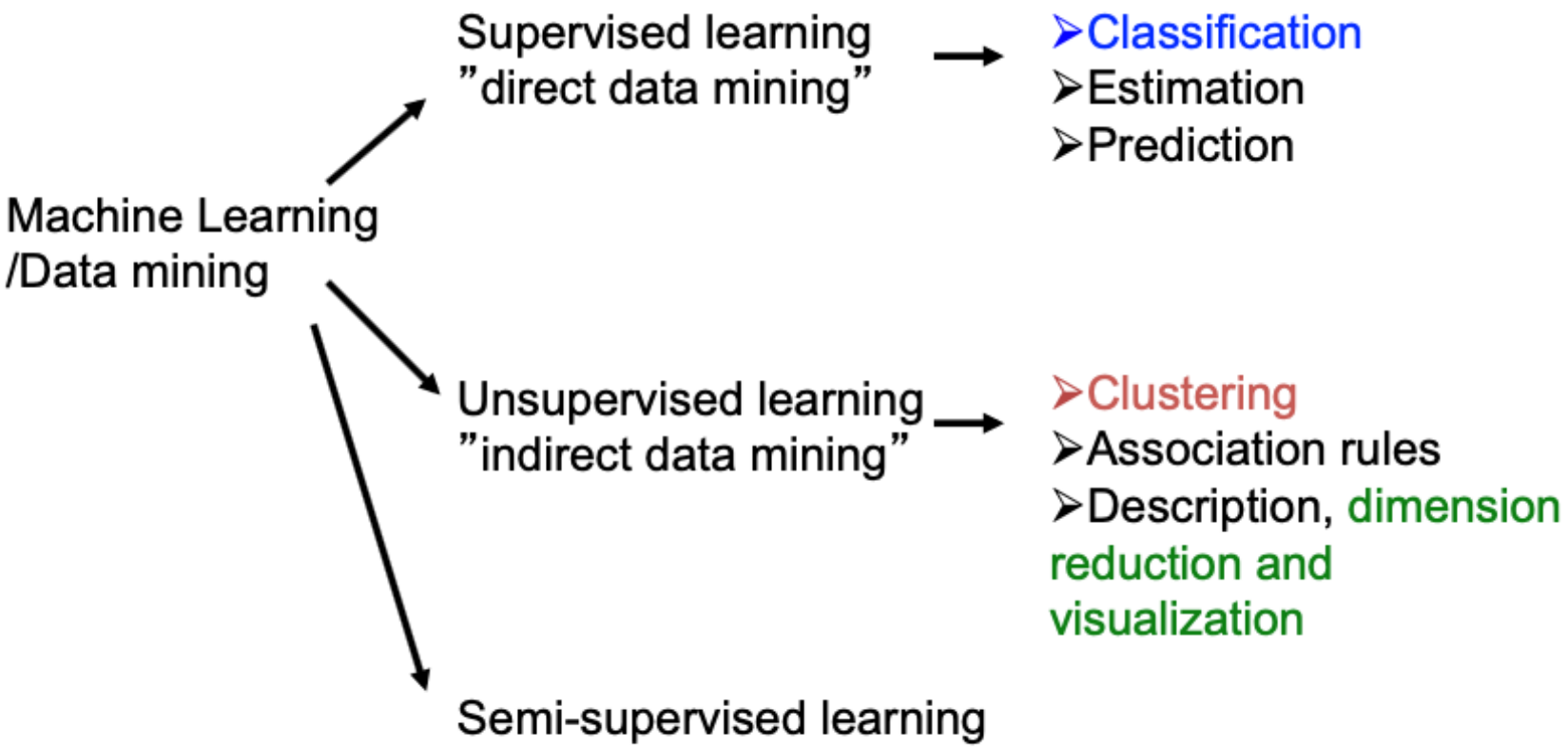
Supervised vs. Unsupervised

$\mathbf{X}$: independent variables, predictors, explanatory variables

$\mathbf{y}$: dependent variables, outcomes, response variables



Supervised vs. Unsupervised



1

[1] Modified from Figure 1.1 from Data Clustering by Gan, Ma and Wu

Supervised Learning

In supervised learning, the problem is well-defined:

- Given a set of observations $\mathbf{X}, \mathbf{Y}$
- Estimate the density $Pr(\mathbf{Y}, \mathbf{X})$
- Usually the goal is to find the model/parameters to minimize a loss, $L(\mathbf{Y}, f(\mathbf{X}))$
- A *common loss* is the **Expected Prediction Error**:

$$EPE(f) = E(\mathbf{Y} - f(\mathbf{X}))^2$$

- Objective Criteria **exist** to measure the success of a supervised learning method

Unsupervised Learning

In unsupervised learning, there is no output variable, all we observe is a set X.

The goal is to infer $Pr(X)$ and/or some of its properties.

- Low dimension: non-parametric density estimation if possible
- High dimension: need to simply properties without density estimation, or apply strong assumptions to estimate the density.

There is **no objective criteria** from the data themself

- Heuristic arguments
- External information
- Evaluate based on properties of the data

Classification

Classification

The general scheme.

An example.

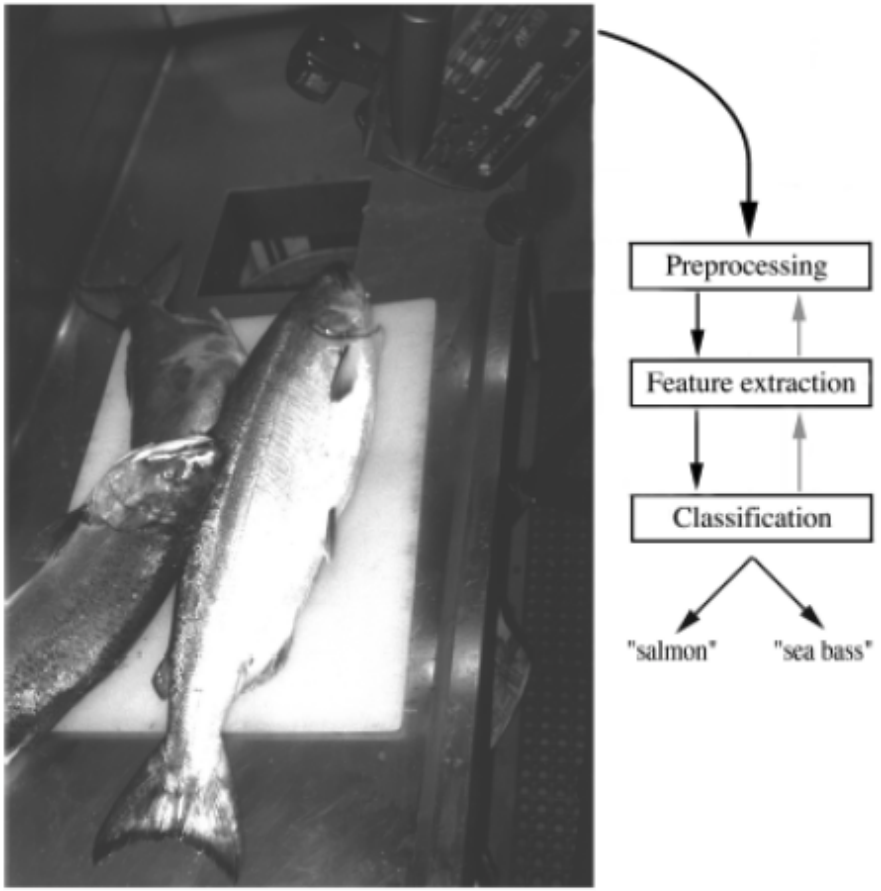


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classification

In most cases, a single feature is not enough to generate a good classifier.

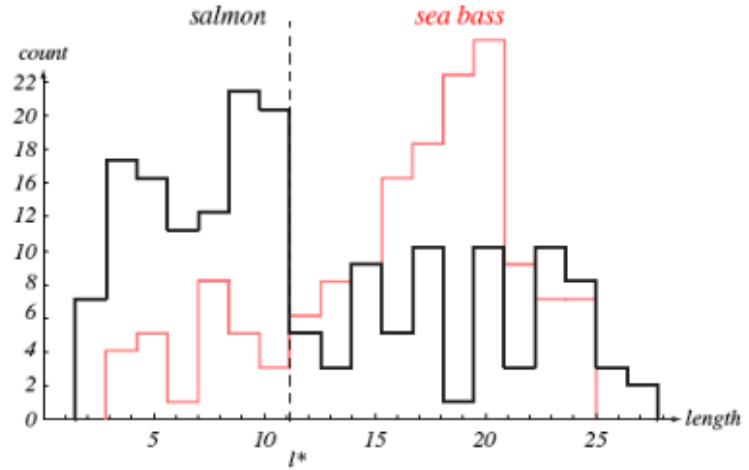


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

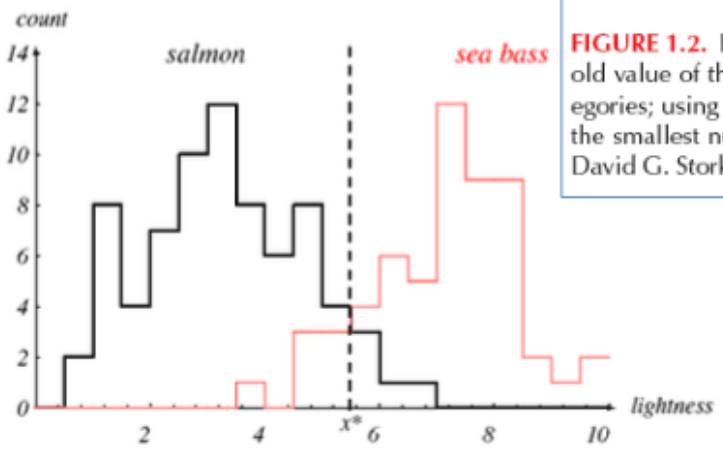


FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classification

Classification

Two extremes: overly rigid and overly flexible classifiers.

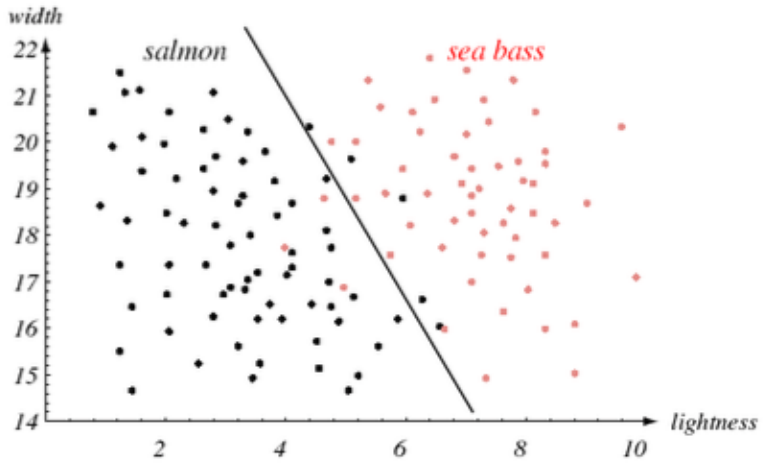


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

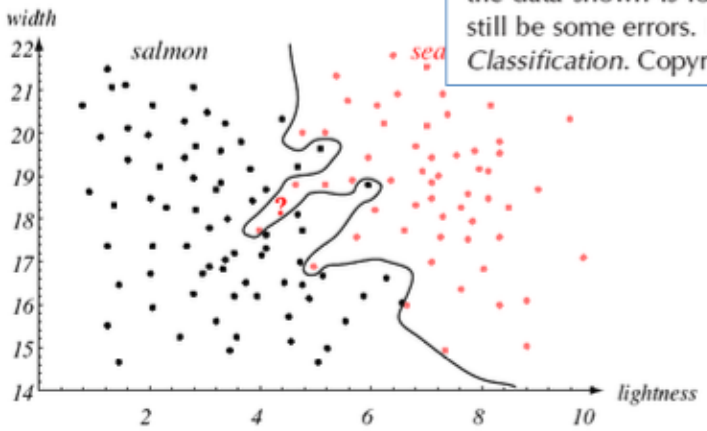


FIGURE 1.5. Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked **?** is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea bass. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classification

Goal: an optimal trade-off between model simplicity and training set performance.

This is similar to the AIC / BIC / …… model selection in regression.

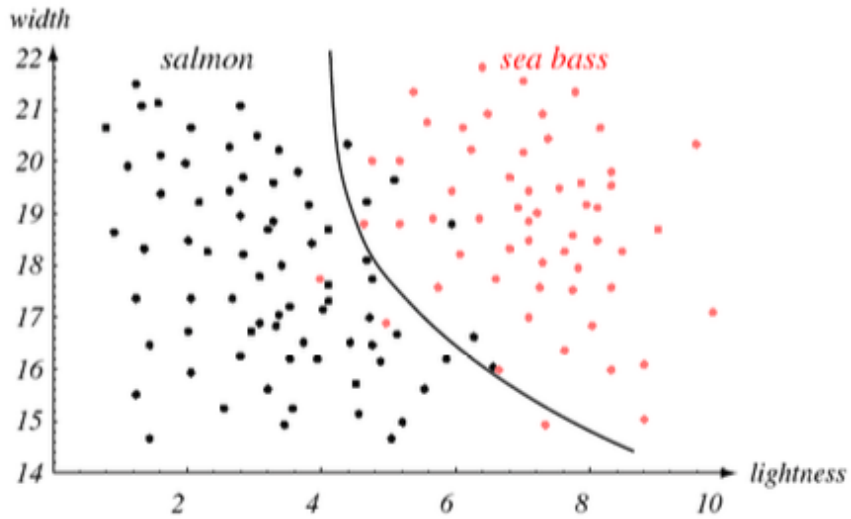


FIGURE 1.6. The decision boundary shown might represent the optimal tradeoff between performance on the training set and simplicity of classifier, thereby giving the highest accuracy on new patterns. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classification

Classification

A classification project:

a systematic view.

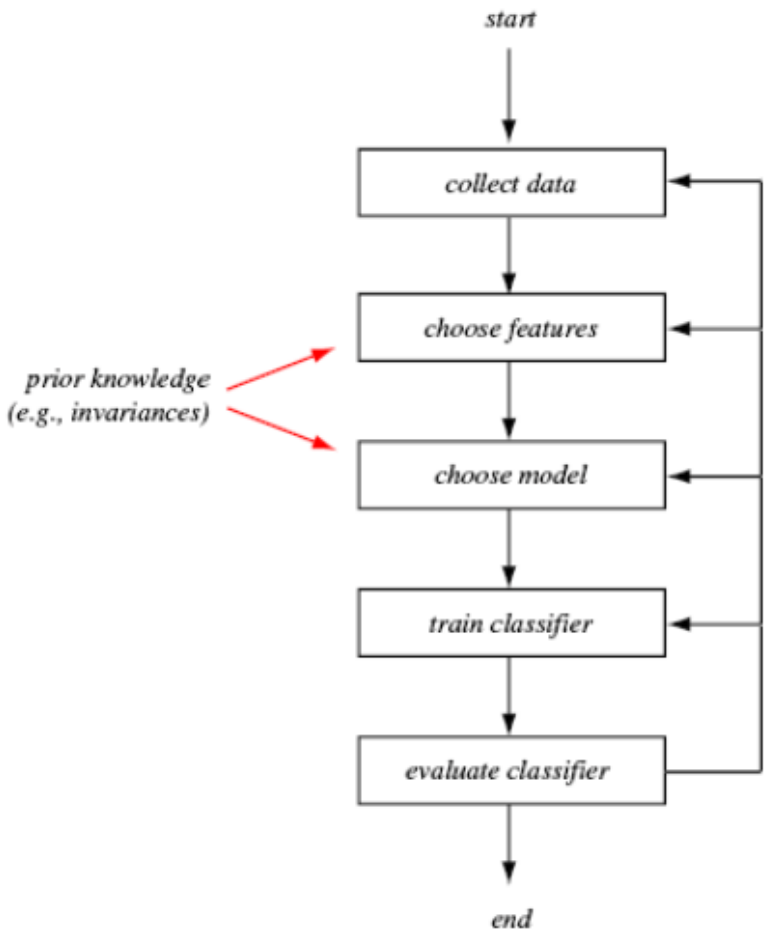
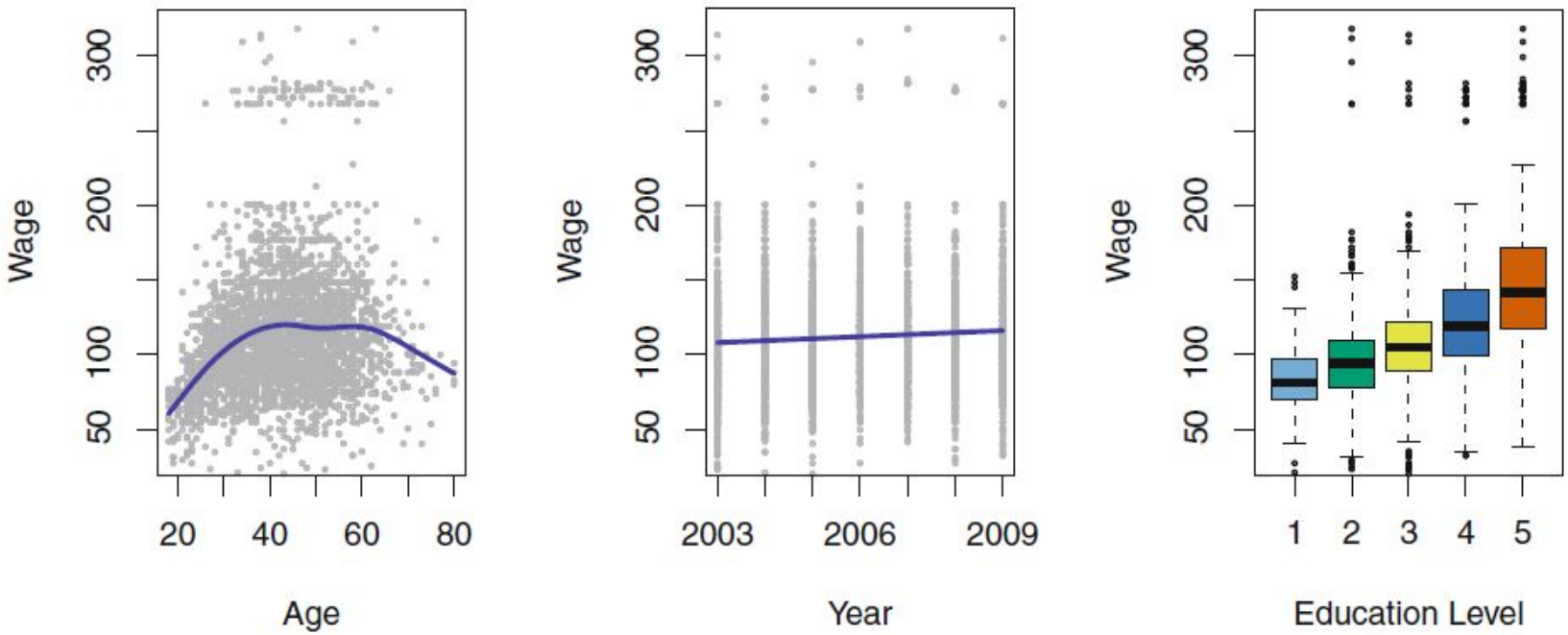


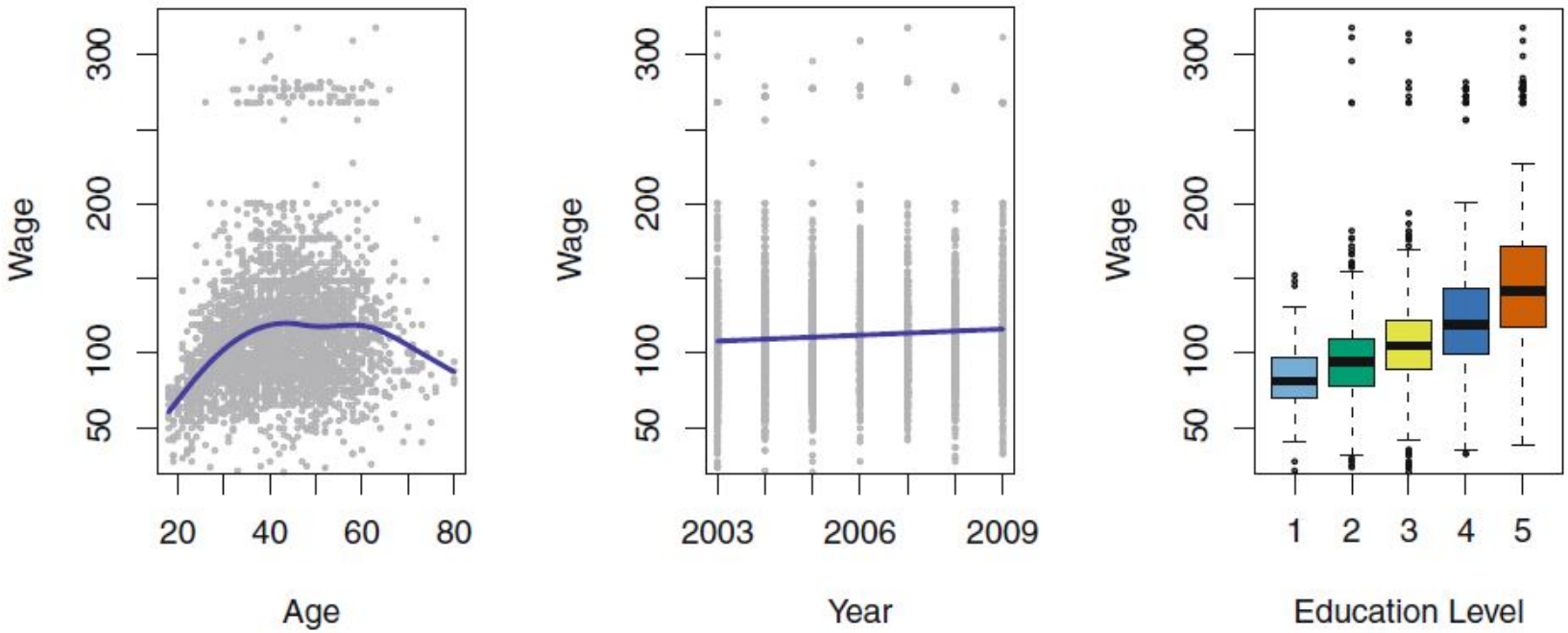
FIGURE 1.8. The design of a pattern recognition system involves a design cycle similar to the one shown here. Data must be collected, both to train and to test the system. The characteristics of the data impact both the choice of appropriate discriminating features and the choice of models for the different categories. The training process uses some or all of the data to determine the system parameters. The results of evaluation may call for repetition of various steps in this process in order to obtain satisfactory results. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Wage Prediction



Wages for a group of males from the Atlantic region.

Wage Prediction



Left panel: wage $\sim$ age

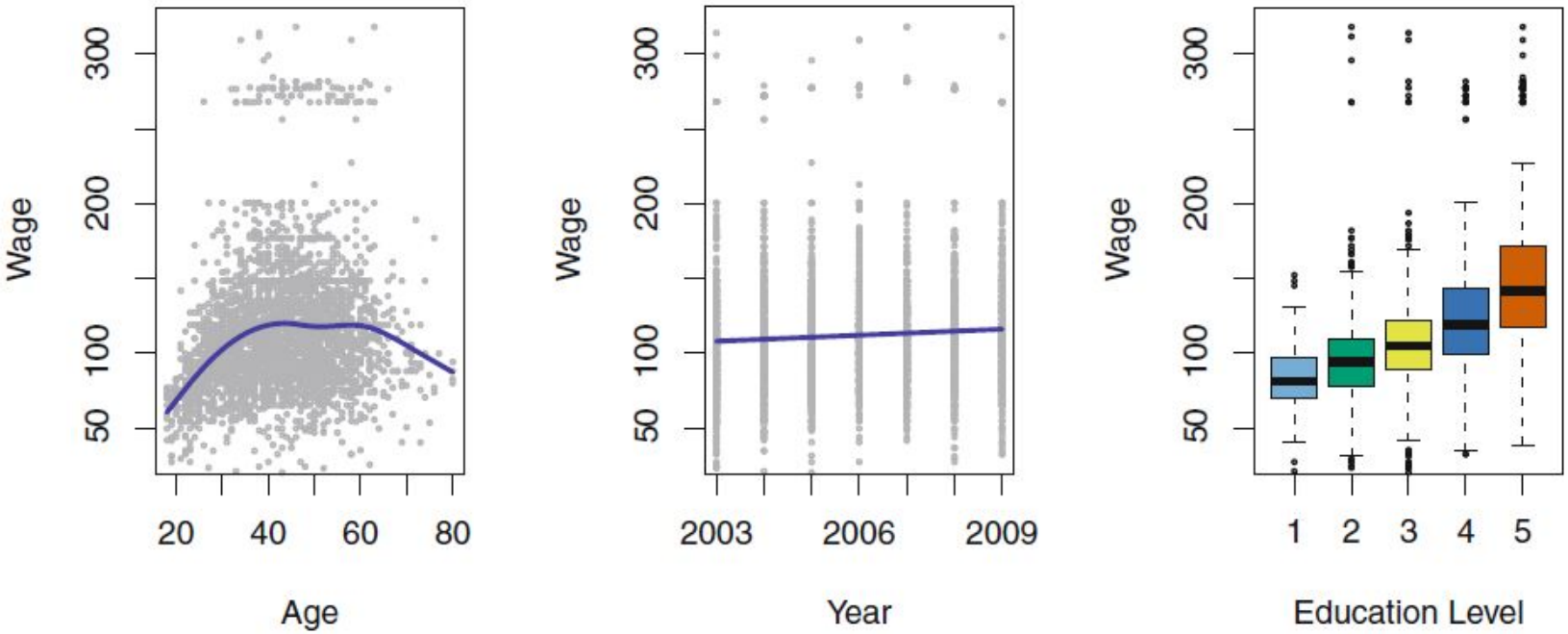
wage increases with age, but decrease after 60

Use the blue curve to predict wage using age

significant amount of variability around average wage value

"Age" alone is insufficient for an accurate prediction.

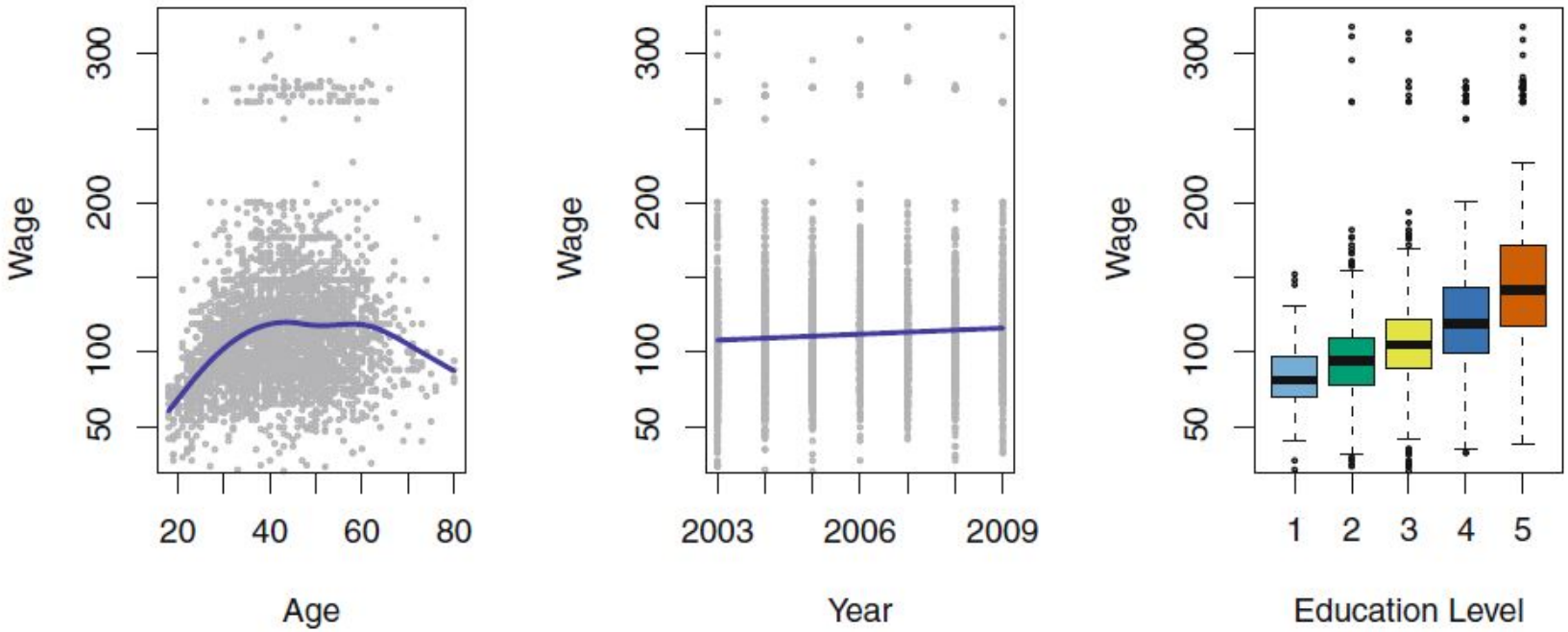
Wage Prediction



Mid panel: wage $\sim$ year

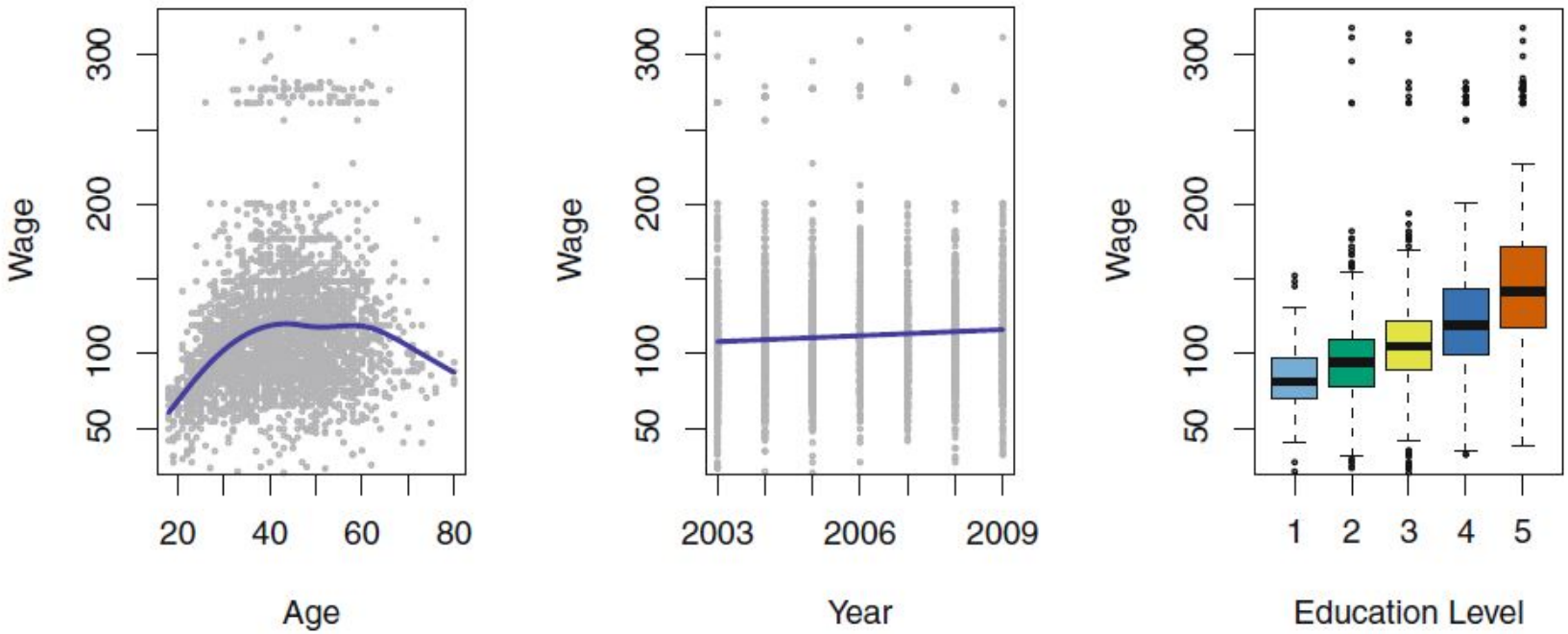
Around \$10k wage increase per year on average under linear model

Wage Prediction



Right panel: wage $\sim$ education
(1) — no high school diploma (5) — Advanced graduate degree
Men with higher education levels tend to have higher wagers

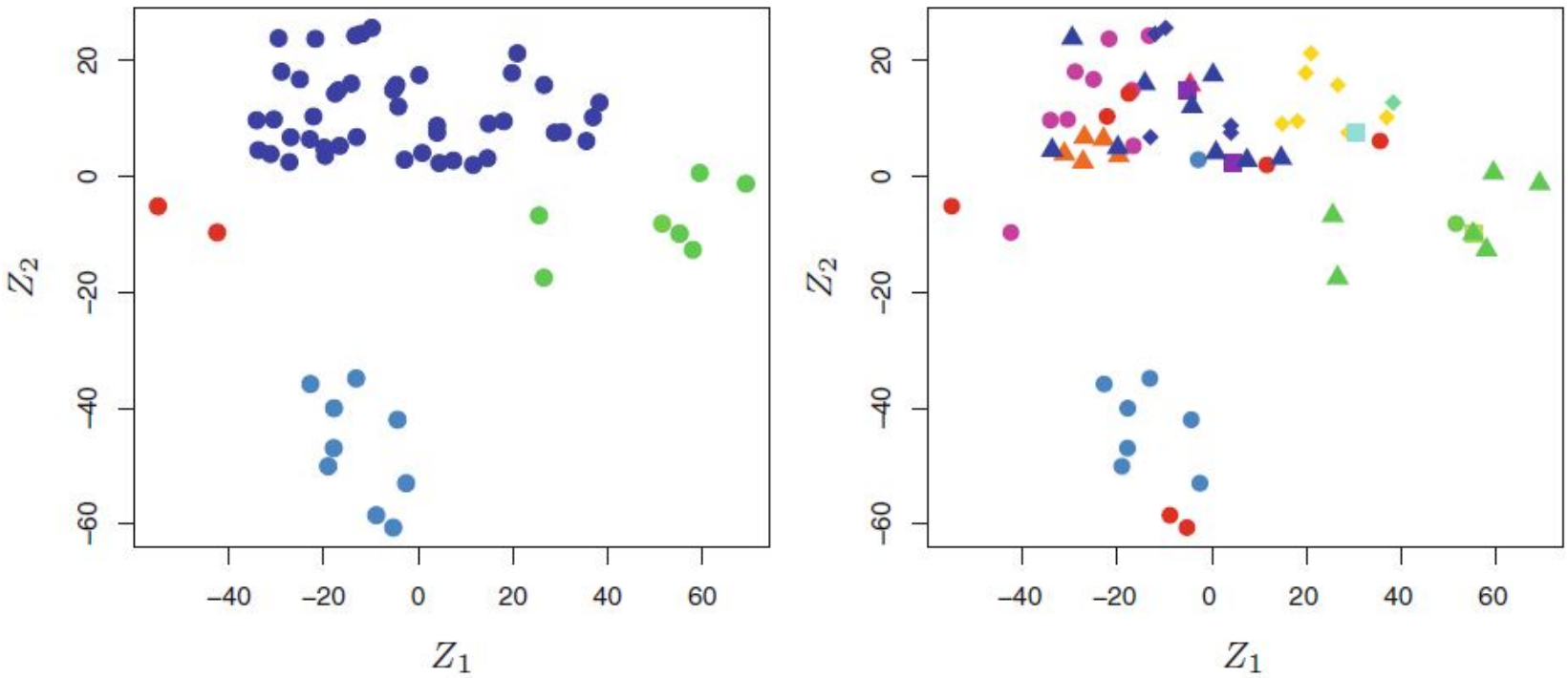
Wage Prediction



Ideally, more accurate wage prediction of wage can be obtained by combining age, year and education.

Consider non-linear modeling: polynomial regression, spline, Generalized Additive Models (GAM)

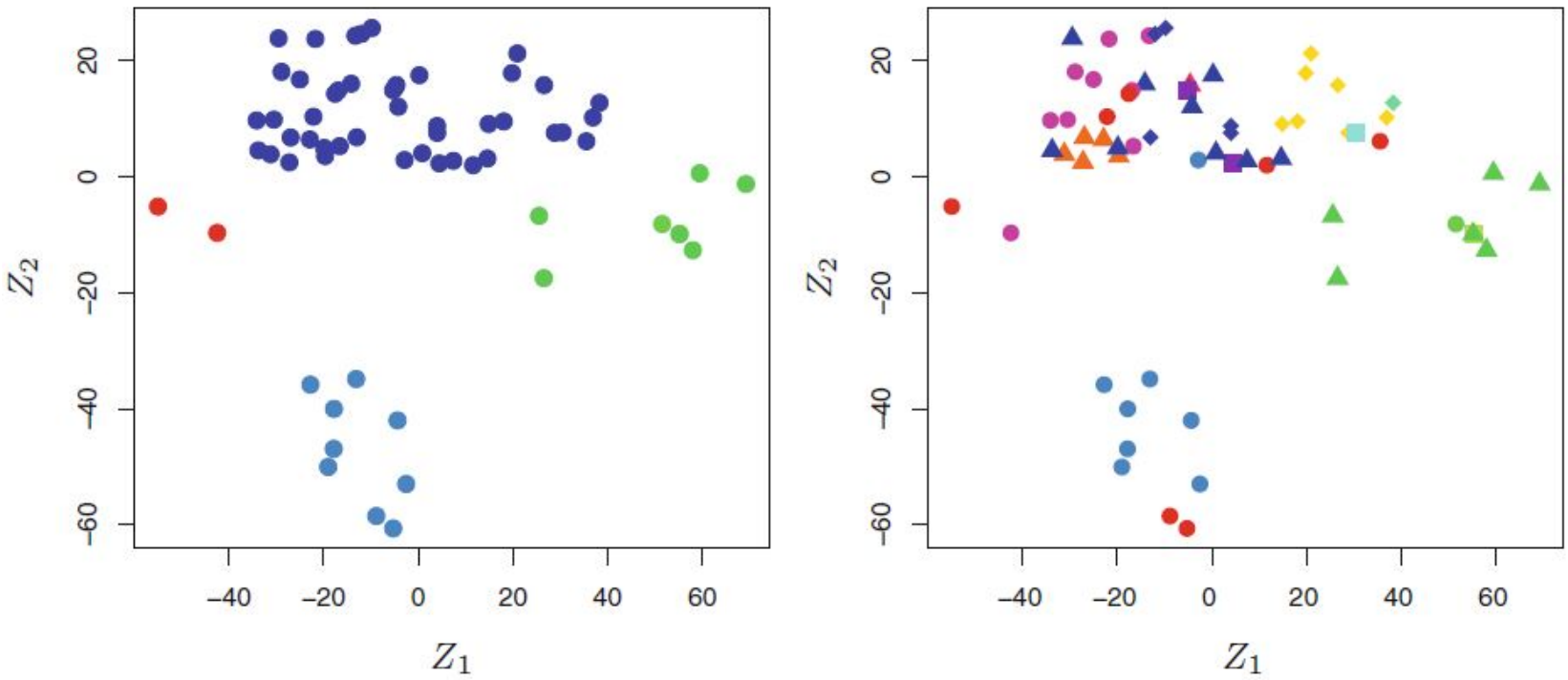
Gene Expression Data Clustering



Dataset NCI60: A matrix of 6,830 gene expression measurements for each of 64 cancer cell lines.

Are there any groups (clusters), among the cell lines based on their gene expression measurements?

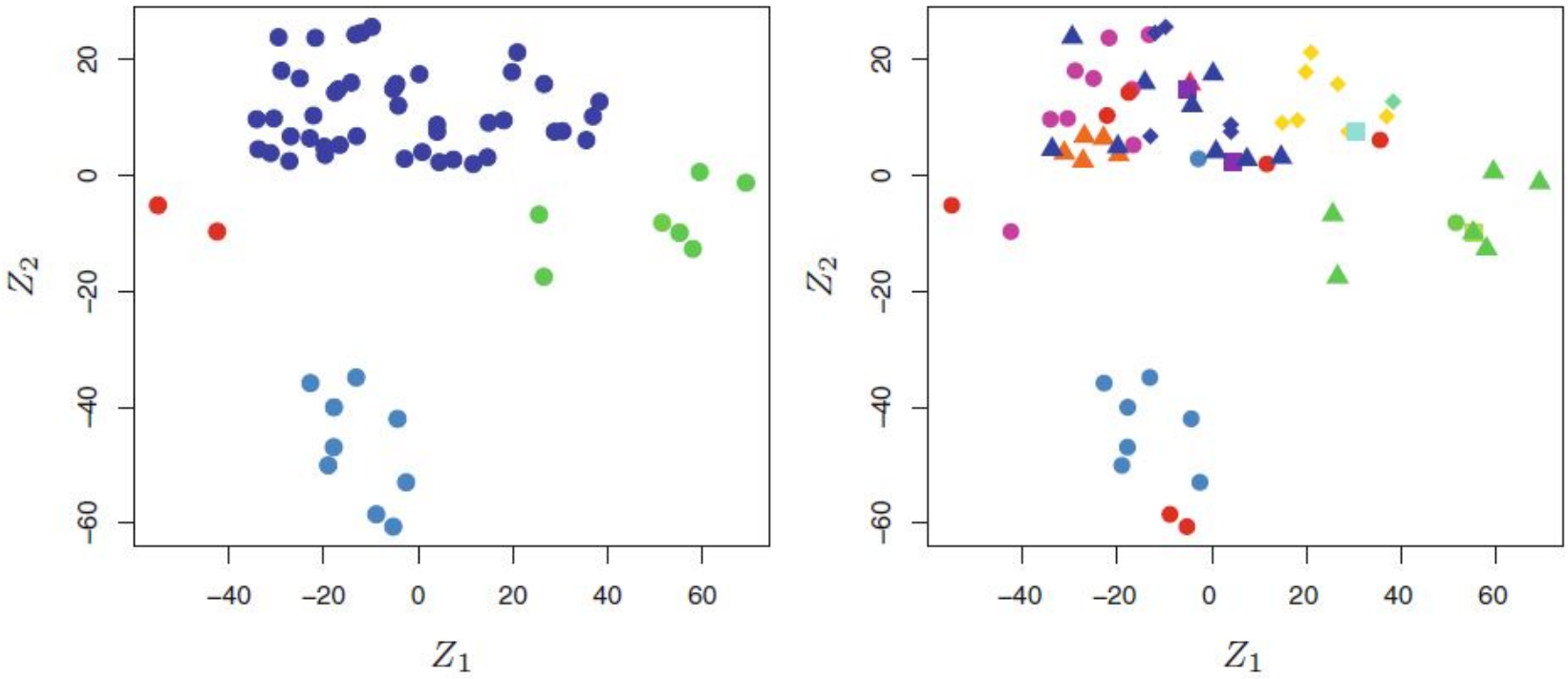
Gene Expression Data Clustering



Principal Component Analysis was adopted

Z_1: 1st PC dimension; Z_2: 2nd PC dimension

Gene Expression Data Clustering



Left: Potentially 4 data-driven clusters
Right: 14 types of cancer labeled (external information)
Cell lines with same cencer type tend to be located near each other

A brief history of statistical learning

- Early 1800: *method of least squares* by Legendre and Gauss. Now a.k.a. *linear regression*
  - First applied in astronomy
  - Robust, easy to interpret, powerful
- 1936: *linear discriminant analysis* by Fisher
- 1940s: *logistic regression*
- 1970s: *generalized linear models (GLM)* Nelder and Wedderburn
- 1980s: Improved computing power enables non-linear methods
- 1980s: *classification and regression trees, cross-validation*
- 1990s: R language, continuous improvement in computing

The Era of Big Data

E.O. Wilson

"We are drowning in information, while starving for wisdom. "

The Era of Big Data

E.O. Wilson

"We are drowning in information, while starving for wisdom. "

Andreas Buja

"There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea."

Four premises of this course

1

Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.

Four premises of this course

1

Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.

2

Statistical learning should not be viewed as a series of black boxes.

Four premises of this course

1

Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.

2

Statistical learning should not be viewed as a series of black boxes.

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 inside the box!

Four premises of this course

1

Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.

2

Statistical learning should not be viewed as a series of black boxes.

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 inside the box!

4

Remember to apply statistical learning methods to real-world problems.

Review of Linear Algebra

Notations

- n: sample size, # of observations
- p: # of variables

For example, in the wage dataset, we have $n = 3,000$ people and $p = 12$ variables (such as $year$, age, $education$, and more)

Matrix

- **Scalar:** ONE numerical value alone
  - $5, 100, n = 3,000, a^2$
- **Vector:** A vector of length n is denoted as $\mathbf{a} = (a_i)_n$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

- If all elements in a vector are 1, the vector is denoted as $\mathbf{1_n}$
- We will stick to the convention that a vector is always a **column vector**.

Matrix

- **Matrix:** $m \times n$ matrix with elements a_{ij} is denoted as $\mathbf{A} = (a_{ij})_{m \times n}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

- **Diagonal Matrix:**

$$diag(a_1, a_2, ..., a_n) \equiv \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & a_n \end{bmatrix}$$

Matrix

- **Identity Matrix:** $\mathbf{I_n} := diag(1, 1, ..., 1)$

$$\mathbf{I_n} \equiv \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & 1 & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- **Matrix Transpose:** If $\mathbf{A} = (a_{ij})_{m \times n}$, then $\mathbf{A}^T$ is an $n \times m$ matrix, where $a_{ij}^T = a_{ji}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \text{ then } \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

Matrix

- Matrix Transpose example:

  If $\mathbf{A} = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{bmatrix}$, then $\mathbf{A}^T = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$

- **Symmetric Matrix:** If $\mathbf{A} = \mathbf{A}^T$ then $\mathbf{A}$ is symmetric.

  $\mathbf{A} = \begin{bmatrix} 1 & 9 & 5 & 6 \\ 9 & 2 & 6 & 7 \\ 5 & 6 & 3 & 8 \\ 6 & 7 & 8 & 4 \end{bmatrix}$ is a symmetric matrix.

  Apparently, **identity matrix** and **diagonal matrix** are always symmetric matrices.

Matrix

- **Matrix Sum:** If $\mathbf{A} = (a_{ij})_{m\times n}$ and $\mathbf{B} = (b_{ij})_{m\times n}$ then

$$\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij})_{m\times n}$$

- **Matrix Product:** If $\mathbf{A} = (a_{ij})_{m\times n}$ and $\mathbf{B} = (b_{ij})_{n\times p}$, then

$$\mathbf{AB} = (c_{ij})_{m\times p}$$

where

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

- Matrix products satisfy $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$

Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix} \text{ Then}$$

$$\mathbf{AB} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \times 7 + 2 \times 8 + 3 \times 9 & 1 \times 10 + 2 \times 11 + 3 \times 12 \\ 4 \times 7 + 5 \times 8 + 6 \times 9 & 4 \times 10 + 5 \times 11 + 6 \times 12 \end{bmatrix}$$

$$= \begin{bmatrix} 50 & 68 \\ 122 & 167 \end{bmatrix}$$

It is only possible to compute $\mathbf{AB}$ if :
$\#$ of columns in $\mathbf{A} = \#$ of rows in $\mathbf{B}$

Matrix Inverse

- **Matrix Inverse Definition:** An $n \times n$ matrix $\mathbf{A}$ is invertible (or non-singular) if there is a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$
- An $n \times n$ matrix $\mathbf{A}$ is invertible if and only if $rank(\mathbf{A}) = n$
- Inverse of Product: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if $\mathbf{A}$ and $\mathbf{B}$ are invertible

Vector Product, Norm and Orthogonality

- **Inner Product:**

$$\mathbf{a}^{\mathrm{T}}\mathbf{b} = \sum_i a_i b_i$$

  where $\mathbf{a} = (a_i)$ and $\mathbf{b} = (b_i)$ are vectors with the same length.

- **Vector norm:**

$$||\mathbf{a}|| = \sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}}$$

- **Orthogonal vectors:** $\mathbf{a} = (a_i)$ and $\mathbf{b} = (b_i)$ are orthogonal vectors if $\mathbf{a}^{\mathrm{T}}\mathbf{b} = 0$ (i.e. they are perpendicular)

Vector Product, Norm and Orthogonality

- **Inner Product:**

$$\mathbf{a}^{\mathrm{T}}\mathbf{b} = \sum_i a_i b_i$$

  where $\mathbf{a} = (a_i)$ and $\mathbf{b} = (b_i)$ are vectors with the same length.

- **Vector norm:**

$$||\mathbf{a}|| = \sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}}$$

- **Orthogonal vectors:** $\mathbf{a} = (a_i)$ and $\mathbf{b} = (b_i)$ are orthogonal vectors if $\mathbf{a}^{\mathrm{T}}\mathbf{b} = 0$ (i.e. they are perpendicular)

Eigenvalue and Eigenvectors

- **Definition:** Given an $n \times n$ matrix $\mathbf{A}$, then if a scalar λ and a vector $\mathbf{u}$ satisfy

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$$

  then $\mathbf{u}$ is called an eigenvector of $\mathbf{A}$, with associated eigenvalue λ.
- **Intuition:** Eigenvector is *a direction* in multidimensional space. The linear transformation, $\mathbf{A}$, can't rotate this direction like on others. Instead, $\mathbf{A}$ can only stretch/squeeze/reverse this direction.

Eigenvalue and Eigenvectors

Basic Properities:
Let matrix $\mathbf{A}$ be a $n \times n$ square matrix with $rank(\mathbf{A}) = n$.
Define a square matrix $\mathbf{Q} = (\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_n})$, whose columns are the n linearly independent eigenvectors of $\mathbf{A}$.
Define diagonal matrix $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_n)$, where each diagonal element is the eigenvalue associated with the each column of $\mathbf{Q}$. Then from the basic defination we have:

$$\mathbf{AQ} = \mathbf{Q\Lambda}$$

Because the columns of $\mathbf{Q}$ are linearly independent, $\mathbf{Q}$ is full rank and thus invertible. Right multiplying both sides by $\mathbf{Q^{-1}}$:

$$\mathbf{A} = \mathbf{Q\Lambda Q^{-1}}$$

Textbook chapters

Textbook:

ISLR: An Introduction to Statistical Learning: with applications in R

- ISLR chapter 1 & 2