Computational Data Mining

Part 1: Course Introduction

Instructor: Zahra Narimani

Department of Computer Science and Information Technology,
Institute for Advanced Studies in Basic Sciences, Zanjan, Iran

Reference Book

Main reference:

Elden, L., Matrix Methods in Data Mining and Pattern Recognition, SIAM, 2007

Grading policy

Homework ~ 4

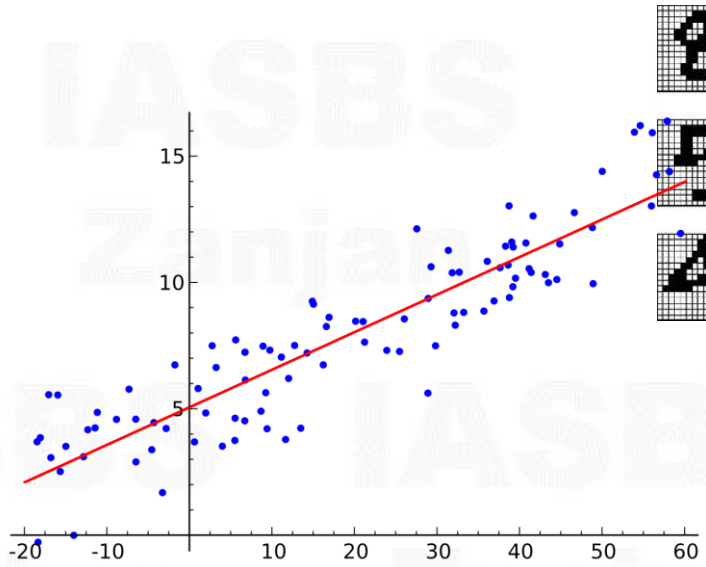
Project ~ 6-7

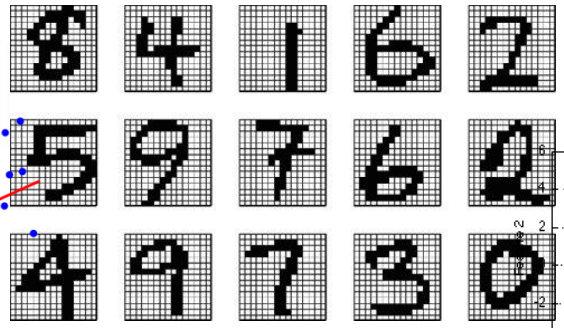
Midterm and Final Exam ~ 9-10

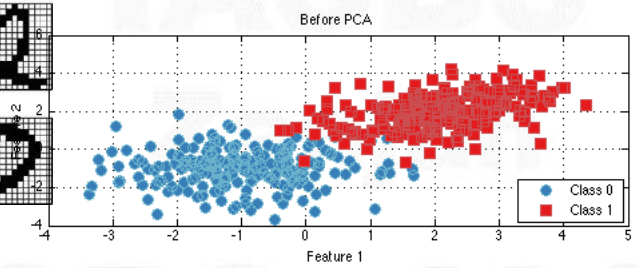
Data mining and pattern recognition

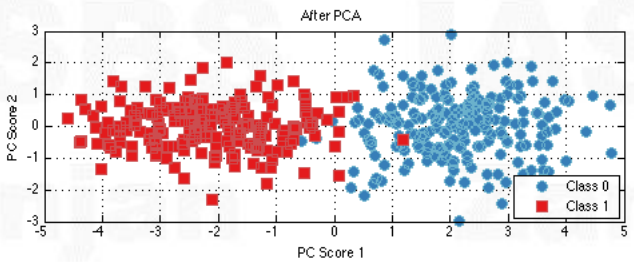
* The science of extracting useful information from large data sets

* The act of taking in raw data and making an action based on the ‘category’ of the pattern









Vectors and matrices

Document 1: The **Google™ matrix** P is a model of the **Internet**.
Document 2: P_{ij} is nonzero if there is a **link** from **Web page** j to i.
Document 3: The **Google matrix** is used to **rank** all **Web pages**.
Document 4: The **ranking** is done by solving a **matrix eigenvalue** problem.
Document 5: **England** dropped out of the top 10 in the **FIFA ranking**.

$$q = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{10}$$

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
eigenvalue	0	0	0	1	0
England	0	0	0	0	1
FIFA	0	0	0	0	1
Google	1	0	1	0	0
Internet	1	0	0	0	0
link	0	1	0	0	0
matrix	1	0	1	1	0
page	0	1	1	0	0
rank	0	0	1	1	1
Web	0	1	1	0	0

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Matrix as linear transformation

$$a_{.j} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}, \qquad j = 1, 2, \ldots, n$$

$$A = \begin{pmatrix} a_{.1} & a_{.2} & \cdots & a_{.n} \end{pmatrix}$$

$$y = Ax = \begin{pmatrix} a_{.1} & a_{.2} & \ldots & a_{.n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{j=1}^{n} x_j a_{.j}$$

System of linear equations

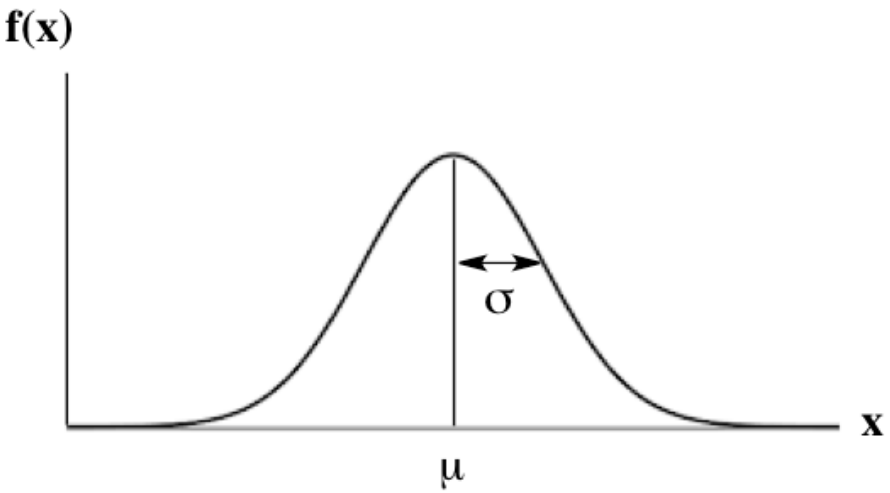
$$2x_1 + 3x_2 + 5x_3 = 1$$
$$4x_1 - 2x_2 - 7x_3 = 8$$
$$9x_1 + 5x_2 - 3x_3 = 2$$

System of linear equations

$$
\begin{aligned}
2x_1 + 3x_2 + 5x_3 &= 1 \\
4x_1 - 2x_2 - 7x_3 &= 8 \\
9x_1 + 5x_2 - 3x_3 &= 2
\end{aligned}
$$

$$
\begin{bmatrix} 2 & 3 & 5 \\ 4 & -2 & -7 \\ 9 & 5 & -3 \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}
=
\begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix}
$$

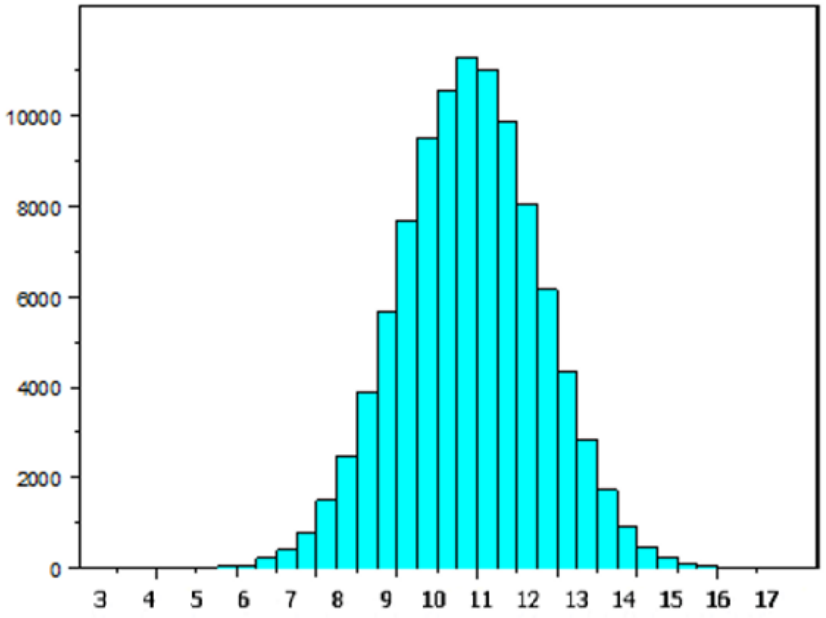
Normal distribution

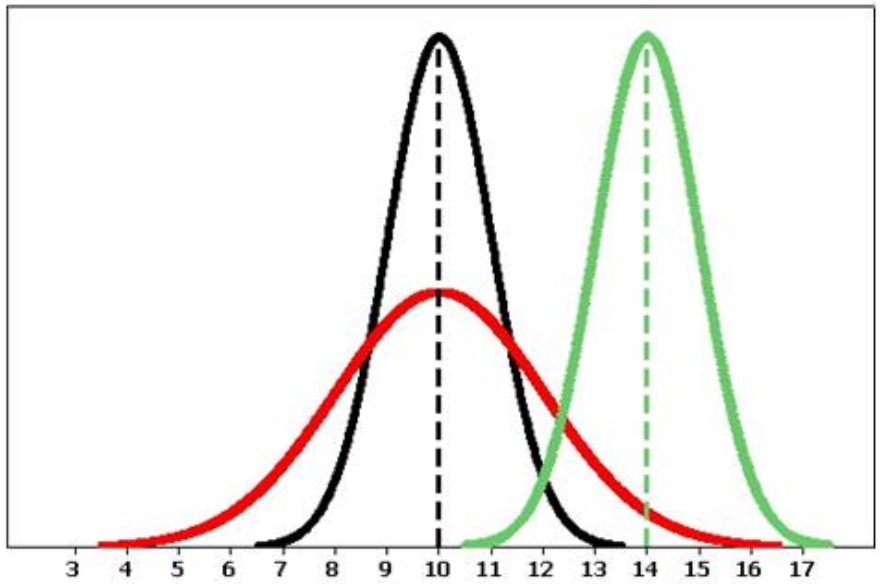


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

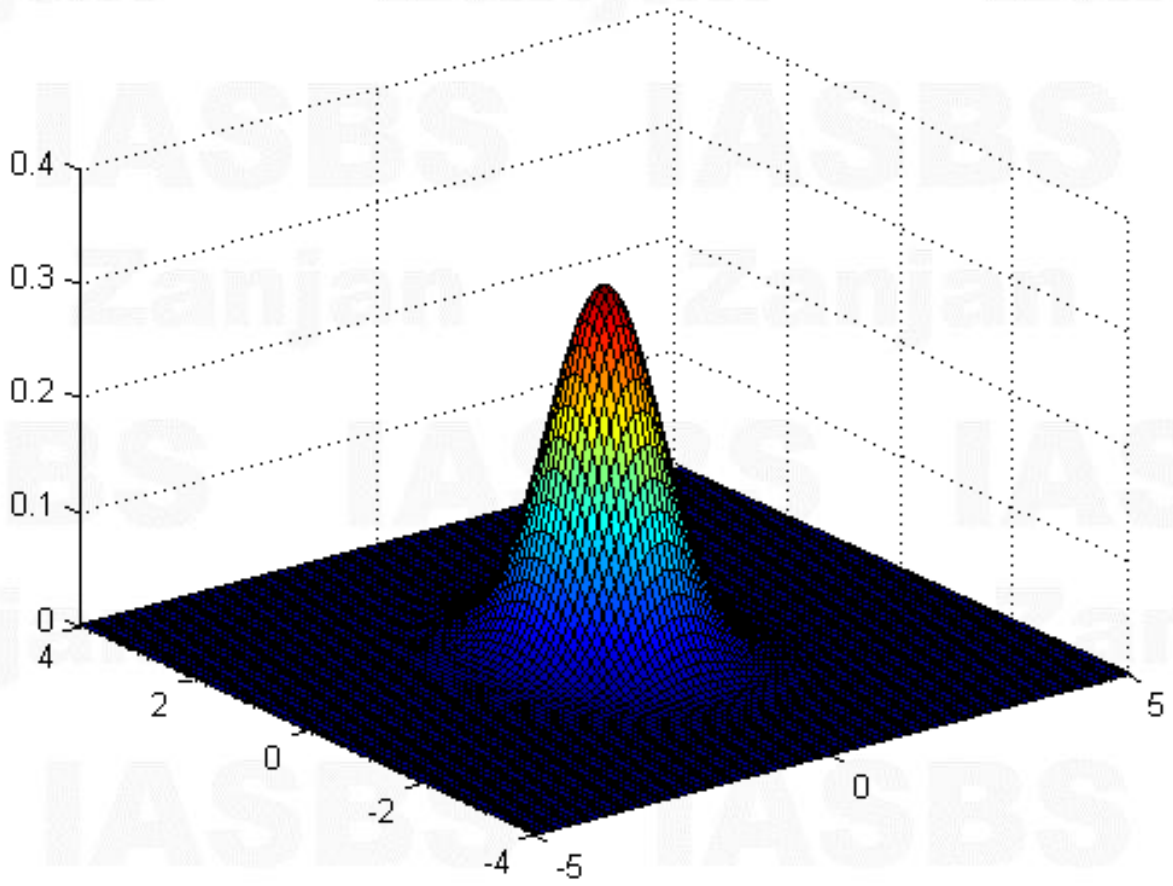
$$e = 2.71828$$

Optimization problem: find parameters of a distribution





Search the space to find optimal solution



Any Question?