Data Mining and Analysis: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 1: Data Mining and Analysis

Data Matrix

Data can often be represented or abstracted as an $n \times d$ *data matrix*, with n rows and d columns, given as

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- **Rows**: Also called *instances*, *examples*, *records*, *transactions*, *objects*, *points*, *feature-vectors*, etc. Given as a d-tuple

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

- **Columns**: Also called *attributes*, *properties*, *features*, *dimensions*, *variables*, *fields*, etc. Given as an n-tuple

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Iris Dataset Extract

	Sepal length X_1	**Sepal width** X_2	**Petal length** X_3	**Petal width** X_4	**Class** X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Attributes

Attributes may be classified into two main types

- **Numeric Attributes**: real-valued or integer-valued domain
  - *Interval-scaled*: only differences are meaningful
    e.g., `temperature`
  - *Ratio-scaled*: differences and ratios are meaningful
    e..g, `Age`
- **Categorical Attributes**: set-valued domain composed of a set of symbols
  - *Nominal*: only equality is meaningful
    e.g., domain(`Sex`) = { `M, F`}
  - *Ordinal*: both equality (are two values the same?) and inequality (is one value less than another?) are meaningful
    e.g., domain(`Education`) = { `High School`, `BS`, `MS`, `PhD`}

Data: Algebraic and Geometric View

For numeric data matrix **D**, each row or point is a d-dimensional column vector:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}^T \in \mathbb{R}^d$$

whereas each column or attribute is a n-dimensional column vector:

$$X_j = \begin{pmatrix} x_{1j} & x_{2j} & \cdots & x_{nj} \end{pmatrix}^T \in \mathbb{R}^n$$

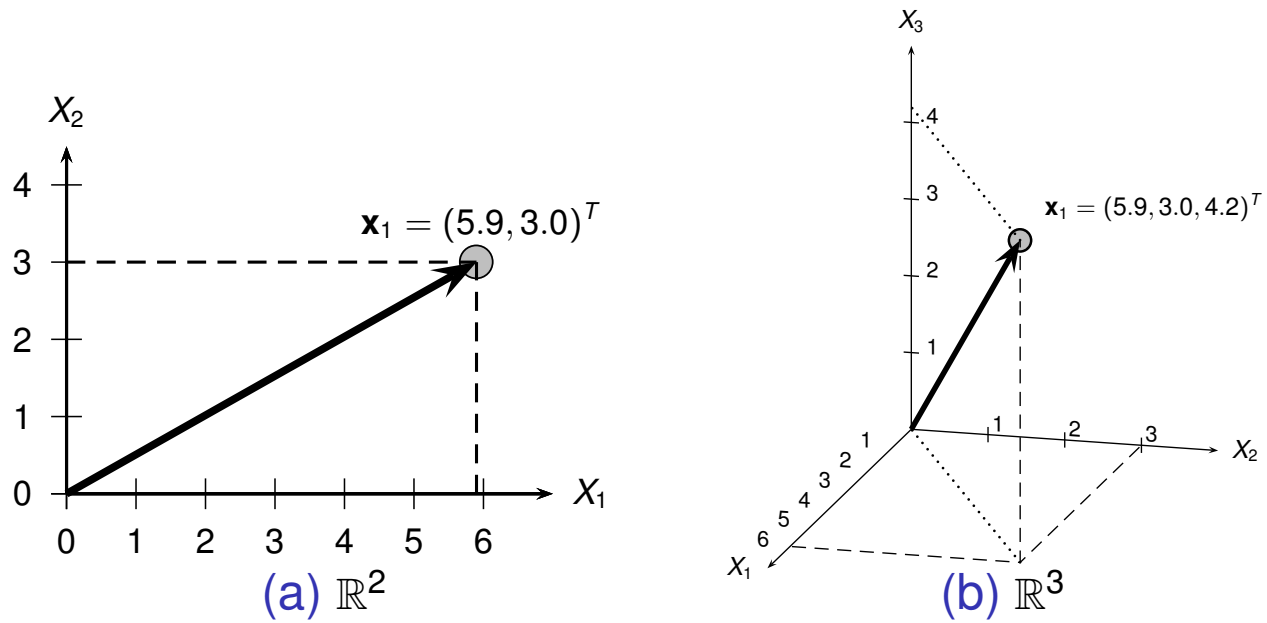
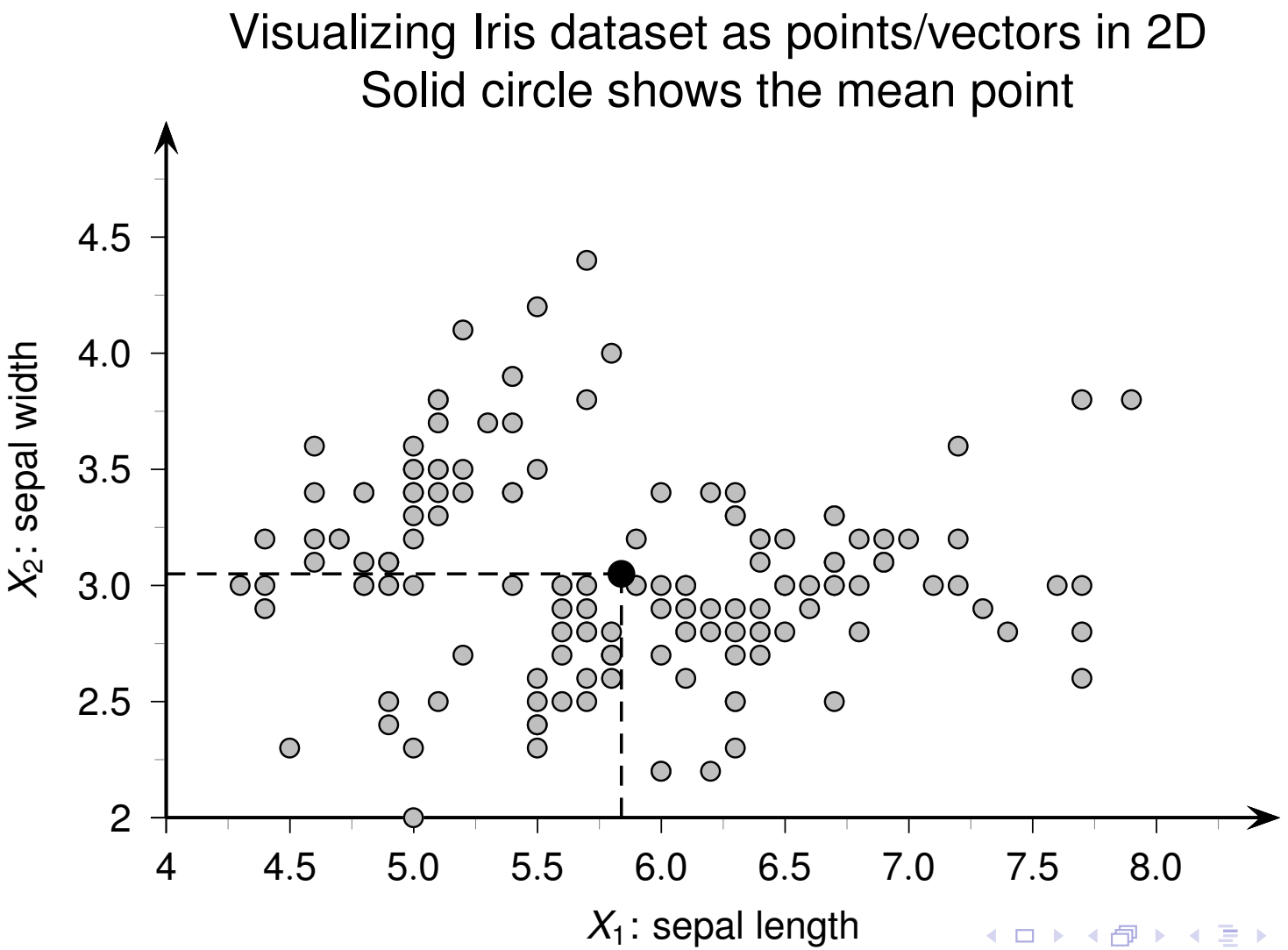


Figure: Projections of $\mathbf{x}_1 = (5.9, 3.0, 4.2, 1.5)^T$ in 2D and 3D

Scatterplot: 2D Iris Dataset

`sepal length` **versus** `sepal width`.



Numeric Data Matrix

If all attributes are numeric, then the data matrix $\mathbf{D}$ is an $n \times d$ matrix, or equivalently a set of n row vectors $\mathbf{x}_i^T \in \mathbb{R}^d$ or a set of d column vectors $X_j \in \mathbb{R}^n$

$$\mathbf{D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} — \ \mathbf{x}_1^T \ — \\ — \ \mathbf{x}_2^T \ — \\ \vdots \\ — \ \mathbf{x}_n^T \ — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

The *mean* of the data matrix $\mathbf{D}$ is the average of all the points:

$$mean(\mathbf{D}) = \boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

The *centered data matrix* is obtained by subtracting the mean from all the points:

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}^T \\ \boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T - \boldsymbol{\mu}^T \\ \mathbf{x}_2^T - \boldsymbol{\mu}^T \\ \vdots \\ \mathbf{x}_n^T - \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{pmatrix} \tag{1}$$

where $\mathbf{z}_i = \mathbf{x}_i - \boldsymbol{\mu}$ is a centered point, and $\mathbf{1} \in \mathbb{R}^n$ is the vector of ones.

Norm, Distance and Angle

Given two points $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, their *dot product* is defined as the scalar

$$\mathbf{a}^T\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_mb_m = \sum_{i=1}^{m} a_ib_i$$

The *Euclidean norm* or *length* of a vector $\mathbf{a}$ is defined as

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

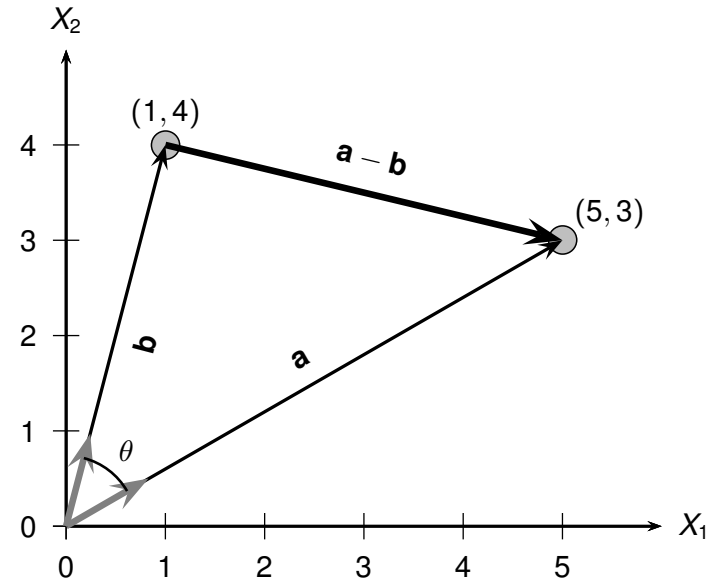
The *unit vector* in the direction of $\mathbf{a}$ is $\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$ with $\|\mathbf{a}\| = 1$.

Distance between $\mathbf{a}$ and $\mathbf{b}$ is given as

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

Angle between $\mathbf{a}$ and $\mathbf{b}$ is given as

$$\cos\theta = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|} = \left(\frac{\mathbf{a}}{\|\mathbf{a}\|}\right)^T \left(\frac{\mathbf{b}}{\|\mathbf{b}\|}\right)$$



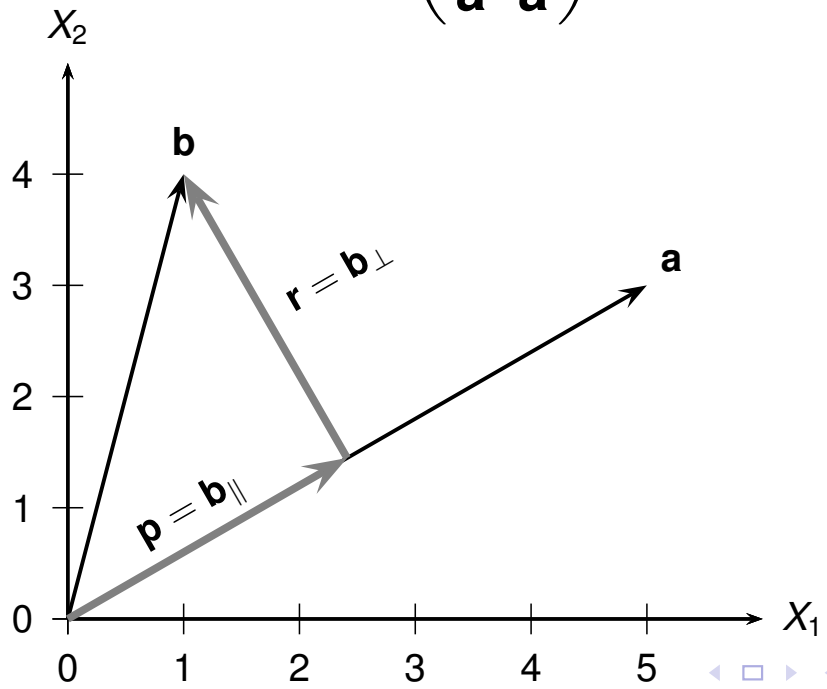
Orthogonal Projection

Two vectors **a** and **b** are *orthogonal* iff $\mathbf{a}^T\mathbf{b} = 0$, i.e., the angle between them is 90°. Orthogonal projection of **b** on **a** comprises the vector $\mathbf{p} = \mathbf{b}_{\parallel}$ parallel to **a**, and $\mathbf{r} = \mathbf{b}_{\perp}$ perpendicular or orthogonal to **a**, given as

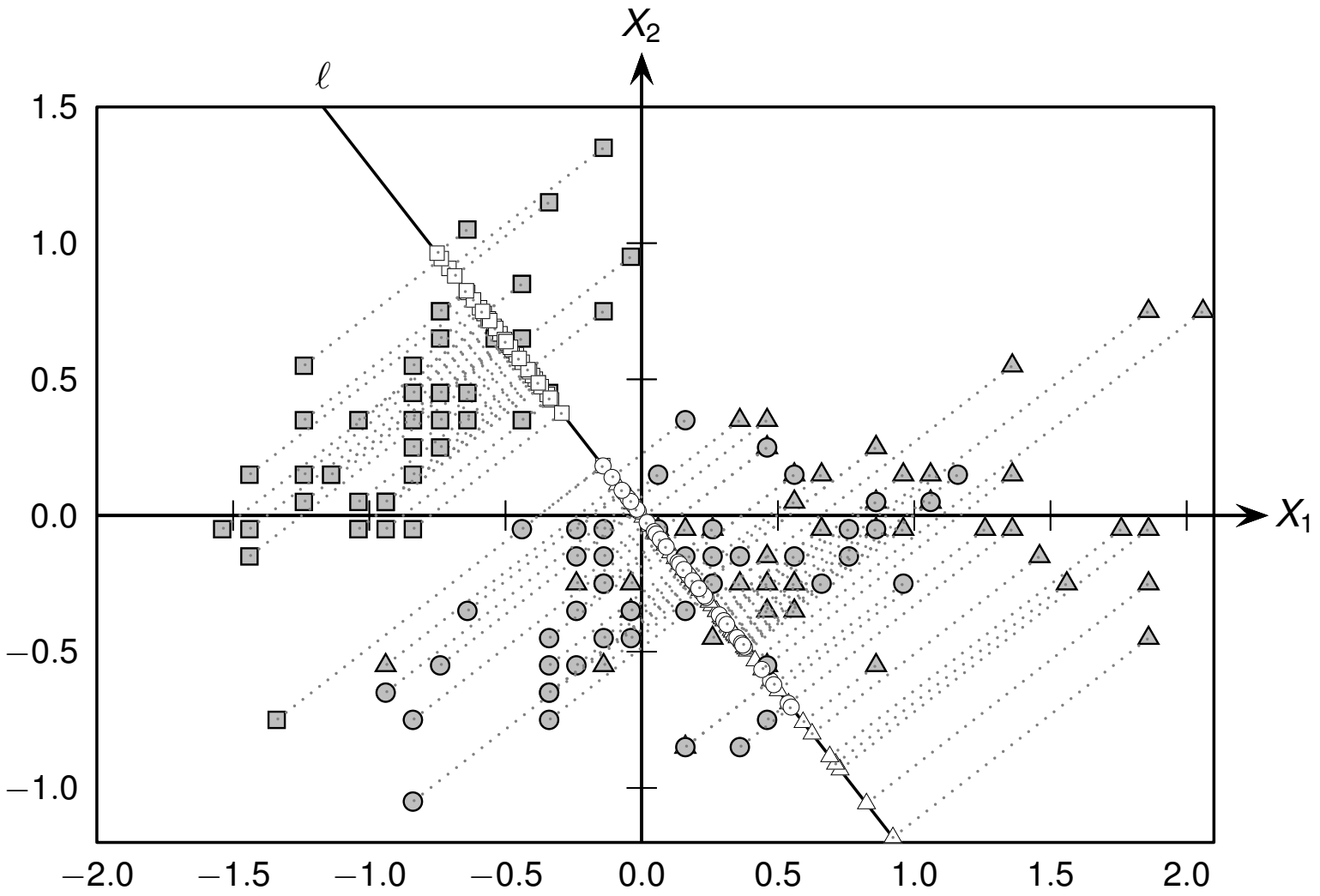
$$\mathbf{b} = \mathbf{b}_{\parallel} + \mathbf{b}_{\perp} = \mathbf{p} + \mathbf{r}$$

where

$$\mathbf{p} = \mathbf{b}_{\parallel} = \left(\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}\right)\mathbf{a}$$



Projection of Centered Iris Data Onto a Line ℓ.



Data: Probabilistic View

A *random variable* X is a function $X : \mathcal{O} \to \mathbb{R}$, where $\mathcal{O}$ is the set of all possible outcomes of the experiment, also called the *sample space*.

A *discrete random variable* takes on only a finite or countably infinite number of values, whereas a *continuous random variable* if it can take on any value in $\mathbb{R}$.

By default, a numeric attribute X_j is considered as the identity random variable given as

$$X(v) = v$$

for all $v \in \mathcal{O}$. Here $\mathcal{O} = \mathbb{R}$.

Discrete Variable: Long Sepal Length

Define random variable A, denoting long sepal length (7cm or more) as follows:

$$A(v) = \begin{cases} 0 & \text{if } v < 7 \\ 1 & \text{if } v \geq 7 \end{cases}$$

The sample space of A is $\mathcal{O} = [4.3, 7.9]$, and its range is $\{0, 1\}$. Thus, A is discrete.

Probability Mass Function

If X is discrete, the *probability mass function* of X is defined as

$$f(x) = P(X = x) \qquad \text{for all } x \in \mathbb{R}$$

f must obey the basic rules of probability. That is, f must be non-negative:

$$f(x) \geq 0$$

and the sum of all probabilities should add to 1:

$$\sum_x f(x) = 1$$

Intuitively, for a discrete variable X, the probability is concentrated or massed at only discrete values in the range of X, and is zero for all other values.

Sepal Length: Bernoulli Distribution

Iris Dataset Extract: `sepal length` (in centimeters)

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

Define random variable A as follows: $A(v) = \begin{cases} 0 & \text{if } v < 7 \\ 1 & \text{if } v \geq 7 \end{cases}$

We find that only 13 Irises have sepal length of at least 7 cm. Thus, the probability mass function of A can be estimated as:

$$f(1) = P(A = 1) = \frac{13}{150} = 0.087 = p$$

and

$$f(0) = P(A = 0) = \frac{137}{150} = 0.913 = 1 - p$$

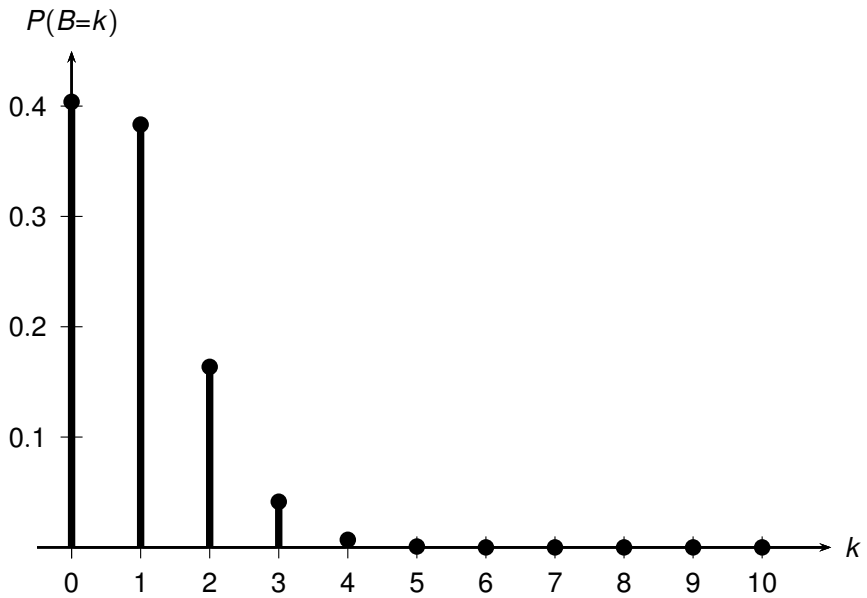
A has a *Bernoulli distribution* with parameter $p \in [0, 1]$, which denotes the probability of a *success*, that is, the probability of picking an Iris with a long sepal length at random from the set of all points.

Sepal Length: Binomial Distribution

Define discrete random variable B, denoting the number of Irises with long sepal length in m independent Bernoulli trials with probability of success p. In this case, B takes on the discrete values $[0, m]$, and its probability mass function is given by the *Binomial distribution*

$$f(k) = P(B = k) = \binom{m}{k} p^k (1-p)^{m-k}$$

Binomial distribution for long sepal length ($p = 0.087$) for $m = 10$ trials



Probability Density Function

If X is continuous, the *probability density function* of X is defined as

$$P\big(X \in [a,b]\big) = \int_a^b f(x)\, dx$$

f must obey the basic rules of probability. That is, f must be non-negative:

$$f(x) \geq 0$$

and the sum of all probabilities should add to 1:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Note that $P(X = v) = 0$ for all $v \in \mathbb{R}$ since there are infinite possible values in the sample space. What it means is that the probability mass is spread so thinly over the range of values that it can be measured only over intervals $[a, b] \subset \mathbb{R}$, rather than at specific points.

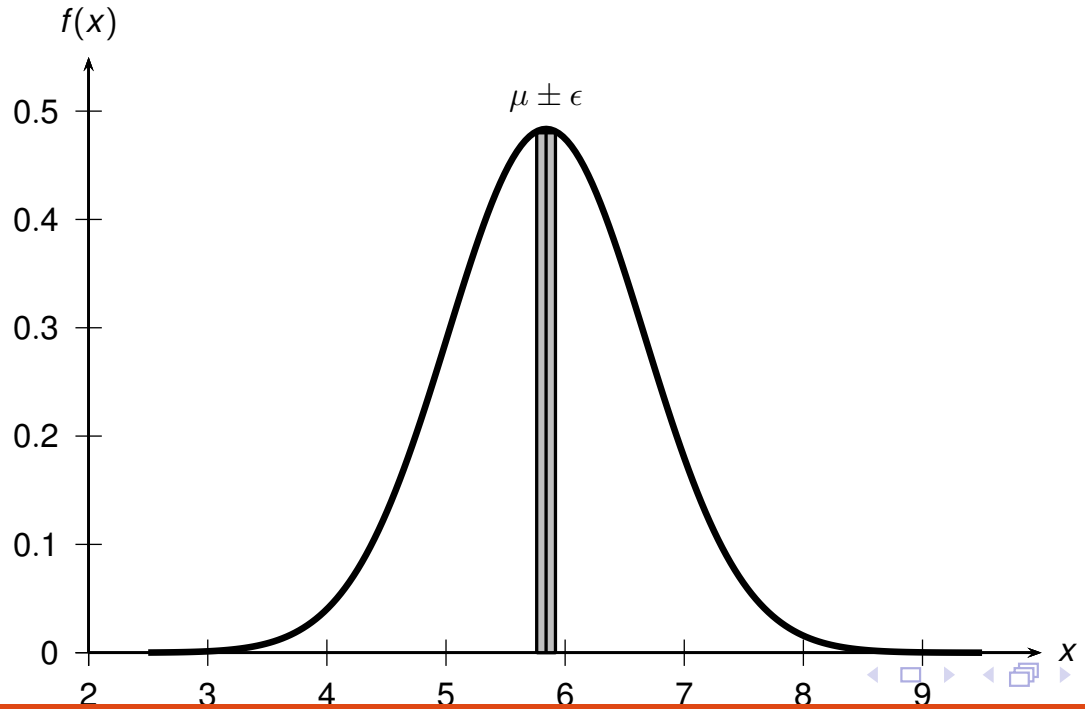
Sepal Length: Normal Distribution

We model sepal length via the *Gaussian* or *normal* density function, given as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean value, and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is the variance.

Normal distribution for sepal length: $\mu = 5.84$, $\sigma^2 = 0.681$



Cumulative Distribution Function

For random variable X, its *cumulative distribution function (CDF)* $F : \mathbb{R} \to [0, 1]$, gives the probability of observing a value at most some given value x:

$$F(x) = P(X \leq x) \qquad \text{for all } -\infty < x < \infty$$

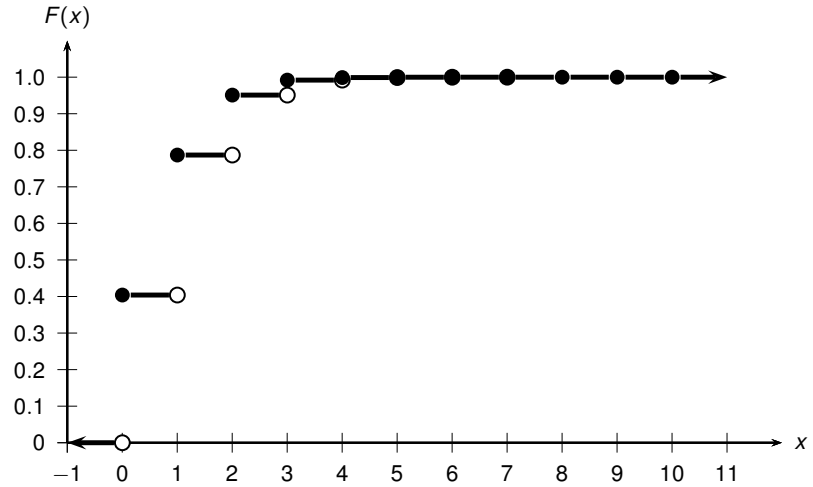
When X is discrete, F is given as

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u)$$

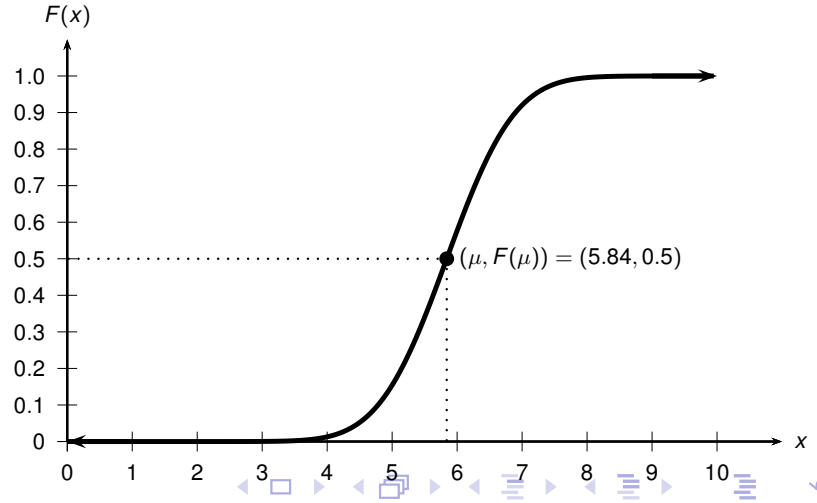
When X is continuous, F is given as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

CDF for binomial distribution ($p = 0.087, m = 10$)



CDF for the normal distribution ($\mu = 5.84, \sigma^2 = 0.681$)



Bivariate Random Variable: Joint Probability Mass Function

Define discrete random variables

$$\text{long sepal length: } X_1(v) = \begin{cases} 1 & \text{if } v \geq 7 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{long sepal width: } X_2(v) = \begin{cases} 1 & \text{if } v \geq 3.5 \\ 0 & \text{otherwise} \end{cases}$$

The bivariate random variable

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

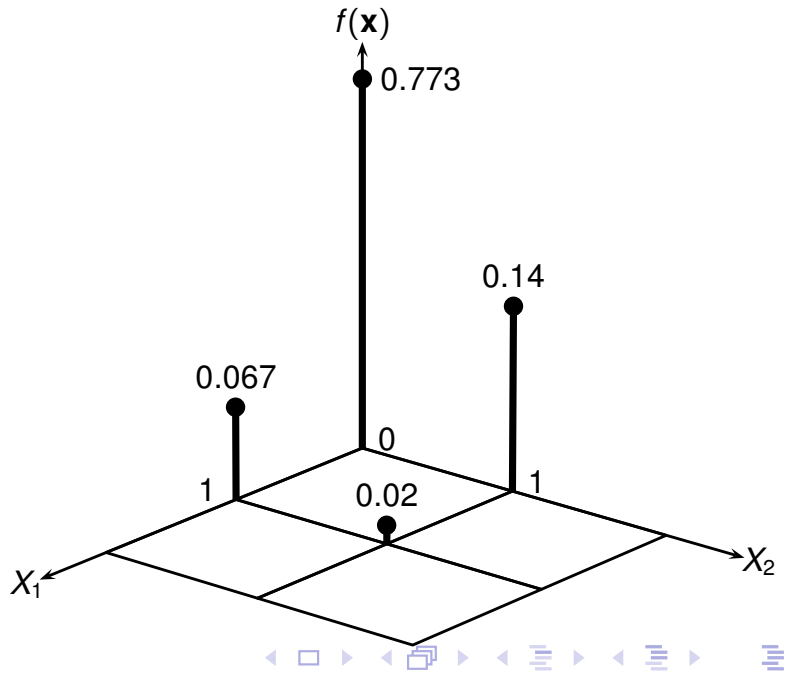
has the joint probability mass function

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$$

$$\text{i.e., } f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

Iris: joint PMF for long sepal length and sepal width

$$f(0,0) = P(X_1 = 0, X_2 = 0) = 116/150 = 0.773$$
$$f(0,1) = P(X_1 = 0, X_2 = 1) = 21/150 = 0.140$$
$$f(1,0) = P(X_1 = 1, X_2 = 0) = 10/150 = 0.067$$
$$f(1,1) = P(X_1 = 1, X_2 = 1) = 3/150 = 0.020$$



Bivariate Random Variable: Probability Density Function

Bivariate Normal: modeling joint distribution for long sepal length (X_1) and sepal width (X_2)

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right\}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ specify the 2D mean and covariance matrix:

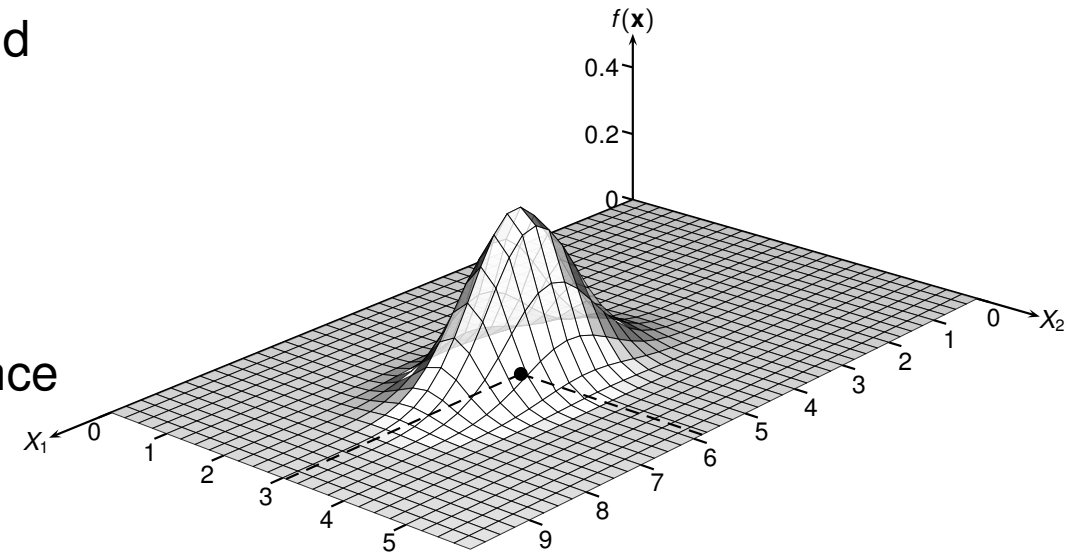
$$\boldsymbol{\mu} = (\mu_1, \mu_2)^T \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

with mean $\mu_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ and covariance $\sigma_{ij} = \frac{1}{n}\sum_{k=1}(x_{ki} - \mu_i)(x_{kj} - \mu_j)$. Also, $\sigma_i^2 = \sigma_{ii}$.

Bivariate Normal

$$\boldsymbol{\mu} = (5.843, 3.054)^T$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$



Random Sample and Statistics

Given a random variable X, a *random sample* of size n from X is defined as a set of n *independent and identically distributed (IID)* random variables

$$S_1, S_2, \ldots, S_n$$

The S_i's have the same probability distribution as X, and are statistically independent.
Two random variables X_1 and X_2 are (statistically) *independent* if, for every $W_1 \subset \mathbb{R}$ and $W_2 \subset \mathbb{R}$, we have

$$P(X_1 \in W_1 \text{ and } X_2 \in W_2) = P(X_1 \in W_1) \cdot P(X_2 \in W_2)$$

which also implies that

$$F(\mathbf{x}) = F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$$
$$f(\mathbf{x}) = f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

where F_i is the cumulative distribution function, and f_i is the probability mass or density function for random variable X_i.

Multivariate Sample

Given dataset $\mathbf{D}$, the n data points $\mathbf{x}_i$ (with $1 \leq i \leq n$) constitute a d-dimensional *multivariate random sample* drawn from the vector random variable $\mathbf{X} = (X_1, X_2, \ldots, X_d)$.

Since the $\mathbf{x}_i$ are assumed to be independent and identically distributed, their joint distribution is given as

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}}(\mathbf{x}_i)$$

where $f_{\mathbf{X}}$ is the probability mass or density function for $\mathbf{X}$.

Assuming that the d attributes $X_1, X_2, \ldots, X_d$ are statistically independent, the joint distribution for the entire dataset is given as:

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n} \prod_{j=1}^{d} f_{X_j}(x_{ij})$$

Sample Statistics

Let $\{\mathbf{S}_i\}_{i=1}^m$ be a random sample of size m drawn from a (multivariate) random variable $\mathbf{X}$. A *statistic* $\hat{\theta}$ is a function

$$\hat{\theta}: (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m) \to \mathbb{R}$$

The statistic is an estimate of the corresponding population parameter θ, where the *population* refers to the entire universe of entities under study. The statistic is itself a random variable.

The *sample mean* is a statistic, defined as the average

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

For `sepal length`, we have $\hat{\mu} = 5.84$, which is an estimator for the (unknown) true population mean sepal length.

Sample Statistics: Variance

The *sample variance* is a statistic

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

For sepal length, we have $\hat{\sigma}^2 = 0.681$.

The *total variance* is a multivariate statistic

$$var(\mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \boldsymbol{\mu}\|^2$$

For the Iris data (with 4 attributes: sepal length and width, petal length and width), we have $var(\mathbf{D}) = 0.868$.