Data mining and Machine learning

Part 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

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

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

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

- **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
$\boldsymbol{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\boldsymbol{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\boldsymbol{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\boldsymbol{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\boldsymbol{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\boldsymbol{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\boldsymbol{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\boldsymbol{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\boldsymbol{x}_{150}$	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 $\boldsymbol{D}$, each row or point is a d-dimensional column vector:

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

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

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

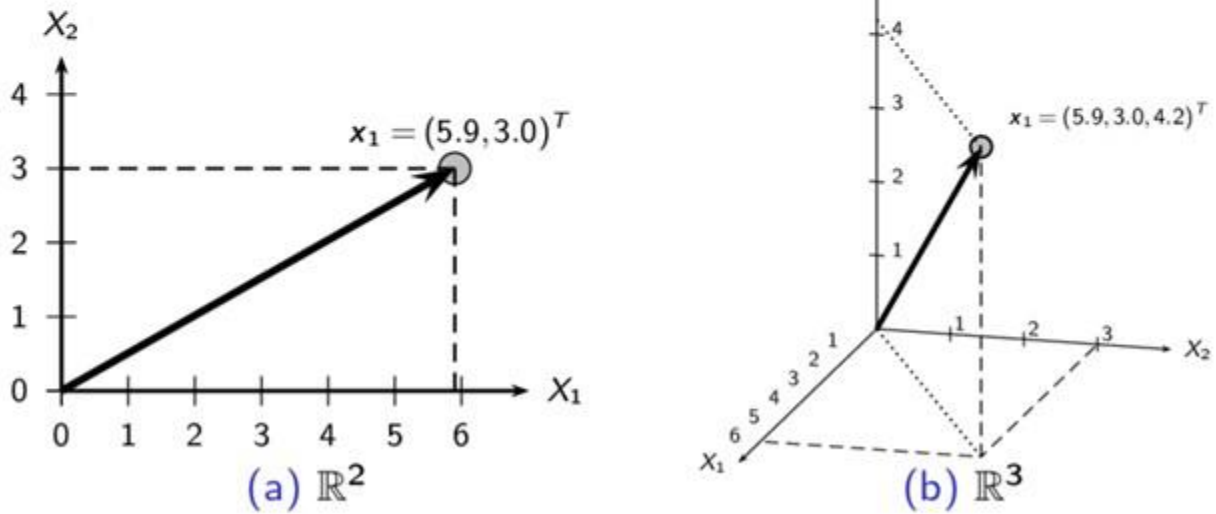
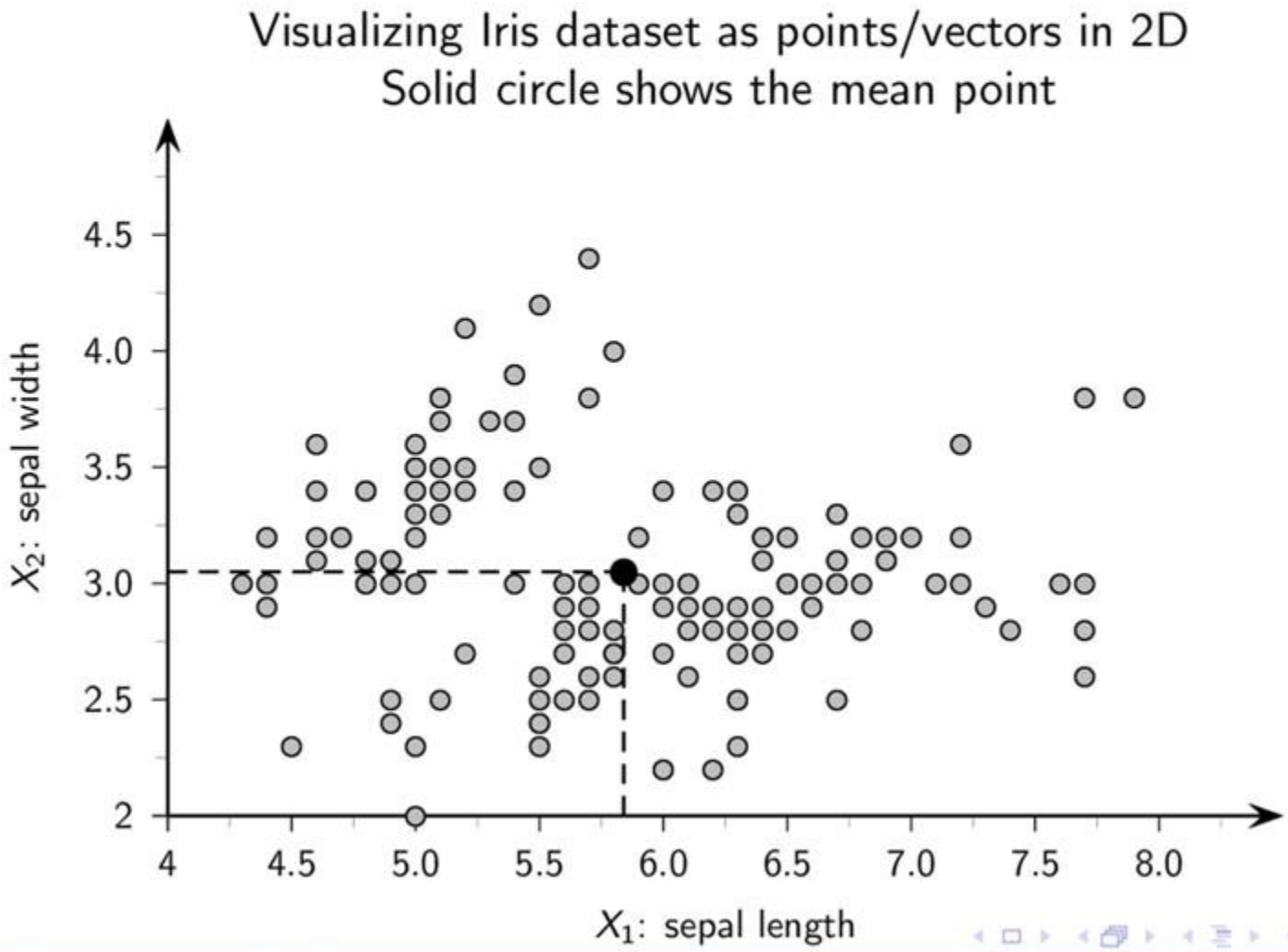


Figure: Projections of $\boldsymbol{x}_1 = (5.9, 3.0, 4.2, 1.5)^T$ in 2D and 3D

Scatterplot: 2D Iris Dataset
sepal length versus sepal width.



Numeric Data Matrix

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

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

The *mean* of the data matrix $\boldsymbol{D}$ is the average of all the points: $\textit{mean}(\boldsymbol{D}) = \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$

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

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

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

Norm, Distance and Angle

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

$$\begin{aligned}\boldsymbol{a}^T \boldsymbol{b} &= a_1 b_1 + a_2 b_2 + \cdots + a_m b_m \\ &= \sum_{i=1}^{m} a_i b_i\end{aligned}$$

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

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

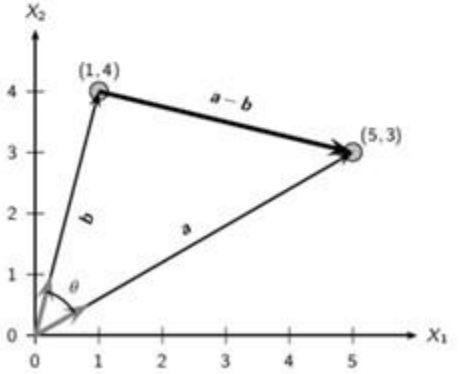
The *unit vector* in the direction of $\boldsymbol{a}$ is $\boldsymbol{u} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}$ with $\|\boldsymbol{a}\| = 1$.

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

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

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

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



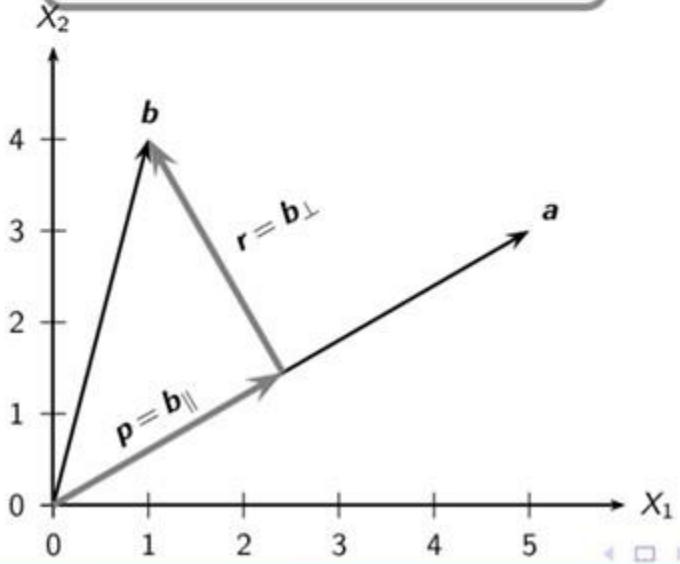
Orthogonal Projection

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

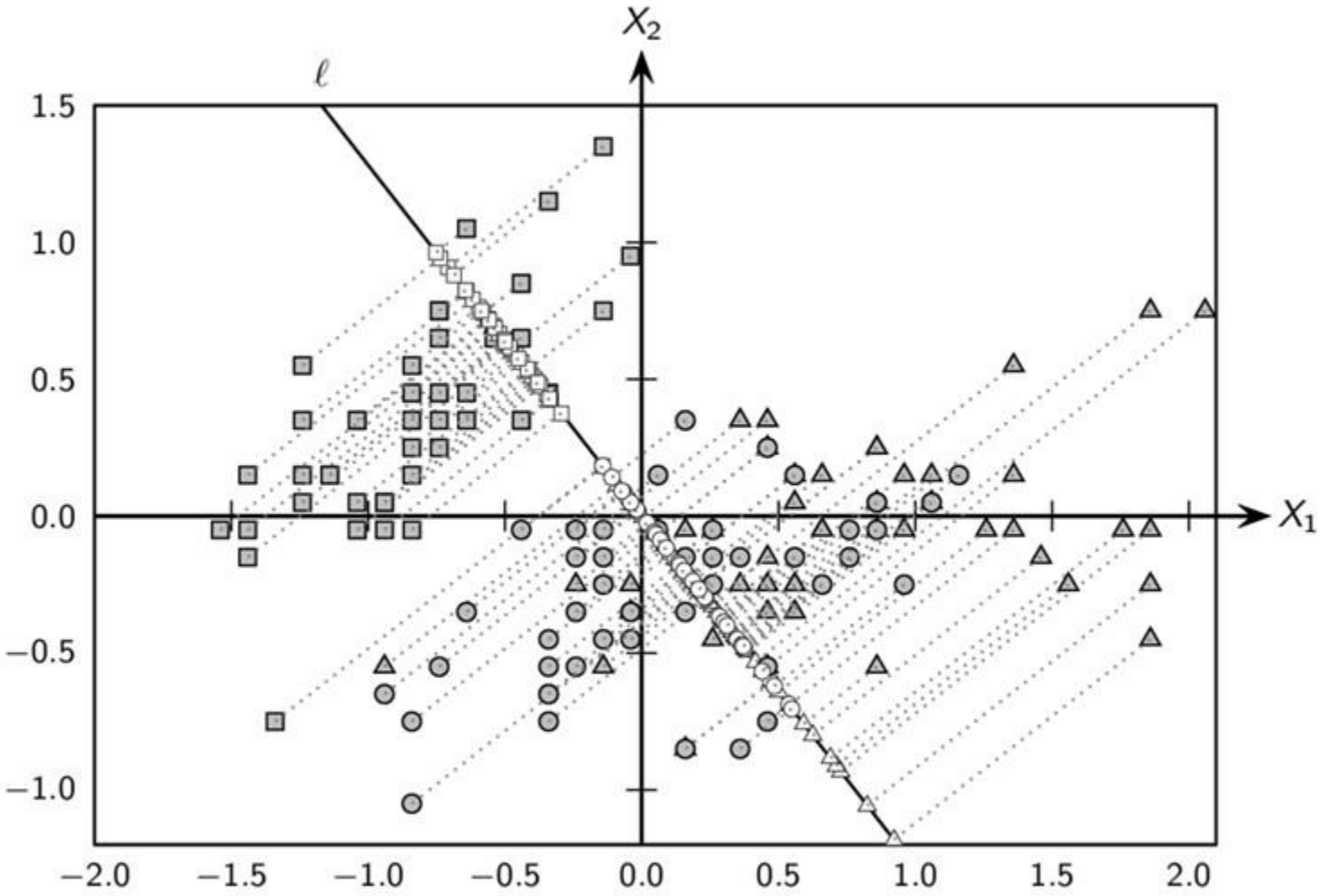
$$\boldsymbol{b} = \boldsymbol{b}_{\parallel} + \boldsymbol{b}_{\perp} = \boldsymbol{p} + \boldsymbol{r}$$

where

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



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

$$X(v) = v$$

$\forall v \in \mathcal{O}$. 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

$$f(x) = P(X = x) \qquad \forall x \in \mathbb{R}$$

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

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

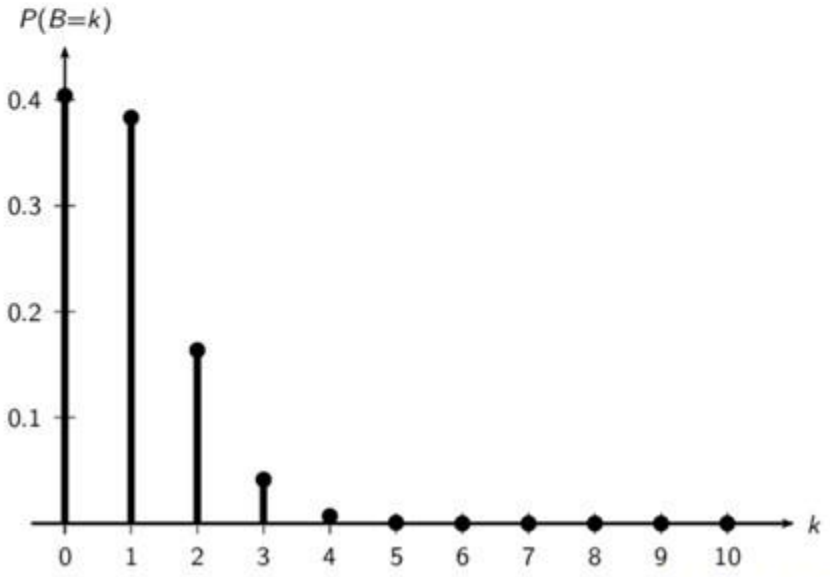
A has a *Bernoulli distribution* with parameter $p \in [0, 1]$, which denotes the probability of a *success*, that is, the probability of picking an Iris with a long sepal length at random from the set of all points.

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

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

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 $P(X = v) = 0 \forall v \in \mathbb{R}$ 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.

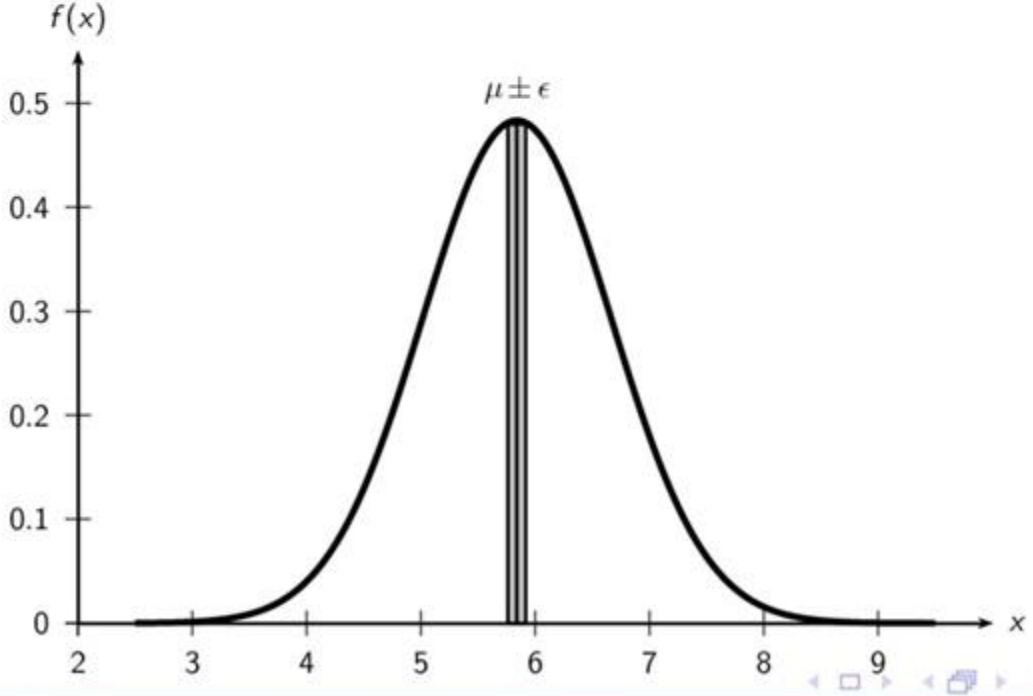
Sepal Length: Normal Distribution

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:

$$F(x) = P(X \leq x) \qquad \forall x | -\infty < x < \infty$$

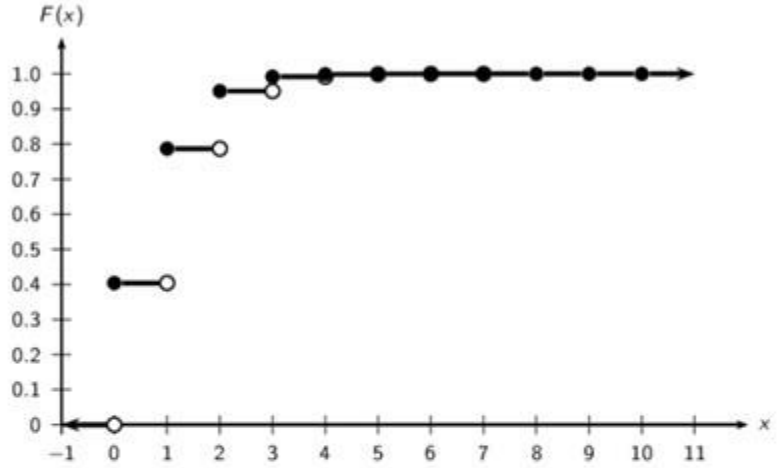
When X is discrete, F is given as

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

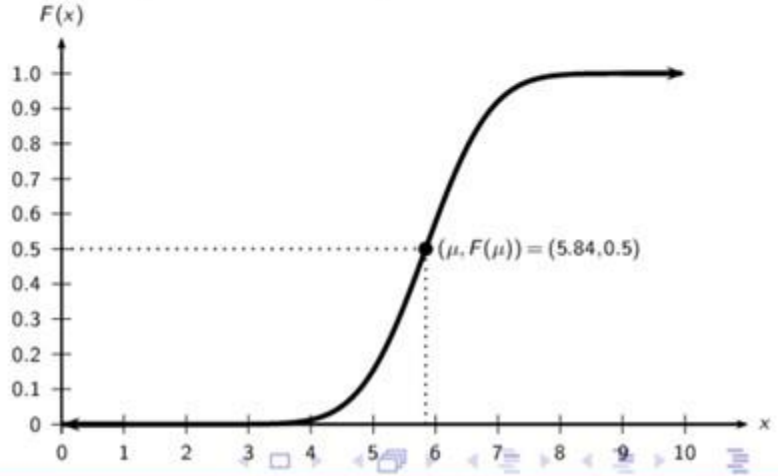
When X is continuous, F is given as

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

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

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

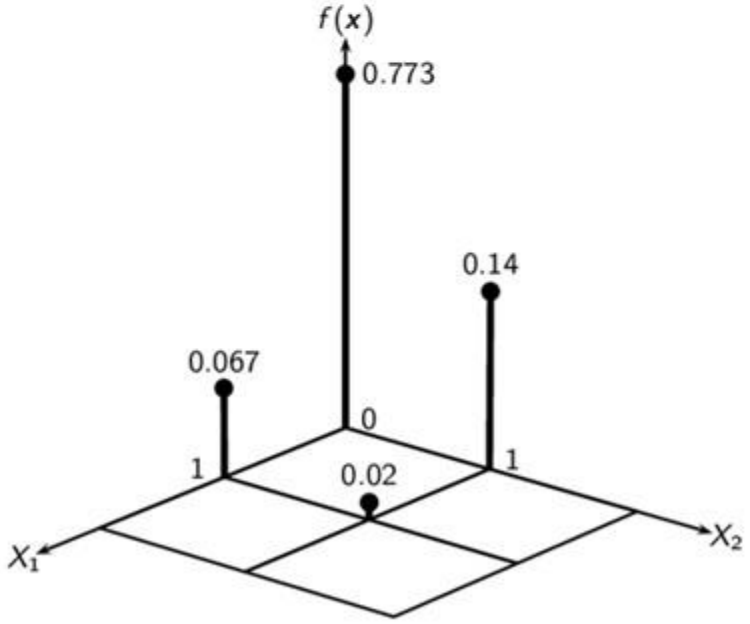
has the joint probability mass function

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

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

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)

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

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

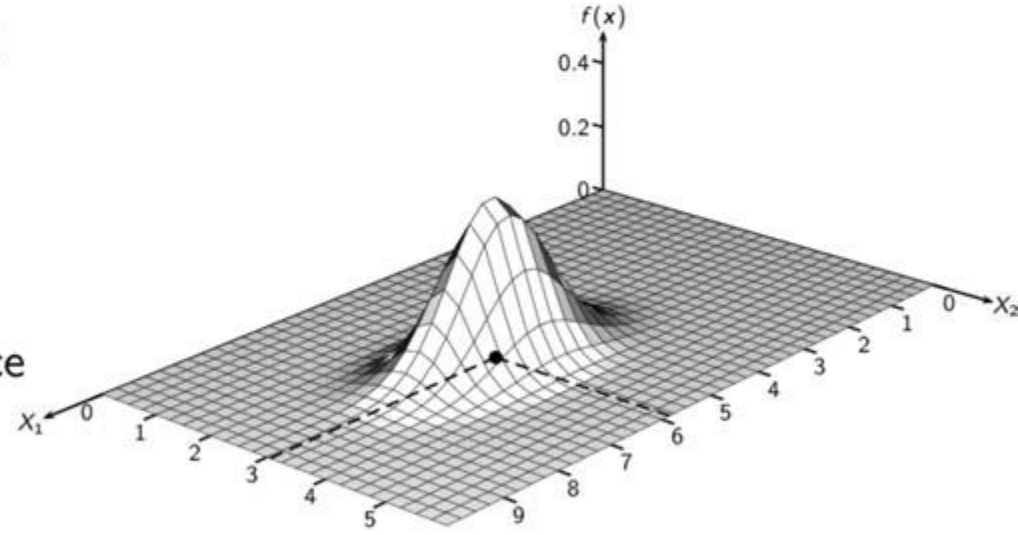
$$\boldsymbol{\mu} = (\mu_1, \mu_2)^T \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

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

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

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



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

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

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

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

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

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

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:

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

Sample Statistics

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

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

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

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

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

Data mining and Machine learning

Part 2. Numeric Attributes

Univariate Analysis

Univariate analysis focuses on a single attribute at a time. The data matrix $\boldsymbol{D}$ is an $n \times 1$ matrix,

$$\boldsymbol{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where X is the numeric attribute of interest, with $x_i \in \mathbb{R}$.

X is assumed to be a random variable, and the observed data a random sample drawn from X, i.e., x_i's are independent and identically distributed as X.

In the vector view, we treat the sample as an n-dimensional vector, and write $X \in \mathbb{R}^n$.

Empirical Probability Mass Function

The *empirical probability mass function (PMF)* of X is given as

$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x)$$

where the indicator variable I takes on the value 1 when its argument is true, and 0 otherwise. The empirical PMF puts a probability mass of $\frac{1}{n}$ at each point x_i.

The *empirical cumulative distribution function (CDF)* of X is given as

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

The *inverse cumulative distribution function* or *quantile function* for X is defined as follows:

$$F^{-1}(q) = \min\{x \mid \hat{F}(x) \geq q\} \qquad \text{for } q \in [0, 1]$$

The inverse CDF gives the least value of X, for which q fraction of the values are higher, and $1 - q$ fraction of the values are lower.

Mean

The *mean* or *expected value* of a random variable X is the arithmetic average of the values of X. It provides a one-number summary of the *location* or *central tendency* for the distribution of X.

If X is discrete, it is defined as

$$\mu = E[X] = \sum_x x \cdot f(x)$$

where $f(x)$ is the probability mass function of X.

If X is continuous it is defined as

$$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

where $f(x)$ is the probability density function of X.

Sample Mean

The *sample mean* is a statistic, that is, a function $\hat{\mu} : \{x_1, x_2, \ldots, x_n\} \to \mathbb{R}$, defined as the average value of x_i's:

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

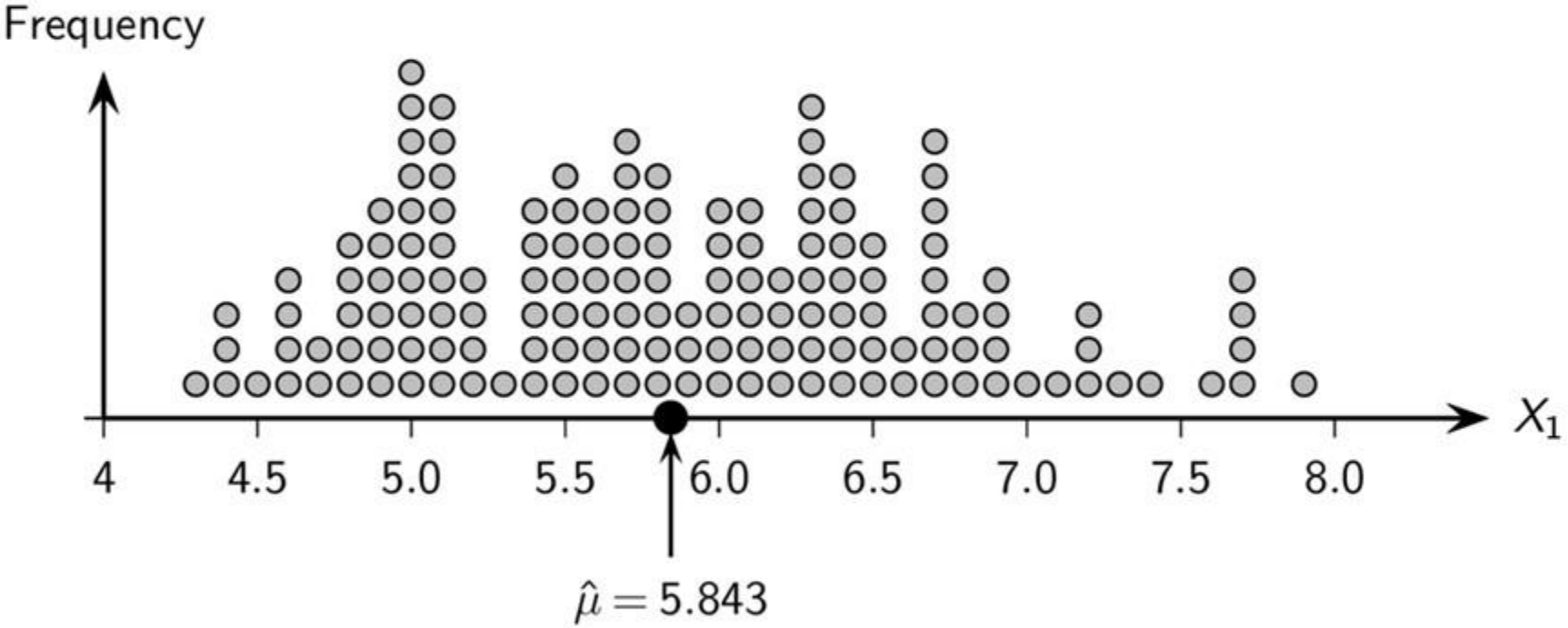
It serves as an estimator for the unknown mean value μ of X.

An estimator $\hat{\theta}$ is called an *unbiased estimator* for parameter θ if $E[\hat{\theta}] = \theta$ for every possible value of θ. The sample mean $\hat{\mu}$ is an unbiased estimator for the population mean μ, as

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

We say that a statistic is *robust* if it is not affected by extreme values (such as outliers) in the data. The sample mean is not robust because a single large value can skew the average.

Sample Mean: Iris sepal length



Median

The *median* of a random variable is defined as the value m such that

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

The median m is the "middle-most" value; half of the values of X are less and half of the values of X are more than m.

In terms of the (inverse) cumulative distribution function, the median is the value m for which

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

The *sample median* is given as

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

Median is robust, as it is not affected very much by extreme values.

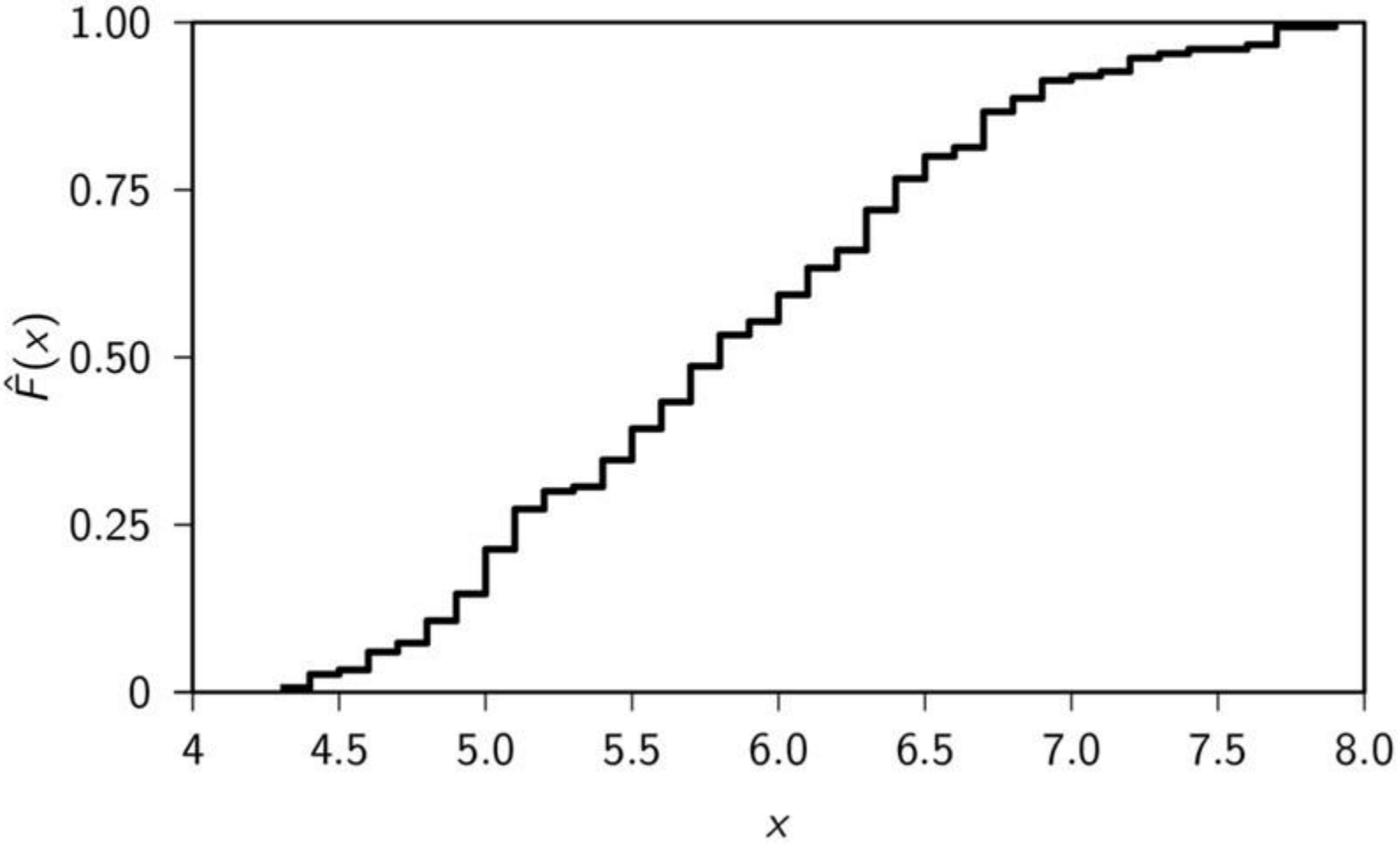
Mode

The *mode* of a random variable X is the value at which the probability mass function or the probability density function attains its maximum value, depending on whether X is discrete or continuous, respectively.

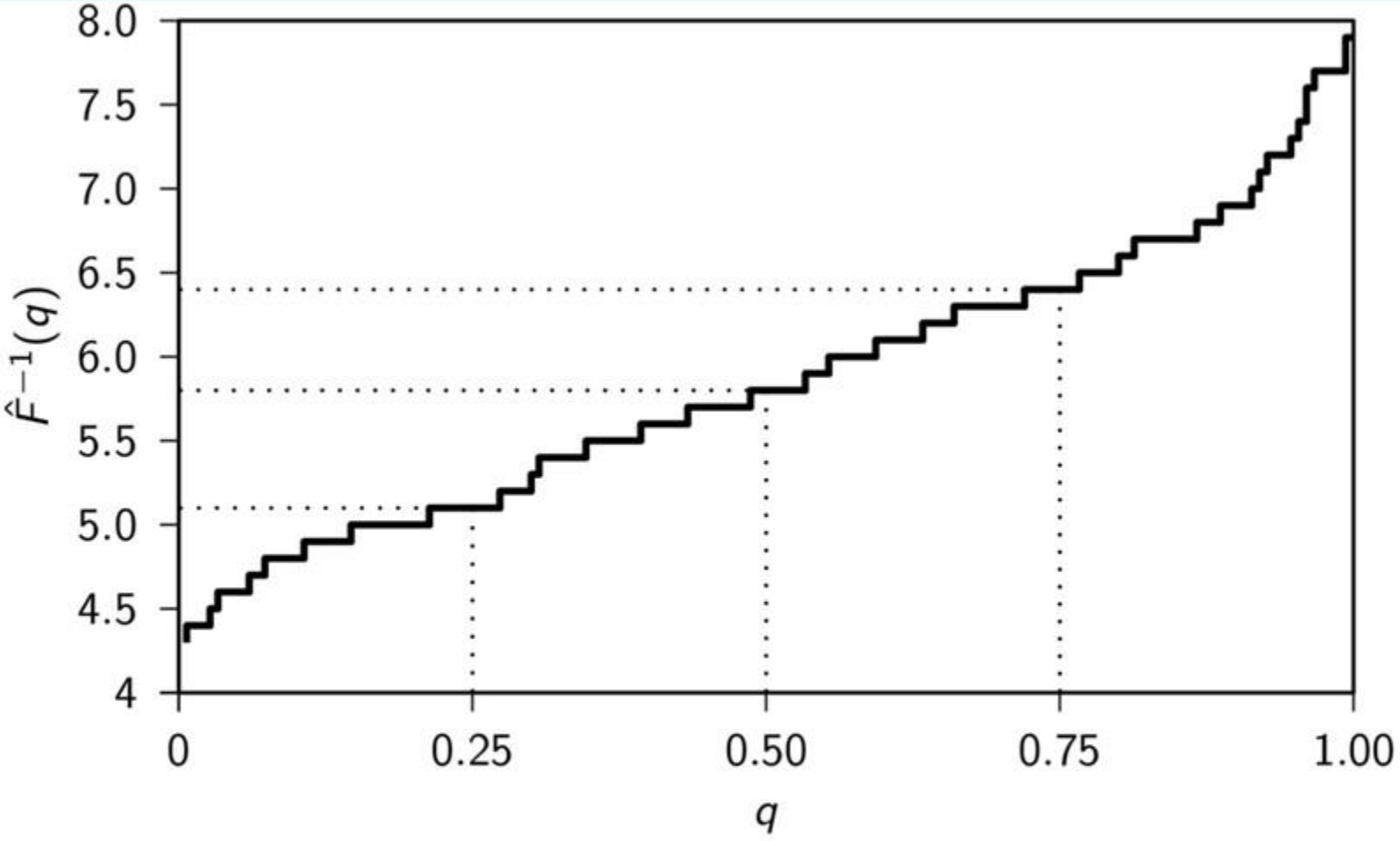
The *sample mode* is a value for which the empirical probability mass function attains its maximum, given as

$$\text{mode}(X) = \arg\max_x \hat{f}(x)$$

Empirical CDF: sepal length



Empirical Inverse CDF: sepal length



The median is 5.8, since

$$\hat{F}(5.8) = 0.5 \text{ or } 5.8 = \hat{F}^{-1}(0.5)$$

Range

The *value range* or simply *range* of a random variable X is the difference between the maximum and minimum values of X, given as

$$r = \max\{X\} - \min\{X\}$$

The *sample range* is a statistic, given as

$$\hat{r} = \max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\}$$

Range is sensitive to extreme values, and thus is not robust.

A more robust measure of the dispersion of X is the *interquartile range (IQR)*, defined as

$$IQR = F^{-1}(0.75) - F^{-1}(0.25)$$

The *sample IQR* is given as

$$\widehat{IQR} = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

Variance and Standard Deviation

The *variance* of a random variable X provides a measure of how much the values of X deviate from the mean or expected value of X

$$\sigma^2 = \text{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

The *standard deviation* σ, is the positive square root of the variance, σ^2.

The *sample variance* is defined as

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

and the *sample standard deviation* is

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

Geometric Interpretation of Sample Variance

The sample values for X comprise a vector in n-dimensional space, where n is the sample size. Let Z denote the centered sample

$$Z = X - 1 \cdot \hat{\mu} = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$$

where $1 \in \mathbb{R}^n$ is the vector of ones.
Sample variance is squared magnitude of the centered attribute vector, normalized by the sample size:

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

Variance of the Sample Mean and Bias

Sample mean $\hat{\mu}$ is itself a statistic. We can compute its mean value and variance

$$E[\hat{\mu}] = \mu$$

$$var(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$$

The sample mean $\hat{\mu}$ varies or deviates from the mean μ in proportion to the population variance σ^2. However, the deviation can be made smaller by considering larger sample size n.

The sample variance is a *biased estimator* for the true population variance, since

$$E[\hat{\sigma}^2] = \left(\frac{n-1}{n}\right)\sigma^2$$

But it is asymptotically unbiased, since

$$E[\hat{\sigma}^2] \to \sigma^2 \qquad \text{as } n \to \infty$$

Bivariate Analysis

In bivariate analysis, we consider two attributes at the same time. The data $\boldsymbol{D}$ comprises an $n \times 2$ matrix:

$$\boldsymbol{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Geometrically, $\boldsymbol{D}$ comprises n points or vectors in 2-dimensional space

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$$

$\boldsymbol{D}$ can also be viewed as two points or vectors in an n-dimensional space:

$$X_1 = (x_{11}, x_{21}, \ldots, x_{n1})^T$$
$$X_2 = (x_{12}, x_{22}, \ldots, x_{n2})^T$$

In the probabilistic view, $\boldsymbol{X} = (X_1, X_2)^T$ is a bivariate vector random variable, and the points $\boldsymbol{x}_i$ $(1 \leq i \leq n)$ are a random sample drawn from $\boldsymbol{X}$, that is, $\boldsymbol{x}_i$'s IID with $\boldsymbol{X}$.

Bivariate Mean and Variance

The bivariate mean is defined as the expected value of the vector random variable $\boldsymbol{X}$:

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

The sample mean vector is given as

$$\hat{\boldsymbol{\mu}} = \sum_{\boldsymbol{x}} \boldsymbol{x} \hat{f}(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \boldsymbol{x} \left(\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{x}_i = \boldsymbol{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

Covariance

The *covariance* between two attributes X_1 and X_2 provides a measure of the association or linear dependence between them, and is defined as

$$\begin{aligned}\sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1 X_2] - E[X_1]E[X_2]\end{aligned}$$

If X_1 and X_2 are independent, then

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

which implies that $\sigma_{12} = 0$.

The *sample covariance* between X_1 and X_2 is given as

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$

Correlation

The *correlation* between variables X_1 and X_2 is the *standardized covariance*, obtained by normalizing the covariance with the standard deviation of each variable, given as

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

The *sample correlation* for attributes X_1 and X_2 is given as

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n}(x_{i2} - \hat{\mu}_2)^2}}$$

Geometric Interpretation of Sample Covariance and Correlation

Let $\bar{X}_1$ and $\bar{X}_2$ denote the centered attribute vectors in $\mathbb{R}^n$:

$$\bar{X}_1 = X_1 - 1 \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \qquad \bar{X}_2 = X_2 - 1 \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

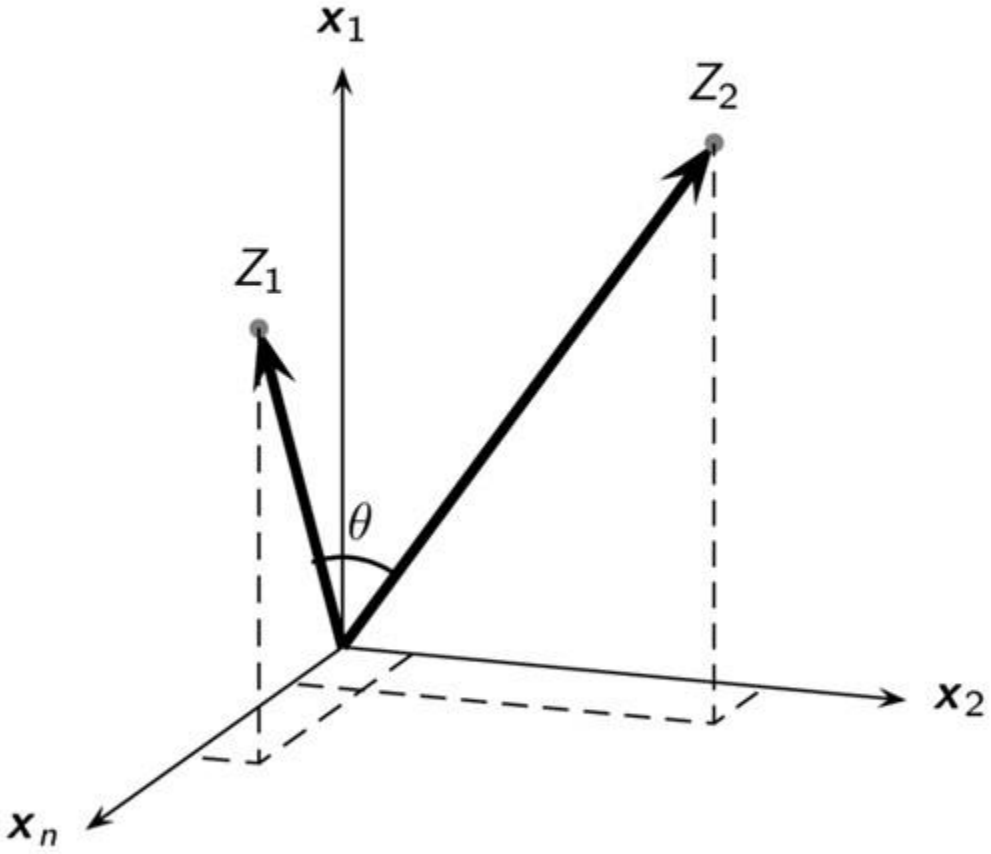
The sample covariance and the sample correlation are given as

$$\hat{\sigma}_{12} = \frac{\bar{X}_1^T \bar{X}_2}{n}$$

$$\hat{\rho}_{12} = \frac{\bar{X}_1^T \bar{X}_2}{\sqrt{\bar{X}_1^T \bar{X}_1}\sqrt{\bar{X}_2^T \bar{X}_2}} = \frac{\bar{X}_1^T \bar{X}_2}{\|\bar{X}_1\| \, \|\bar{X}_2\|} = \left(\frac{\bar{X}_1}{\|\bar{X}_1\|}\right)^T \left(\frac{\bar{X}_2}{\|\bar{X}_2\|}\right) = \cos\theta$$

The correlation coefficient is simply the cosine of the angle between the two centered attribute vectors.

Geometric Interpretation of Covariance and Correlation



Covariance Matrix

The variance–covariance information for the two attributes X_1 and X_2 can be summarized in the square 2×2 *covariance matrix*

$$\Sigma = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$$
$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

Because $\sigma_{12} = \sigma_{21}$, Σ is *symmetric*.
The *total variance* is given as

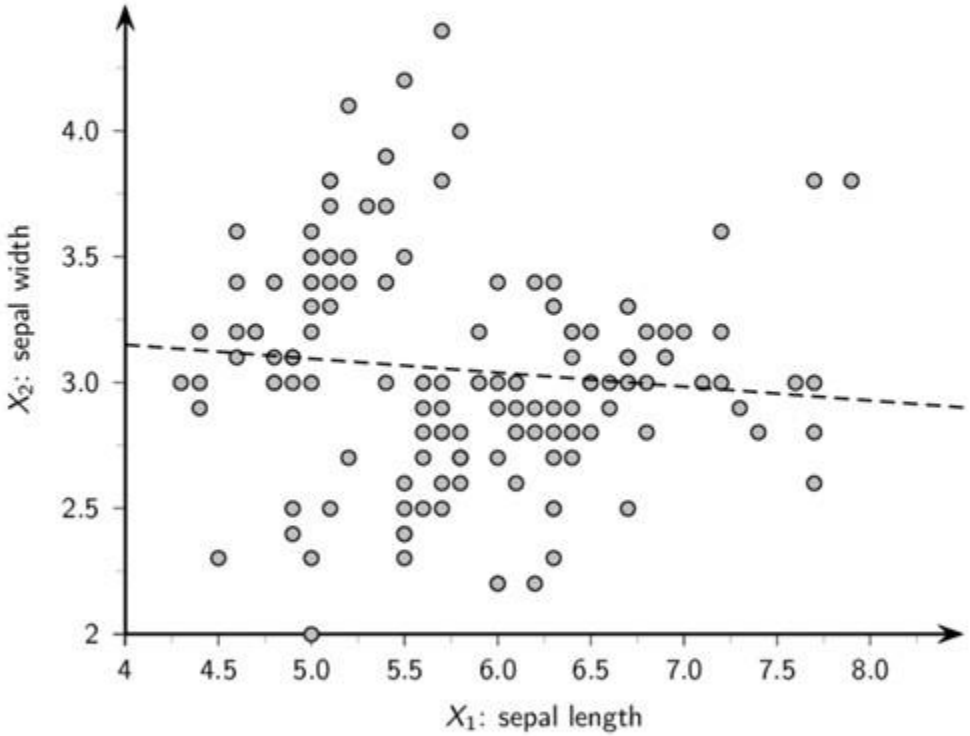
$$var(\boldsymbol{D}) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2$$

We immediately have $tr(\Sigma) \geq 0$.
The *generalized variance* is

$$|\Sigma| = \det(\Sigma) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2 = (1 - \rho_{12}^2)\sigma_1^2\sigma_2^2$$

Note that $|\rho_{12}| \leq 1$ implies that $\det(\Sigma) \geq 0$.

Correlation: sepal length and sepal width



The sample mean is

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

The sample covariance matrix is

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

The sample correlation is

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Multivariate Analysis

In multivariate analysis we consider all the d numeric attributes $X_1, X_2, \ldots, X_d$.

$$\boldsymbol{D} = \begin{array}{c} \begin{pmatrix} X_1 & X_2 & \cdots & X_d \end{pmatrix} \\ \hline \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \end{array}$$

In the row view, the data is a set of n points or vectors in the d-dimensional attribute space

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T \in \mathbb{R}^d$$

In the column view, the data is a set of d points or vectors in the n-dimensional space spanned by the data points

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in \mathbb{R}^n$$

Mean and Covariance

In the probabilistic view, the d attributes are modeled as a vector random variable, $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^T$, and the points $\boldsymbol{x}_i$ are considered to be a random sample drawn from $\boldsymbol{X}$, i.e., IID with $\boldsymbol{X}$.
The *multivariate mean vector* is

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_d \end{pmatrix}^T$$

The *sample mean* is

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

The (sample) covariance matrix is a $d \times d$ (square) symmetric matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix} \qquad \hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix}$$

Covariance Matrix is Positive Semidefinite

Σ is a *positive semidefinite* matrix, that is,

$$\boldsymbol{a}^T \Sigma \boldsymbol{a} \geq 0 \text{ for any } d\text{-dimensional vector } \boldsymbol{a}$$

To see this, observe that

$$\begin{aligned} \boldsymbol{a}^T \Sigma \boldsymbol{a} &= \boldsymbol{a}^T E\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\right] \boldsymbol{a} \\ &= E\left[\boldsymbol{a}^T (\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{a}\right] \\ &= E\left[Y^2\right] \\ &\geq 0 \end{aligned}$$

Because Σ is also symmetric, this implies that all the eigenvalues of Σ are real and non-negative, and they can be arranged from the largest to the smallest as follows: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$.

The total variance is given as: $var(\boldsymbol{D}) = \sum_{i=1}^{d} \sigma_i^2$
The generalized variance is $\det(\Sigma) = \prod_{i=1}^{d} \lambda_i \geq 0$

Sample Covariance Matrix: Inner and Outer Product

Let $\overline{\boldsymbol{D}}$ represent the centered data matrix

$$\overline{\boldsymbol{D}} = \boldsymbol{D} - 1 \cdot \hat{\boldsymbol{\mu}}^T = \begin{pmatrix} \boldsymbol{x}_1^T - \hat{\boldsymbol{\mu}}^T \\ \boldsymbol{x}_2^T - \hat{\boldsymbol{\mu}}^T \\ \vdots \\ \boldsymbol{x}_n^T - \hat{\boldsymbol{\mu}}^T \end{pmatrix} = \begin{pmatrix} - & \overline{\boldsymbol{x}}_1^T & - \\ - & \overline{\boldsymbol{x}}_2^T & - \\ & \vdots & \\ - & \overline{\boldsymbol{x}}_n^T & - \end{pmatrix}$$

Inner product and outer product form for sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n}\left(\overline{\boldsymbol{D}}^T\,\overline{\boldsymbol{D}}\right) = \frac{1}{n}\begin{pmatrix} \overline{X}_1^T\overline{X}_1 & \overline{X}_1^T\overline{X}_2 & \cdots & \overline{X}_1^T\overline{X}_d \\ \overline{X}_2^T\overline{X}_1 & \overline{X}_2^T\overline{X}_2 & \cdots & \overline{X}_2^T\overline{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X}_d^T\overline{X}_1 & \overline{X}_d^T\overline{X}_2 & \cdots & \overline{X}_d^T\overline{X}_d \end{pmatrix} \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{x}}_i \cdot \overline{\boldsymbol{x}}_i^T$$

i.e., $\hat{\Sigma}$ is given as the pairwise *inner or dot products* of the centered attribute vectors, normalized by the sample size, or as a sum of rank-one matrices obtained as the *outer product* of each centered point.

Data Normalization

If the attribute values are in vastly different scales, then it is necessary to normalize them.

Range Normalization: Let X be an attribute and let $x_1, x_2, \ldots, x_n$ be a random sample drawn from X. In *range normalization* each value is scaled by the sample range $\hat{r}$ of X:

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

After transformation the new attribute takes on values in the range $[0, 1]$.

Standard Score Normalization: Also called z-normalization; each value is replaced by its z-score:

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

where $\hat{\mu}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance of X. After transformation, the new attribute has mean $\hat{\mu}' = 0$, and standard deviation $\hat{\sigma}' = 1$.

Normalization Example

x_i	Age (X_1)	Income (X_2)
x_1	12	300
x_2	14	500
x_3	18	1000
x_4	23	2000
x_5	27	3500
x_6	28	4000
x_7	34	4300
x_8	37	6000
x_9	39	2500
x_{10}	40	2700

Since `Income` is much larger, it dominates `Age`. The sample range for `Age` is $\hat{r} = 40 - 12 = 28$, whereas for `Income` it is $6000 - 300 = 5700$. For range normalization, the point $x_2 = (14, 500)$ is scaled to

$$x_2' = \left(\frac{14 - 12}{28}, \frac{500 - 300}{5700}\right) = (0.071, 0.035)$$

For z-normalization, we have

	Age	Income
$\hat{\mu}$	27.2	2680
$\hat{\sigma}$	9.77	1726.15

Thus, $x_2 = (14, 500)$ is scaled to

$$x_2' = \left(\frac{14 - 27.2}{9.77}, \frac{500 - 2680}{1726.15}\right) = (-1.35, -1.26)$$

Univariate Normal Distribution

The normal distribution plays an important role as the parametric distribution of choice in clustering, density estimation, and classification.

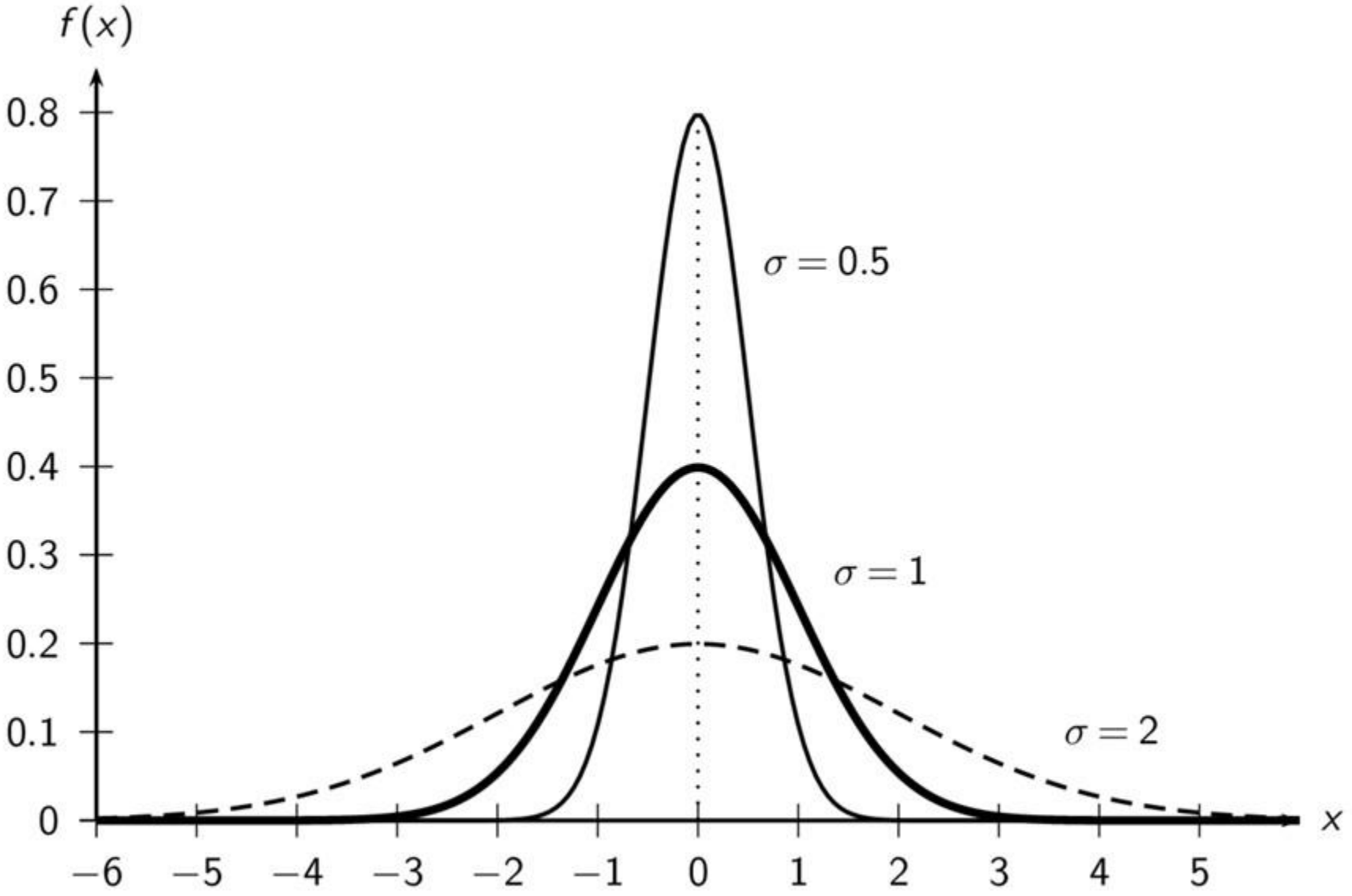
A random variable X has a normal distribution, with the parameters mean μ and variance σ^2, if the probability density function of X is given as follows:

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

The term $(x-\mu)^2$ measures the distance of a value x from the mean μ of the distribution, and thus the probability density decreases exponentially as a function of the distance from the mean.

The maximum value of the density occurs at the mean value $x = \mu$, given as $f(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}$, which is inversely proportional to the standard deviation σ of the distribution.

Normal Distribution: $\mu = 0$, and Different Variances



Multivariate Normal Distribution

Given the d-dimensional vector random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^T$, it has a multivariate normal distribution, with the parameters mean $\boldsymbol{\mu}$ and covariance matrix Σ, if its joint multivariate probability density function is given as follows:

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

where $|\Sigma|$ is the determinant of the covariance matrix.

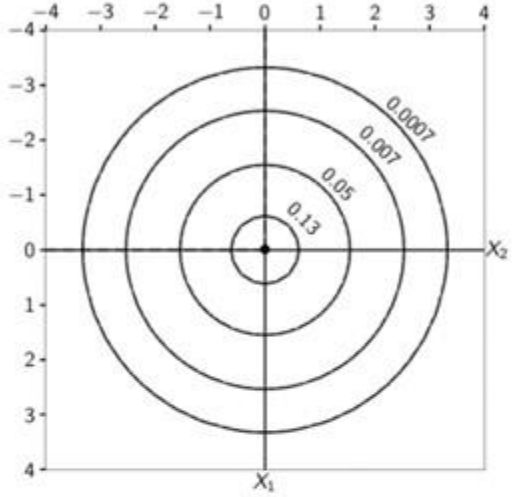
The term

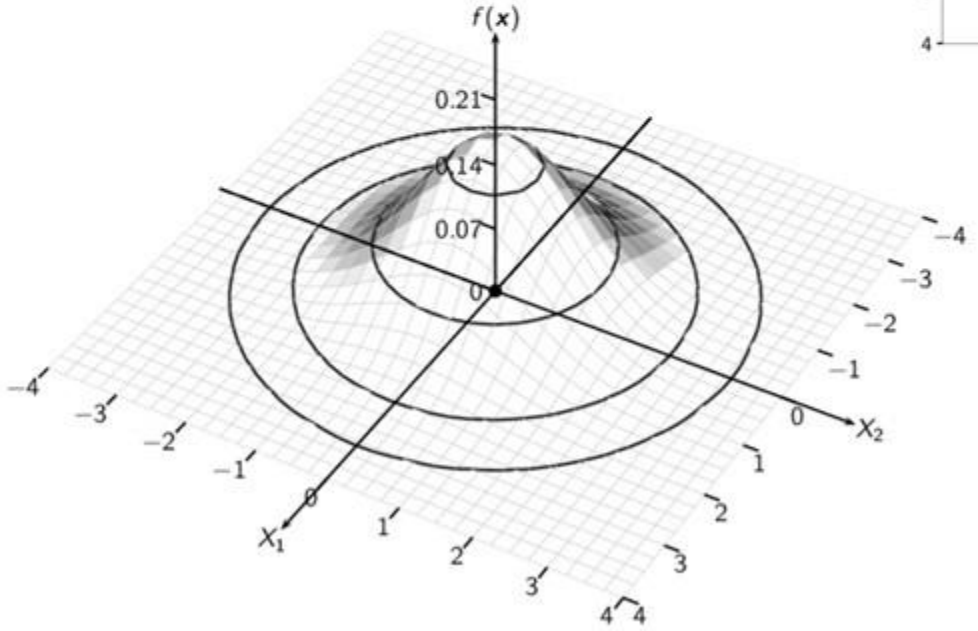
$$(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

measures the distance, called the *Mahalanobis distance* of the point $\boldsymbol{x}$ from the mean $\boldsymbol{\mu}$ of the distribution, taking into account all of the variance–covariance information between the attributes.

Standard Bivariate Normal Density

Parameters: $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$





Geometry of the Multivariate Normal

Compared to the standard multivariate normal, the mean $\boldsymbol{\mu}$ translates the center of the distribution, whereas the covariance matrix Σ scales and rotates the distribution. The eigen-decomposition of Σ is given as

$$\Sigma \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$ are the eigenvalues and $\boldsymbol{u}_i$ the corresponding eigenvectors. This can be expressed compactly as follows:

$$\Sigma = \boldsymbol{U} \Lambda \boldsymbol{U}^T$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \qquad \boldsymbol{U} = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_d \\ | & | & & | \end{pmatrix}$$

The eigenvectors represent the new basis vectors, with the covariance matrix given by Λ (all covariances become zero). Since the trace of a square matrix is invariant to similarity transformation, such as a change of basis, we have

$$var(\boldsymbol{D}) = tr(\Sigma) = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i = tr(\Lambda)$$

Bivariate Normal for Iris: sepal length and sepal width

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

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

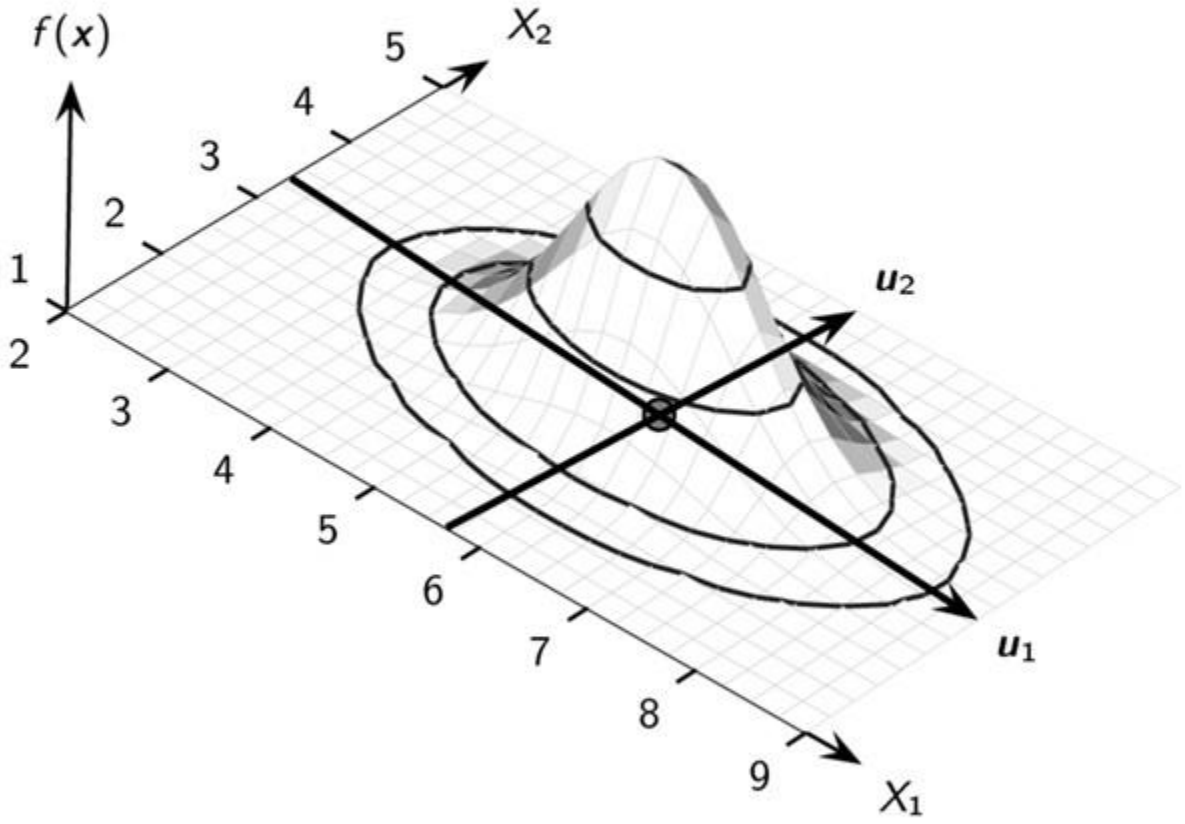
We have

$$\hat{\Sigma} = \boldsymbol{U} \Lambda \boldsymbol{U}^T$$

$$\boldsymbol{U} = \begin{pmatrix} -0.997 & -0.078 \\ 0.078 & -0.997 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix}$$

Angle of rotation is:
$\cos\theta = \boldsymbol{e}_1^T \boldsymbol{u}_1 = -0.997$
or $\theta = 175.5^\circ$



Data mining and Machine learning

Part 3. Categorical Attributes

Univariate Analysis: Bernoulli Variable

Consider a single categorical attribute, X, with domain $dom(X) = \{a_1, a_2, \ldots, a_m\}$ comprising m symbolic values. The data $\boldsymbol{D}$ is an $n \times 1$ symbolic data matrix given as

$$\boldsymbol{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where each point $x_i \in dom(X)$.

Bernoulli Variable: Special case when $m = 2$

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

i.e., $dom(X) = \{0, 1\}$.

Bernoulli Variable: Mean and Variance

The probability mass function (PMF) of X is given as

$$P(X = x) = f(x) = p^x(1-p)^{1-x}$$

The expected value of X is given as

$$\mu = E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

and the variance of X is given as

$$\sigma^2 = var(X) = p(1-p)$$

Assume that each symbolic point has been mapped to its binary value. The set $\{x_1, x_2, \ldots, x_n\}$ is a random sample drawn from X.

The sample mean is given as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n} = \hat{p}$$

where n_i is the number of points with $x_j = i$ in the random sample (equal to the number of occurrences of symbol a_i).

The sample variance is given as

$$\hat{\sigma}^2 = \hat{p}(1-\hat{p})$$

Binomial Distribution: Number of Occurrences

Given the Bernoulli variable X, let $\{x_1, x_2, \ldots, x_n\}$ be a random sample of size n. Let N be the random variable denoting the number of occurrences of the symbol a_1 (value $X = 1$). N has a binomial distribution, given as

$$f(N = n_1 | \, n, p) = \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}$$

N is the sum of the n independent Bernoulli random variables x_i IID with X, that is, $N = \sum_{i=1}^{n} x_i$. The mean or expected number of occurrences of a_1 is

$$\mu_N = E[N] = E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x_i] = \sum_{i=1}^{n} p = np$$

The variance of N is

$$\sigma_N^2 = var(N) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Multivariate Bernoulli Variable

For the general case when $dom(X) = \{a_1, a_2, \ldots, a_m\}$, we model X as an m-dimensional or *multivariate Bernoulli random variable* $\boldsymbol{X} = (A_1, A_2, \ldots, A_m)^T$, where each A_i is a Bernoulli variable with parameter p_i denoting the probability of observing symbol a_i.

However, X can assume only one of the symbolic values at any one time. Thus,

$$\boldsymbol{X}(v) = \boldsymbol{e}_i \text{ if } v = a_i$$

where $\boldsymbol{e}_i$ is the i-th standard basis vector in m dimensions. The range of $\boldsymbol{X}$ consists of m distinct vector values $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m\}$.

The PMF of $\boldsymbol{X}$ is

$$P(\boldsymbol{X} = \boldsymbol{e}_i) = f(\boldsymbol{e}_i) = p_i = \prod_{j=1}^{m} p_j^{e_{ij}}$$

with $\sum_{i=1}^{m} p_i = 1$.

Multivariate Bernoulli: Mean

The mean or expected value of $\boldsymbol{X}$ can be obtained as

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = \sum_{i=1}^{m} \boldsymbol{e}_i f(\boldsymbol{e}_i) = \sum_{i=1}^{m} \boldsymbol{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \boldsymbol{p}$$

The sample mean is

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = \sum_{i=1}^{m} \frac{n_i}{n} \boldsymbol{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\boldsymbol{p}}$$

where n_i is the number of occurrences of the vector value $\boldsymbol{e}_i$ in the sample, i.e., the number of occurrences of the symbol a_i. Furthermore, $\sum_{i=1}^{m} n_i = n$.

Multivariate Bernoulli Variable: sepal length

Bins	Domain	Counts
$[4.3, 5.2]$	Very Short (a_1)	$n_1 = 45$
$(5.2, 6.1]$	Short (a_2)	$n_2 = 50$
$(6.1, 7.0]$	Long (a_3)	$n_3 = 43$
$(7.0, 7.9]$	Very Long (a_4)	$n_4 = 12$

We model sepal length as a multivariate Bernoulli variable $\boldsymbol{X}$

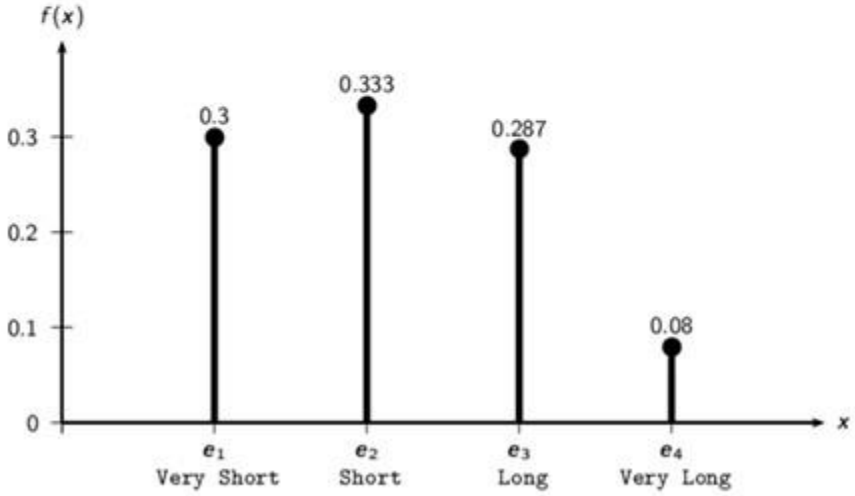
$$\boldsymbol{X}(v) = \begin{cases} \boldsymbol{e}_1 = (1,0,0,0) & \text{if } v = a_1 \\ \boldsymbol{e}_2 = (0,1,0,0) & \text{if } v = a_2 \\ \boldsymbol{e}_3 = (0,0,1,0) & \text{if } v = a_3 \\ \boldsymbol{e}_4 = (0,0,0,1) & \text{if } v = a_4 \end{cases}$$

For example, the symbolic point $x_1 = \texttt{Short} = a_2$ is represented as the vector $(0,1,0,0)^T = \boldsymbol{e}_2$.

Probability Mass Function

The total sample size is $n = 150$; the estimates $\hat{p}_i$ are:

$$\hat{p}_1 = 45/150 = 0.3$$
$$\hat{p}_2 = 50/150 = 0.333$$
$$\hat{p}_3 = 43/150 = 0.287$$
$$\hat{p}_4 = 12/150 = 0.08$$



Multivariate Bernoulli Variable: Covariance Matrix

We have $\boldsymbol{X} = (A_1, A_2, \ldots, A_m)^T$, where A_i is the Bernoulli variable corresponding to symbol a_i. The variance for each Bernoulli variable A_i is

$$\sigma_i^2 = var(A_i) = p_i(1 - p_i)$$

The covariance between A_i and A_j is

$$\sigma_{ij} = E[A_i A_j] - E[A_i] \cdot E[A_j] = 0 - p_i p_j = -p_i p_j$$

Negative relationship since A_i and A_j cannot both be 1 at the same time.
The covariance matrix for $\boldsymbol{X}$ is given as

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 p_m & -p_2 p_m & \cdots & p_m(1-p_m) \end{pmatrix}$$

More compactly $\Sigma = diag(\boldsymbol{p}) - \boldsymbol{p} \cdot \boldsymbol{p}^T$ where $\boldsymbol{\mu} = \boldsymbol{p} = (p_1, \cdots, p_m)^T$.

Categorical, Mapped Binary and Centered Dataset

Modeling as multivariate Bernoulli variable is equivalent to treating $\boldsymbol{X}(x_i)$ as a new $n \times m$ binary data matrix

	X
x_1	Short
x_2	Short
x_3	Long
x_4	Short
x_5	Long

	A_1	A_2
$\boldsymbol{x}_1$	0	1
$\boldsymbol{x}_2$	0	1
$\boldsymbol{x}_3$	1	0
$\boldsymbol{x}_4$	0	1
$\boldsymbol{x}_5$	1	0

	Z_1	Z_2
$\boldsymbol{z}_1$	-0.4	0.4
$\boldsymbol{z}_2$	-0.4	0.4
$\boldsymbol{z}_3$	0.6	-0.6
$\boldsymbol{z}_4$	-0.4	0.4
$\boldsymbol{z}_5$	0.6	-0.6

X is the multivariate Bernoulli variable

$$\boldsymbol{X}(v) = \begin{cases} \boldsymbol{e}_1 = (1,0)^T & \text{if } v = \texttt{Long}(a_1) \\ \boldsymbol{e}_2 = (0,1)^T & \text{if } v = \texttt{Short}(a_2) \end{cases}$$

The sample mean and covariance matrix are

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{p}} = (2/5, 3/5)^T = (0.4, 0.6)^T \qquad \hat{\boldsymbol{\Sigma}} = diag(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^T = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

From the centered data, we have $\boldsymbol{Z} = (Z_1, Z_2)^T$ and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{5}\boldsymbol{Z}^T\boldsymbol{Z} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

Multinomial Distribution: Number of Occurrences

Let $\{x_1, x_2, \ldots, x_n\}$ be a random sample from $\boldsymbol{X}$. Let N_i be the random variable denoting number of occurrences of symbol a_i in the sample, and let $\boldsymbol{N} = (N_1, N_2, \ldots, N_m)^T$. $\boldsymbol{N}$ has a multinomial distribution, given as

$$f(\boldsymbol{N} = (n_1, n_2, \ldots, n_m) \mid \boldsymbol{p}) = \binom{n}{n_1 n_2 \ldots n_m} \prod_{i=1}^{m} p_i^{n_i}$$

The mean and covariance matrix of $\boldsymbol{N}$ are:

$$\boldsymbol{\mu}_N = E[\boldsymbol{N}] = nE[\boldsymbol{X}] = n \cdot \boldsymbol{\mu} = n \cdot \boldsymbol{p} = \begin{pmatrix} np_1 \\ \vdots \\ np_m \end{pmatrix}$$

$$\Sigma_N = n \cdot (diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_m \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_m & -np_2p_m & \cdots & np_m(1-p_m) \end{pmatrix}$$

The sample mean and covariance matrix for $\boldsymbol{N}$ are

$$\hat{\boldsymbol{\mu}}_N = n\hat{\boldsymbol{p}} \qquad \widehat{\Sigma}_N = n\left(diag(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^T\right)$$

Bivariate Analysis

Assume the data comprises two categorical attributes, X_1 and X_2,

$$dom(X_1) = \{a_{11}, a_{12}, \ldots, a_{1m_1}\}$$
$$dom(X_2) = \{a_{21}, a_{22}, \ldots, a_{2m_2}\}$$

We model X_1 and X_2 as multivariate Bernoulli variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with dimensions m_1 and m_2, respectively. The joint distribution of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is modeled as the $m_1 + m_2$ dimensional vector variable $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$

$$\boldsymbol{X}\left((v_1, v_2)^T\right) = \begin{pmatrix} \boldsymbol{X}_1(v_1) \\ \boldsymbol{X}_2(v_2) \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_{1i} \\ \boldsymbol{e}_{2j} \end{pmatrix}$$

provided that $v_1 = a_{1i}$ and $v_2 = a_{2j}$.
The joint PMF for $\boldsymbol{X}$ is given as the $m_1 \times m_2$ matrix

$$\boldsymbol{P}_{12} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m_2} \\ p_{21} & p_{22} & \cdots & p_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} & p_{m_1 2} & \cdots & p_{m_1 m_2} \end{pmatrix}$$

Bivariate Empirical PMF: sepal length and sepal width

X_1:sepal length

Bins	Domain	Counts
[4.3, 5.2]	Very Short (a_1)	$n_1 = 45$
(5.2, 6.1]	Short (a_2)	$n_2 = 50$
(6.1, 7.0]	Long (a_3)	$n_3 = 43$
(7.0, 7.9]	Very Long (a_4)	$n_4 = 12$

X_2:sepal width

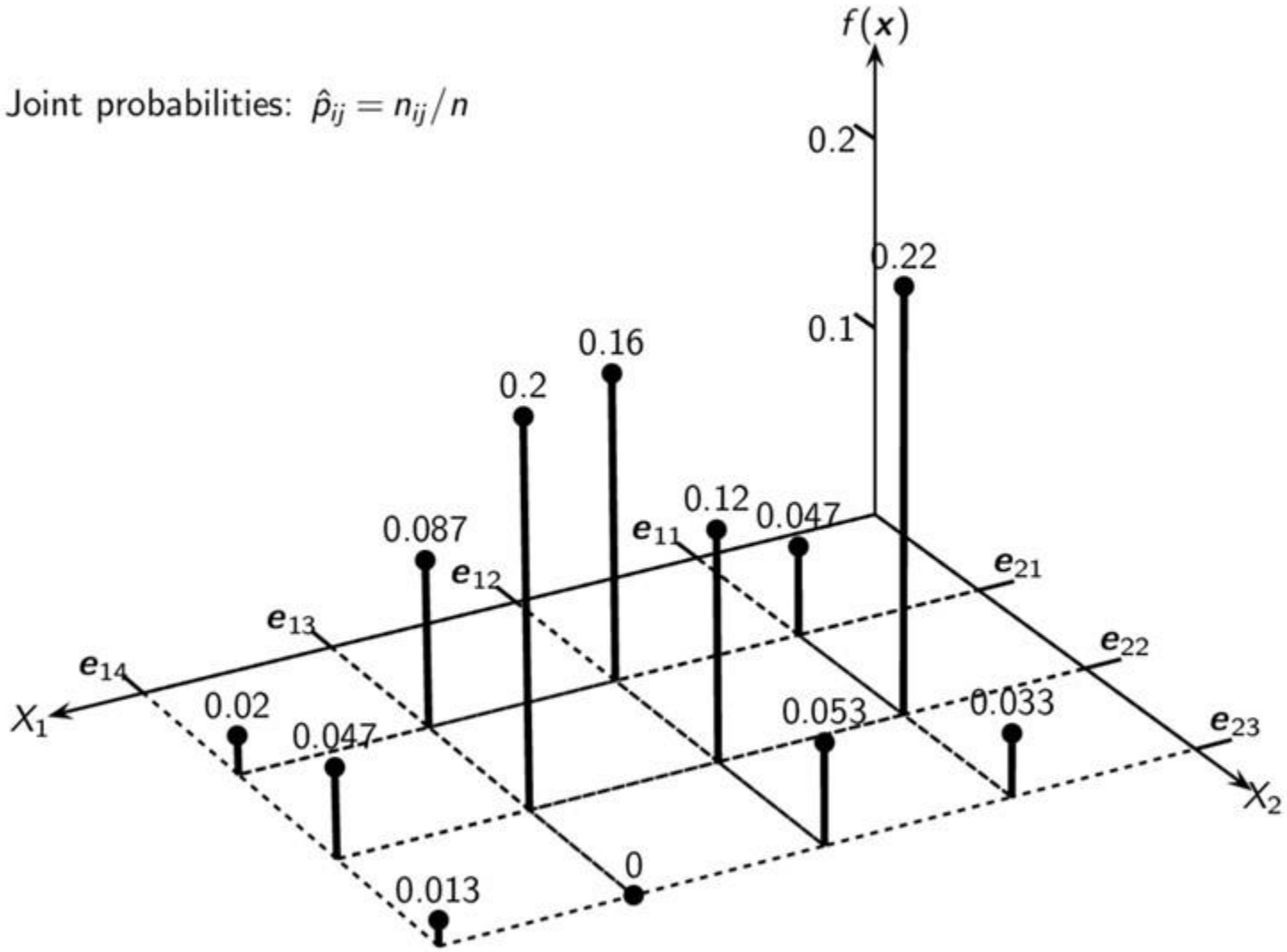
Bins	Domain	Counts
[2.0, 2.8]	Short (a_1)	47
(2.8, 3.6]	Medium (a_2)	88
(3.6, 4.4]	Long (a_3)	15

Observed Counts (n_{ij})

		X_2		
		Short ($\boldsymbol{e}_{21}$)	Medium ($\boldsymbol{e}_{22}$)	Long ($\boldsymbol{e}_{23}$)
X_1	Very Short ($\boldsymbol{e}_{11}$)	7	33	5
	Short ($\boldsymbol{e}_{22}$)	24	18	8
	Long ($\boldsymbol{e}_{13}$)	13	30	0
	Very Long ($\boldsymbol{e}_{14}$)	3	7	2

Bivariate Empirical PMF: sepal length and sepal width

Joint probabilities: $\hat{p}_{ij} = n_{ij}/n$



Attribute Dependence: Contingency Analysis

The *contingency table* for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is the $m_1 \times m_2$ matrix of observed counts n_{ij}

$$\boldsymbol{N}_{12} = n \cdot \widehat{\boldsymbol{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

where $\widehat{\boldsymbol{P}}_{12}$ is the empirical joint PMF for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The contingency table is augmented with row and column marginal counts, as follows:

$$\boldsymbol{N}_1 = n \cdot \hat{\boldsymbol{p}}_1 = \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \end{pmatrix} \qquad \boldsymbol{N}_2 = n \cdot \hat{\boldsymbol{p}}_2 = \begin{pmatrix} n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix}$$

$\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ have a multinomial distribution with parameters $\boldsymbol{p}_1 = (p_1^1, \ldots, p_{m_1}^1)$ and $\boldsymbol{p}_2 = (p_1^2, \ldots, p_{m_2}^2)$, respv.
$\boldsymbol{N}_{12}$ also has a multinomial distribution with parameters $\boldsymbol{P}_{12} = \{p_{ij}\}$, for $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$.

Contingency Table: sepal length vs. sepal width

Sepal length (X_1)	Sepal width (X_2)			
	Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
Very Short (a_{11})	7	33	5	$n_1^1 = 45$
Short (a_{12})	24	18	8	$n_2^1 = 50$
Long (a_{13})	13	30	0	$n_3^1 = 43$
Very Long (a_{14})	3	7	2	$n_4^1 = 12$
Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

Chi-Squared Test for Independence

Assume $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. Then, their joint PMF is

$$\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2$$

The expected frequency for each pair of values is

$$e_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}_i^1 \cdot \hat{p}_j^2 = n \cdot \frac{n_i^1}{n} \cdot \frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n}$$

The χ^2 statistic quantifies the difference between observed and expected counts

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

The sampling distribution for the χ^2 statistic follows the *chi-squared* density function:

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{\frac{q}{2}-1} \mathrm{e}^{-\frac{x}{2}}$$

where q is the degrees of freedom

$$\begin{aligned} q &= |dom(X_1)| \times |dom(X_2)| - (|dom(X_1)| + |dom(X_2)|) + 1 \\ &= m_1 m_2 - m_1 - m_2 + 1 \\ &= (m_1 - 1)(m_2 - 1) \end{aligned}$$

Chi-Squared Test: sepal length and sepal width

	Expected Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Short (a_{23})
X_1	Very Short (a_{11})	14.1	26.4	4.5
	Short (a_{12})	15.67	29.33	5.0
	Long (a_{13})	13.47	25.23	4.3
	Very Long (a_{14})	3.76	7.04	1.2

	Observed Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Long (a_{23})
	Very Short (a_{11})	7	33	5
	Short (a_{12})	24	18	8
	Long (a_{13})	13	30	0
	Very Long (a_{14})	3	7	2

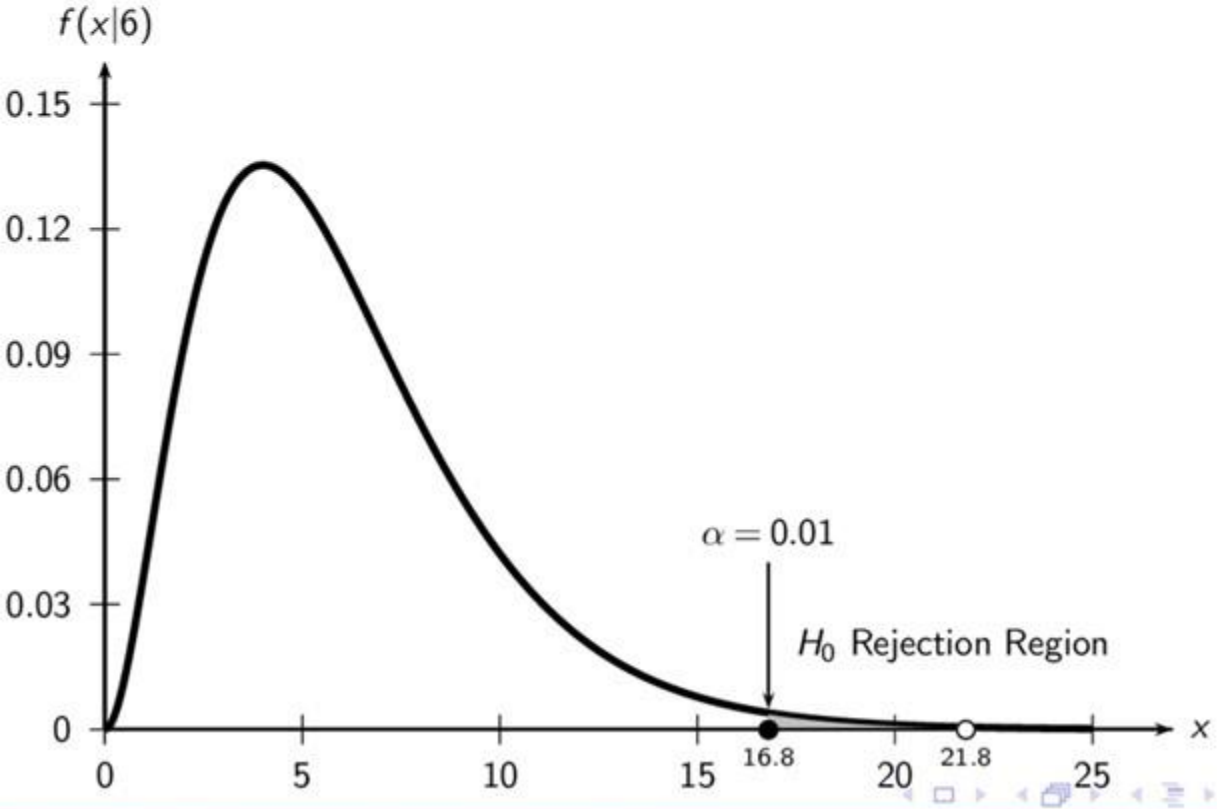
The chi-squared statistic value is $\chi^2 = 21.8$.

The number of degrees of freedom are

$$q = (m_1 - 1) \cdot (m_2 - 1) = 3 \cdot 2 = 6$$

Chi-Squared Distribution ($q = 6$).

The *p-value* of a statistic θ is defined as the probability of obtaining a value at least as extreme as the observed value.
The null hypothesis, that X_1 and X_2 are independent, is rejected if $p\text{-value}(z) \leq \alpha$, say $\alpha = 0.01$. We have $p\text{-value}(21.8) = 0.0013$. Thus, we reject the null hypothesis, and conclude that X_1 and X_2 are dependent.



Multiway Contingency Analysis

Given $\boldsymbol{X} = (X_1, X_2, \cdots, X_d)^T$. The chi-squared statistic is given as

$$\chi^2 = \sum_{\boldsymbol{i}} \frac{(n_{\boldsymbol{i}} - e_{\boldsymbol{i}})^2}{e_{\boldsymbol{i}}} = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \cdots \sum_{i_d=1}^{m_d} \frac{(n_{i_1,i_2,\ldots,i_d} - e_{i_1,i_2,\ldots,i_d})^2}{e_{i_1,i_2,\ldots,i_d}}$$

Under the null hypothesis, that attributes are independent, the expected number of occurrences of the symbol tuple $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$ is given as

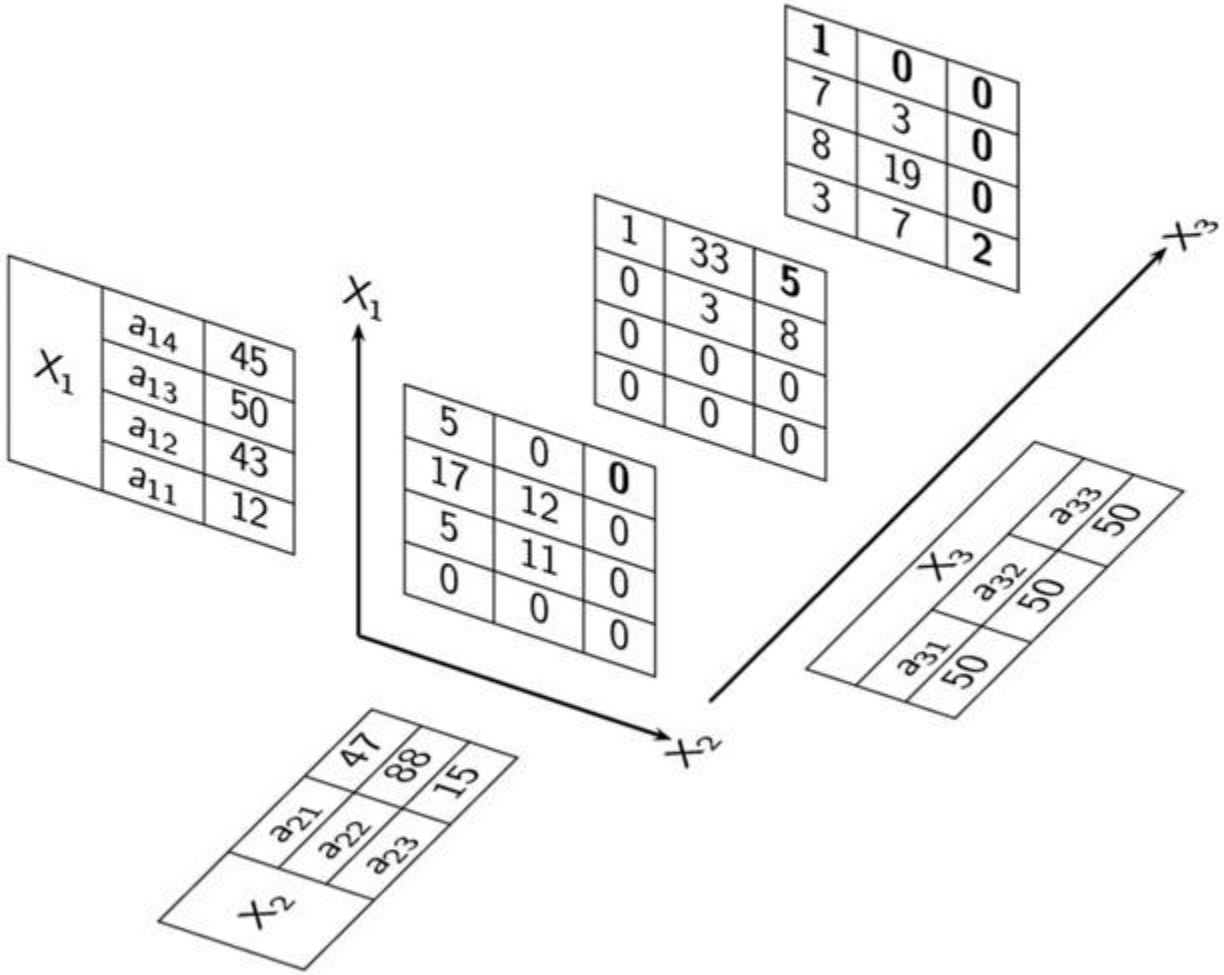
$$e_{\boldsymbol{i}} = n \cdot \hat{p}_{\boldsymbol{i}} = n \cdot \prod_{j=1}^{d} \hat{p}_{i_j}^{j} = \frac{n_{i_1}^1 n_{i_2}^2 \ldots n_{i_d}^d}{n^{d-1}}$$

The total number of degrees of freedom for the chi-squared distribution is given as

$$\begin{aligned} q &= \prod_{i=1}^{d} |dom(X_i)| - \sum_{i=1}^{d} |dom(X_i)| + (d-1) \\ &= \left(\prod_{i=1}^{d} m_i\right) - \left(\sum_{i=1}^{d} m_i\right) + d - 1 \end{aligned}$$

3-Way Contingency Table

X_1: sepal length, X_2: sepal width and X_3: Iris type



3-Way Contingency Analysis

		$X_3(a_{31}/a_{32}/a_{33})$		
		X_2		
		a_{21}	a_{22}	a_{23}
X_1	a_{11}	1.25	2.35	0.40
	a_{12}	4.49	8.41	1.43
	a_{13}	5.22	9.78	1.67
	a_{14}	4.70	8.80	1.50

The value of the χ^2 statistic is $\chi^2 = 231.06$, and the number of degrees of freedom is $q = 4 \cdot 3 \cdot 3 - (4+3+3) + 2 = 36 - 10 + 2 = 28$.

For a significance level of $\alpha = 0.01$, the critical value of the chi-square distribution is $z = 48.28$.

The observed value of $\chi^2 = 231.06$ is much greater than z, and it is thus extremely unlikely to happen under the null hypothesis. We conclude that the three attributes are not 3-way independent, but rather there is some dependence between them.

Distance and Angle

With the modeling of categorical attributes as multivariate Bernoulli variables, it is possible to compute the distance or the angle between any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{e}_{1i_1} \\ \vdots \\ \boldsymbol{e}_{d\,i_d} \end{pmatrix} \qquad \boldsymbol{x}_j = \begin{pmatrix} \boldsymbol{e}_{1j_1} \\ \vdots \\ \boldsymbol{e}_{d\,j_d} \end{pmatrix}$$

The different measures of distance and similarity rely on the number of matching and mismatching values (or symbols) across the d attributes $\boldsymbol{X}_k$.
The number of matching values s is given as:

$$s = \boldsymbol{x}_i^T \boldsymbol{x}_j = \sum_{k=1}^{d} (\boldsymbol{e}_{ki_k})^T \boldsymbol{e}_{kj_k}$$

The number of mismatches is simply $d - s$. Also useful is the norm of each point:

$$\|\boldsymbol{x}_i\|^2 = \boldsymbol{x}_i^T \boldsymbol{x}_i = d$$

Distance and Angle

The *Euclidean distance* between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is given as

$$\delta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| = \sqrt{\boldsymbol{x}_i^T \boldsymbol{x}_i - 2\boldsymbol{x}_i \boldsymbol{x}_j + \boldsymbol{x}_j^T \boldsymbol{x}_j} = \sqrt{2(d-s)}$$

The *Hamming distance* is given as

$$\delta_H(\boldsymbol{x}_i, \boldsymbol{x}_j) = d - s$$

Cosine Similarity: The cosine of the angle is given as

$$\cos\theta = \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \cdot \|\boldsymbol{x}_j\|} = \frac{s}{d}$$

The *Jaccard Coeff icient* is given as

$$J(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

Discretization

Discretization, also called *binning*, converts numeric attributes into categorical ones.

Equal-Width Intervals: Partition the range of X into k *equal-width* intervals. The interval width is simply the range of X divided by k:

$$w = \frac{x_{\max} - x_{\min}}{k}$$

Thus, the ith interval boundary is given as

$$v_i = x_{\min} + iw, \text{ for } i = 1, \ldots, k-1$$

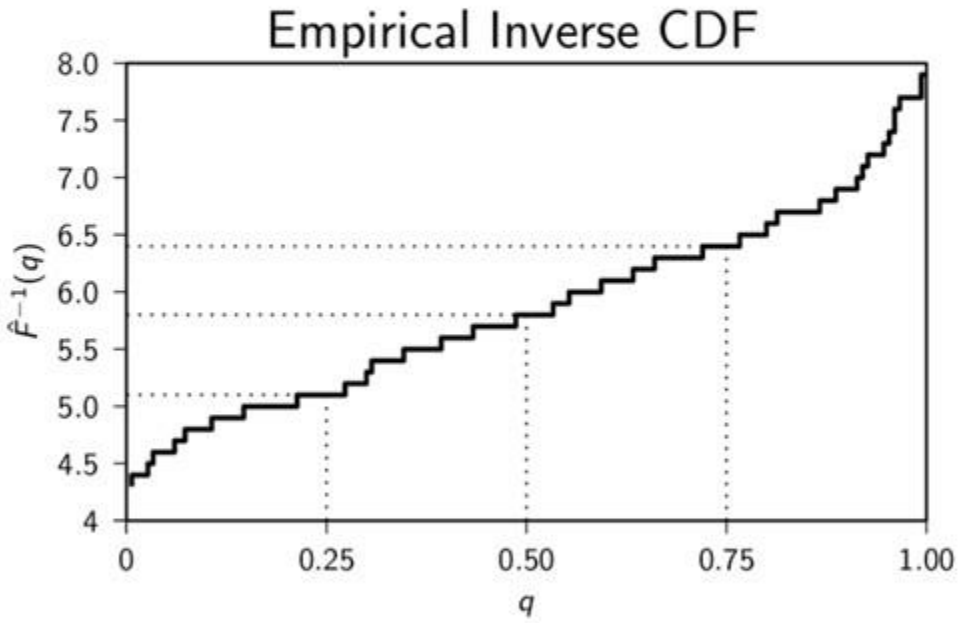
Equal-Frequency Intervals: We divide the range of X into intervals that contain (approximately) equal number of points. The intervals are computed from the empirical quantile or inverse cumulative distribution function

$$\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}$$

We require that each interval contain $1/k$ of the probability mass; therefore, the interval boundaries are given as follows:

$$v_i = \hat{F}^{-1}(i/k) \text{ for } i = 1, \ldots, k-1$$

Equal-Frequency Discretization: sepal length (4 bins)



Quartile values:

$$\hat{F}^{-}1(0.25) = 5.1$$
$$\hat{F}^{-}1(0.5) = 5.8$$
$$\hat{F}^{-}1(0.75) = 6.4$$

Range: $[4.3, 7.9]$

Bin	Width	Count
$[4.3, 5.1]$	0.8	$n_1 = 41$
$(5.1, 5.8]$	0.7	$n_2 = 39$
$(5.8, 6.4]$	0.6	$n_3 = 35$
$(6.4, 7.9]$	1.5	$n_4 = 35$

Data mining and Machine learning

Part 1. Data mining and Analysis

Graphs

A *graph* $G = (V, E)$ comprises a finite nonempty set V of *vertices* or *nodes*, and a set $E \subseteq V \times V$ of *edges* consisting of *unordered* pairs of vertices.

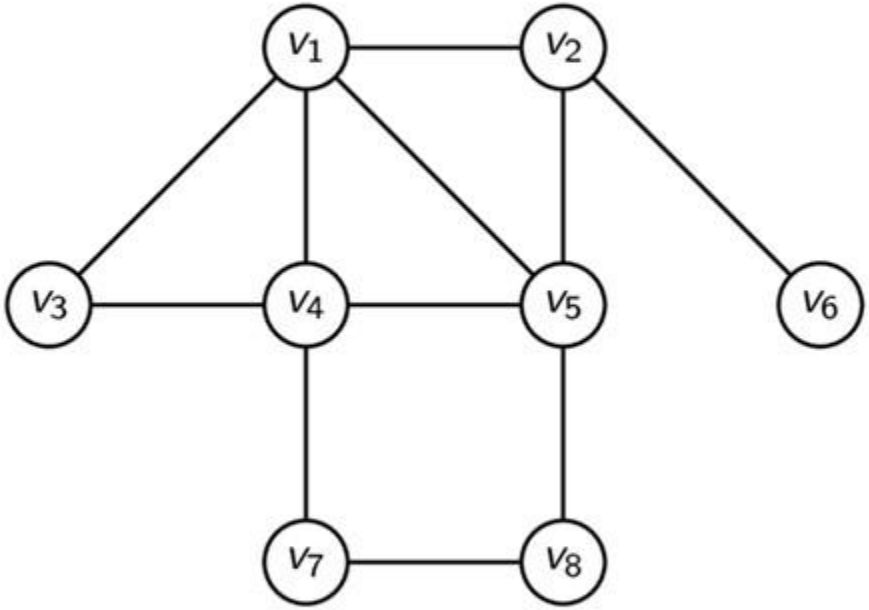
The number of nodes in the graph G, given as $|V| = n$, is called the *order* of the graph, and the number of edges in the graph, given as $|E| = m$, is called the *size* of G.

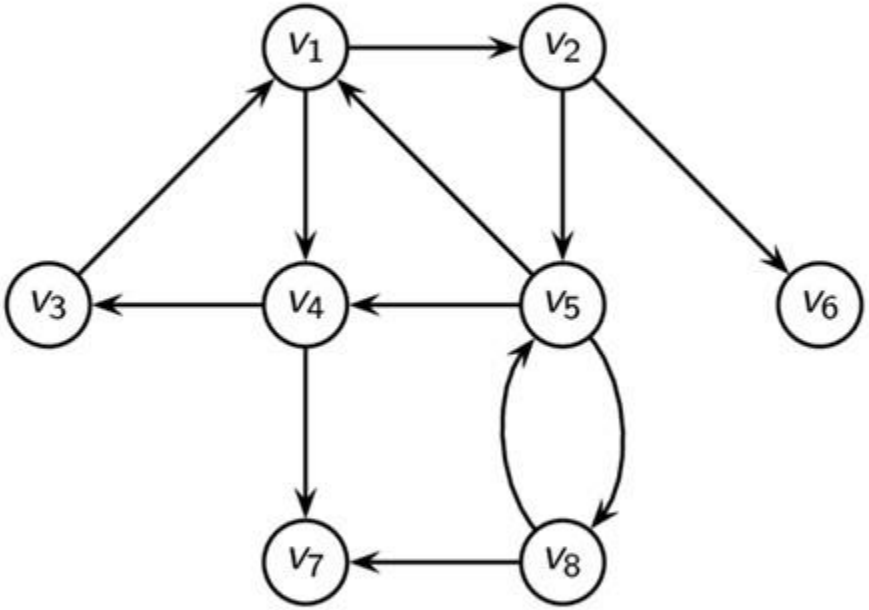
A *directed graph* or *digraph* has an edge set E consisting of *ordered* pairs of vertices.

A *weighted graph* consists of a graph together with a weight w_{ij} for each edge $(v_i, v_j) \in E$.

A graph $H = (V_H, E_H)$ is called a *subgraph* of $G = (V, E)$ if $V_H \subseteq V$ and $E_H \subseteq E$.

Undirected and Directed Graphs





Degree Distribution

The *degree* of a node $v_i \in V$ is the number of edges incident with it, and is denoted as $d(v_i)$ or just d_i.

The *degree sequence* of a graph is the list of the degrees of the nodes sorted in non-increasing order.

Let N_k denote the number of vertices with degree k. The *degree frequency distribution* of a graph is given as

$$(N_0, N_1, \ldots, N_t)$$

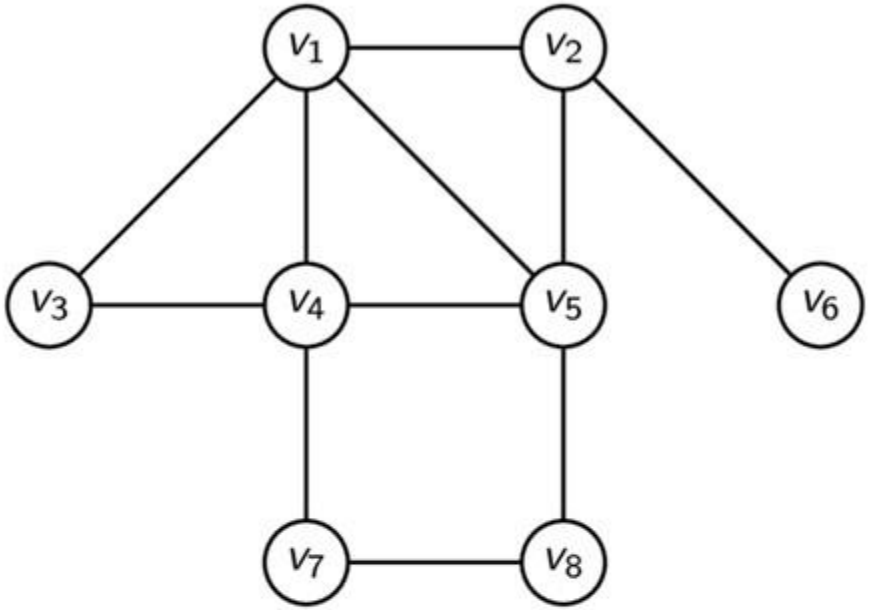
where t is the maximum degree for a node in G.

Let X be a random variable denoting the degree of a node. The *degree distribution* of a graph gives the probability mass function f for X, given as

$$\big(f(0), f(1), \ldots, f(t)\big)$$

where $f(k) = P(X = k) = \frac{N_k}{n}$ is the probability of a node with degree k.

Degree Distribution



The degree sequence of the graph is

$$(4,4,4,3,2,2,2,1)$$

Its degree frequency distribution is

$$(N_0, N_1, N_2, N_3, N_4) = (0,1,3,1,3)$$

The degree distribution is given as

$$\big(f(0), f(1), f(2), f(3), f(4)\big) = (0, 0.125, 0.375, 0.125, 0.375)$$

Path, Distance and Connectedness

A *walk* in a graph G between nodes x and y is an ordered sequence of vertices, starting at x and ending at y,

$$x = v_0, v_1, \ldots, v_{t-1}, v_t = y$$

such that there is an edge between every pair of consecutive vertices, that is, $(v_{i-1}, v_i) \in E$ for all $i = 1, 2, \ldots, t$. The length of the walk, t, is measured in terms of *hops* – the number of edges along the walk.

A *path* is a walk with *distinct* vertices (with the exception of the start and end vertices). A path of minimum length between nodes x and y is called a *shortest path*, and the length of the shortest path is called the *distance* between x and y, denoted as $d(x, y)$. If no path exists between the two nodes, the distance is assumed to be $d(x, y) = \infty$.

Two nodes v_i and v_j are *connected* if there exists a path between them. A graph is *connected* if there is a path between all pairs of vertices. A *connected component*, or just *component*, of a graph is a maximal connected subgraph.

A directed graph is *strongly connected* if there is a (directed) path between all ordered pairs of vertices. It is *weakly connected* if there exists a path between node pairs only by considering edges as undirected.

Adjacency Matrix

A graph $G = (V, E)$, with $|V| = n$ vertices, can be represented as an $n \times n$, symmetric binary *adjacency matrix*, $\boldsymbol{A}$, defined as

$$\boldsymbol{A}(i,j) = \begin{cases} 1 & \text{if } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

If the graph is directed, then the adjacency matrix $\boldsymbol{A}$ is not symmetric.

If the graph is weighted, then we obtain an $n \times n$ *weighted adjacency matrix*, $\boldsymbol{A}$, defined as

$$\boldsymbol{A}(i,j) = \begin{cases} w_{ij} & \text{if } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

where w_{ij} is the weight on edge $(v_i, v_j) \in E$.

Graphs from Data Matrix

Many datasets that are not in the form of a graph can still be converted into one.

Let $\boldsymbol{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ (with $\boldsymbol{x}_i \in \mathbb{R}^d$), be a dataset. Define a weighted graph $G = (V, E)$, with edge weight

$$w_{ij} = sim(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

where $sim(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denotes the similarity between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

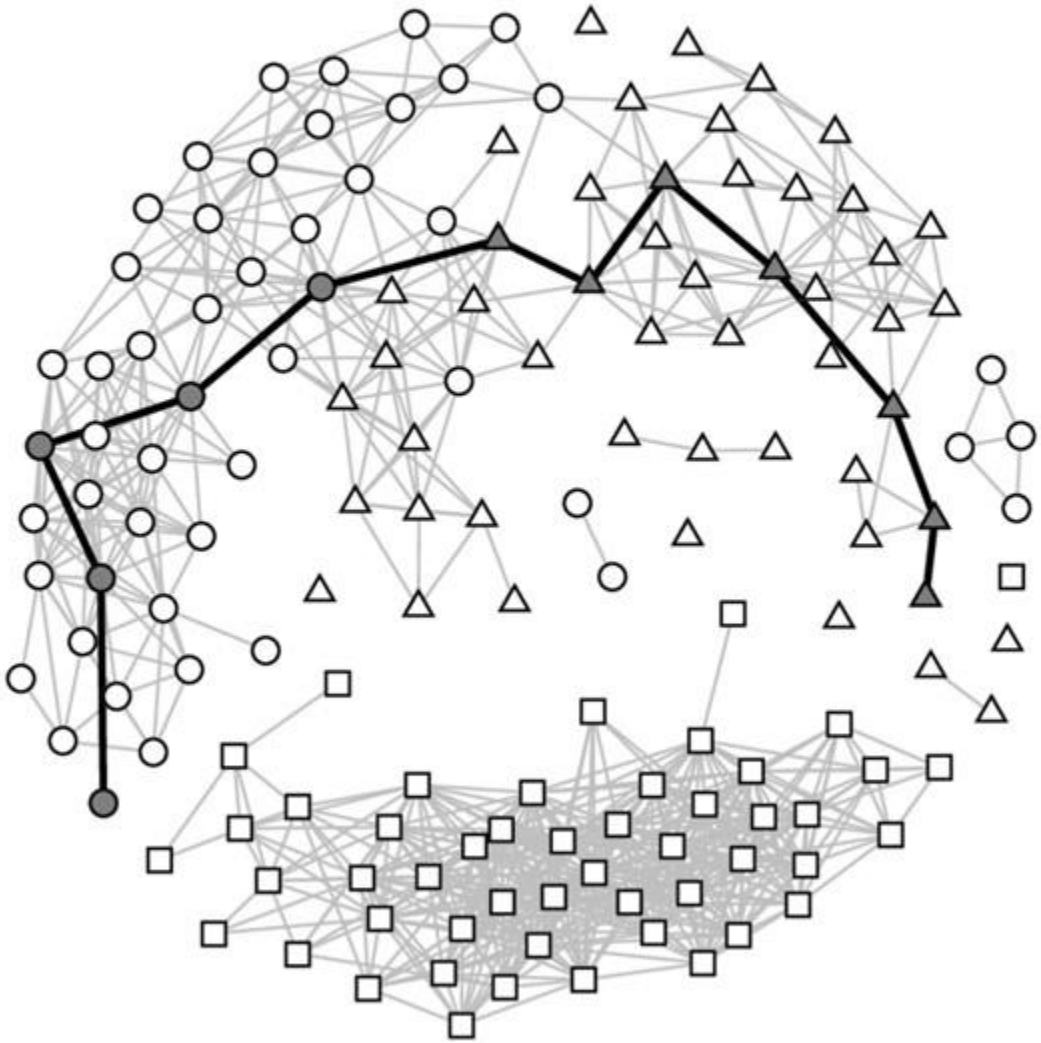
For instance, using the Gaussian similarity

$$w_{ij} = sim(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right\}$$

where σ is the spread parameter.

Iris Similarity Graph: Gaussian Similarity

$\sigma = 1/\sqrt{2}$; edge exists iff $w_{ij} \geq 0.777$
order: $|V| = n = 150$; size: $|E| = m = 753$



Topological Graph Attributes

Graph attributes are *local* if they apply to only a single node (or an edge), and *global* if they refer to the entire graph.

Degree: The degree of a node $v_i \in G$ is defined as

$$d_i = \sum_j \boldsymbol{A}(i,j)$$

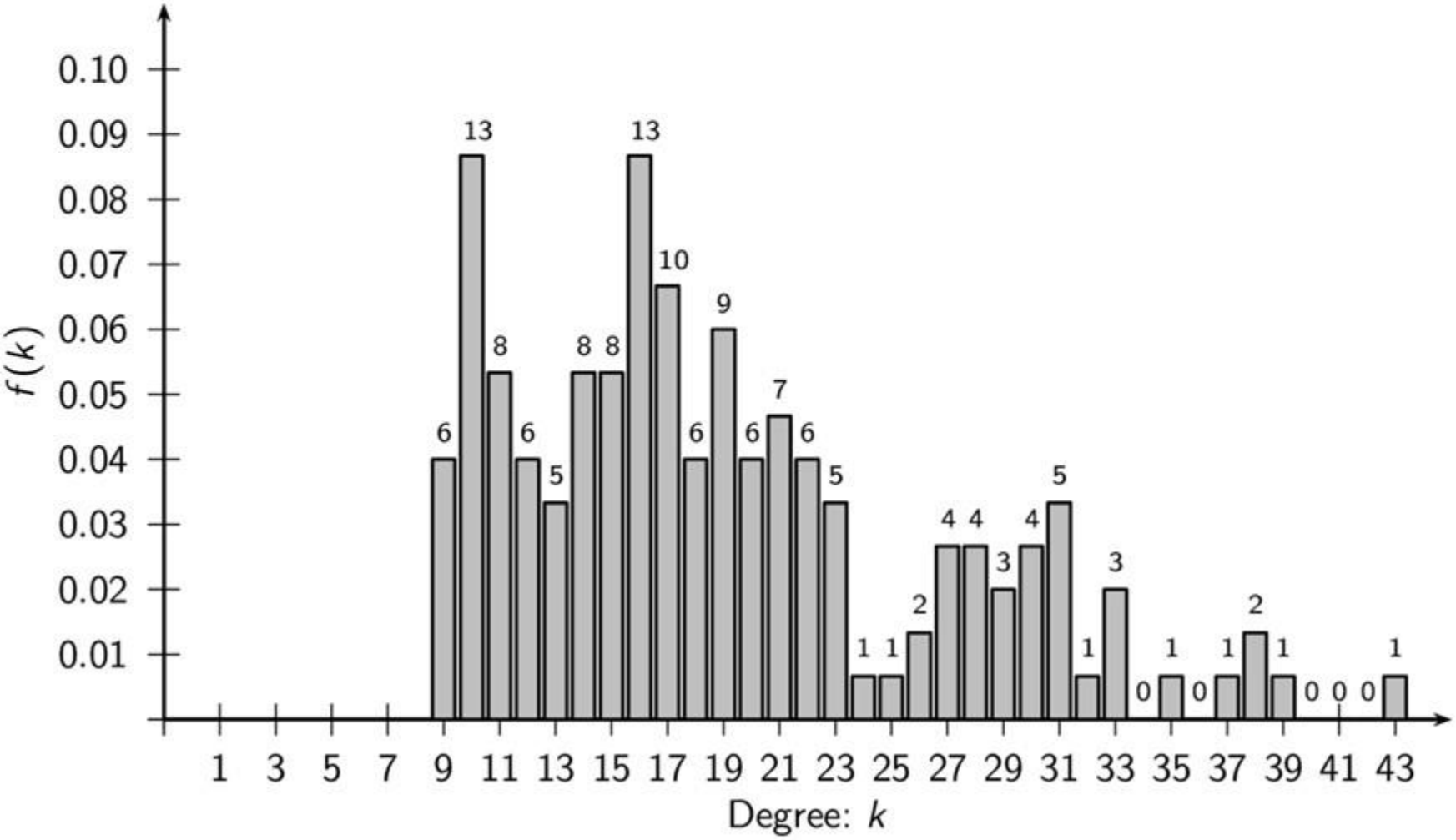
The corresponding global attribute for the entire graph G is the *average degree*:

$$\mu_d = \frac{\sum_i d_i}{n}$$

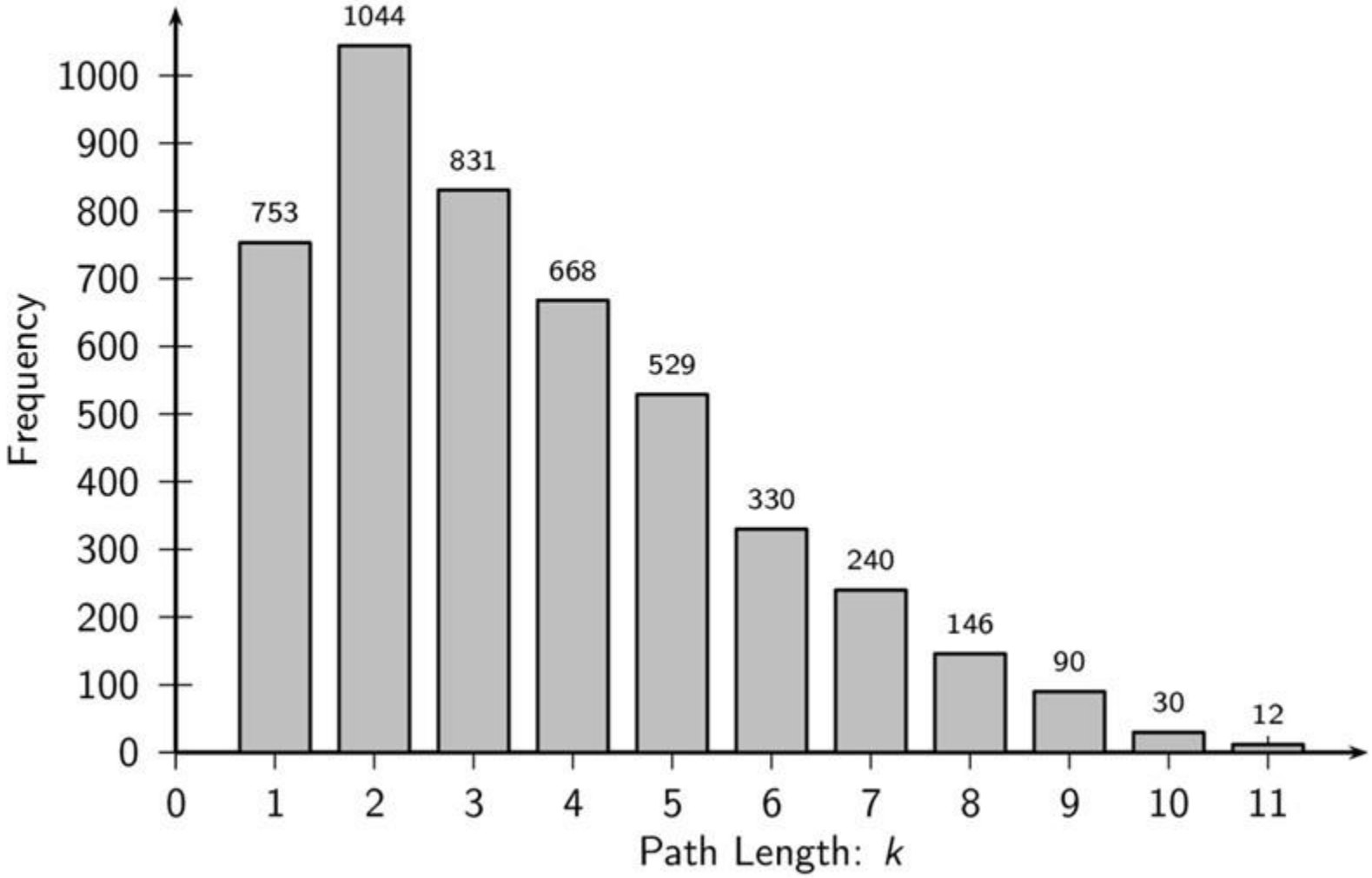
Average Path Length: The *average path length* is given as

$$\mu_L = \frac{\sum_i \sum_{j>i} d(v_i, v_j)}{\binom{n}{2}} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} d(v_i, v_j)$$

Iris Graph: Degree Distribution



Iris Graph: Path Length Histogram



Eccentricity, Radius and Diameter

The *eccentricity* of a node v_i is the maximum distance from v_i to any other node in the graph:

$$e(v_i) = \max_j \{ d(v_i, v_j) \}$$

The *radius* of a connected graph, denoted $r(G)$, is the minimum eccentricity of any node in the graph:

$$r(G) = \min_i \{ e(v_i) \} = \min_i \left\{ \max_j \{ d(v_i, v_j) \} \right\}$$

The *diameter*, denoted $d(G)$, is the maximum eccentricity of any vertex in the graph:

$$d(G) = \max_i \{ e(v_i) \} = \max_{i,j} \{ d(v_i, v_j) \}$$

For a disconnected graph, values are computed over the connected components of the graph.

The diameter of a graph G is sensitive to outliers. *Effective diameter* is more robust; defined as the minimum number of hops for which a large fraction, typically 90%, of all connected pairs of nodes can reach each other.

Clustering Coefficient

The *clustering coefficient* of a node v_i is a measure of the density of edges in the neighborhood of v_i.

Let $G_i = (V_i, E_i)$ be the subgraph induced by the neighbors of vertex v_i. Note that $v_i \notin V_i$, as we assume that G is simple.

Let $|V_i| = n_i$ be the number of neighbors of v_i, and $|E_i| = m_i$ be the number of edges among the neighbors of v_i. The clustering coefficient of v_i is defined as

$$C(v_i) = \frac{\text{no. of edges in } G_i}{\text{maximum number of edges in } G_i} = \frac{m_i}{\binom{n_i}{2}} = \frac{2 \cdot m_i}{n_i(n_i - 1)}$$

The *clustering coefficient* of a graph G is simply the average clustering coefficient over all the nodes, given as

$$C(G) = \frac{1}{n} \sum_i C(v_i)$$

$C(v_i)$ is well defined only for nodes with degree $d(v_i) \geq 2$, thus define $C(v_i) = 0$ if $d_i < 2$.

Transitivity and Efficiency

Transitivity of the graph is defined as

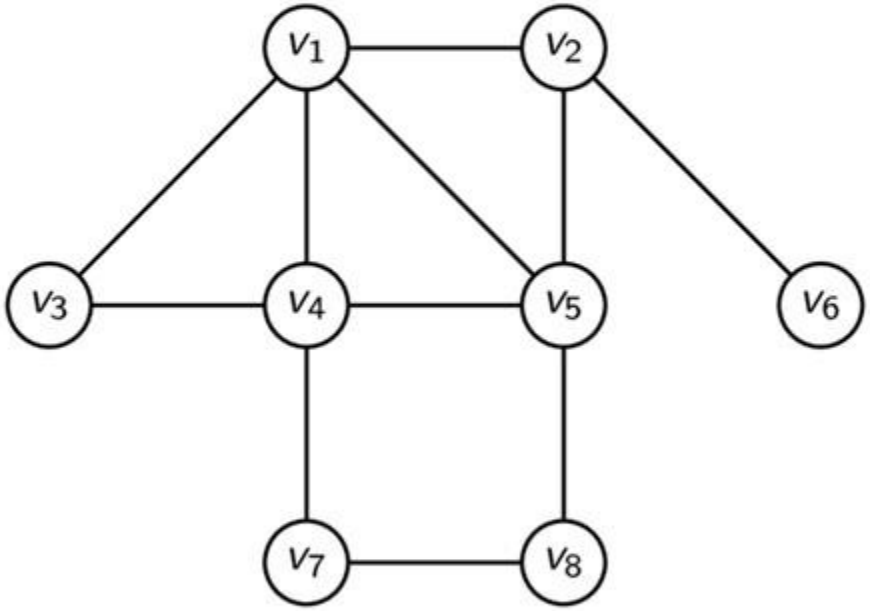
$$T(G) = \frac{3 \times \text{no. of triangles in } G}{\text{no. of connected triples in } G}$$

where the subgraph composed of the edges (v_i, v_j) and (v_i, v_k) is a *connected triple* centered at v_i, and a connected triple centered at v_i that includes (v_j, v_k) is called a *triangle* (a complete subgraph of size 3).

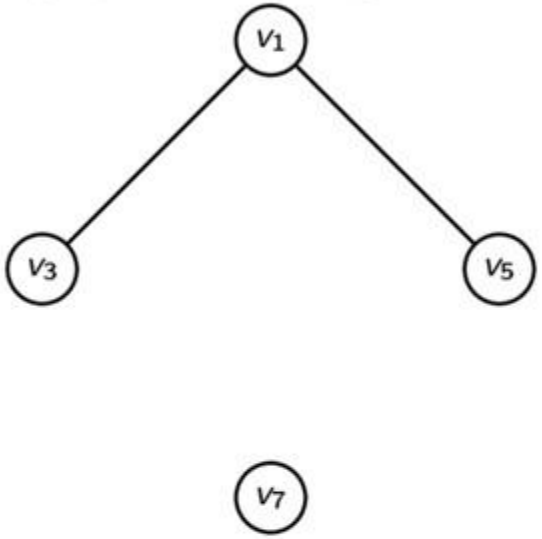
The *efficiency* for a pair of nodes v_i and v_j is defined as $\frac{1}{d(v_i, v_j)}$. If v_i and v_j are not connected, then $d(v_i, v_j) = \infty$ and the efficiency is $1/\infty = 0$.
The *efficiency* of a graph G is the average efficiency over all pairs of nodes, whether connected or not, given as

$$\frac{2}{n(n-1)} \sum_i \sum_{j>i} \frac{1}{d(v_i, v_j)}$$

Clustering Coefficient



Subgraph induced by node v_4:



The clustering coefficient of v_4 is

$$C(v_4) = \frac{2}{\binom{4}{2}} = \frac{2}{6} = 0.33$$

The clustering coefficient for G is

$$C(G) = \frac{1}{8}\left(\frac{1}{2} + \frac{1}{3} + 1 + \frac{1}{3} + \frac{1}{3}\right) = \frac{2.5}{8} = 0.3125$$

Centrality Analysis

A centrality is a function $c: V \to \mathbb{R}$, that induces a ranking on V.

Degree Centrality: The simplest notion of centrality is the degree d_i of a vertex v_i – the higher the degree, the more important or central the vertex.

Eccentricity Centrality: Eccentricity centrality is defined as:

$$c(v_i) = \frac{1}{e(v_i)} = \frac{1}{\max_j \{d(v_i, v_j)\}}$$

The less eccentric a node is, the more central it is.

Closeness Centrality: closeness centrality uses the sum of all the distances to rank how central a node is

$$c(v_i) = \frac{1}{\sum_j d(v_i, v_j)}$$

Betweenness Centrality

The betweenness centrality measures how many shortest paths between all pairs of vertices include v_i. It gives an indication as to the central "monitoring" role played by v_i for various pairs of nodes.

Let η_{jk} denote the number of shortest paths between vertices v_j and v_k, and let $\eta_{jk}(v_i)$ denote the number of such paths that include or contain v_i.

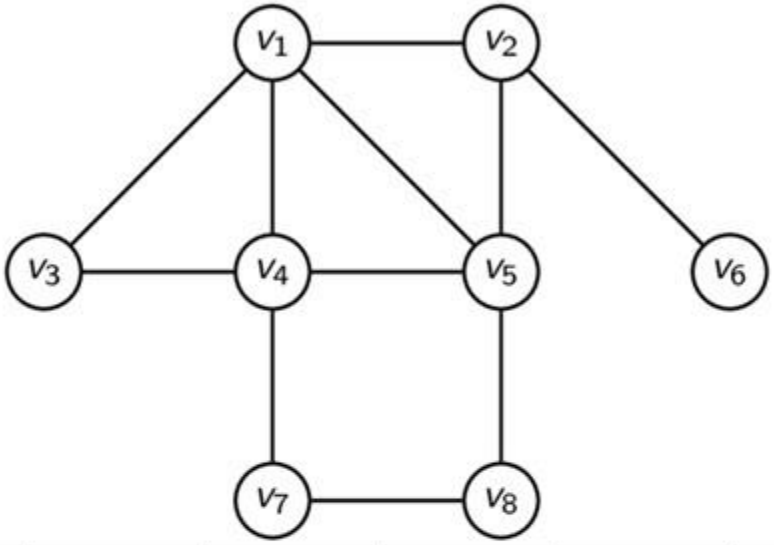
The fraction of paths through v_i is denoted as

$$\gamma_{jk}(v_i) = \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

The betweenness centrality for a node v_i is defined as

$$c(v_i) = \sum_{j \neq i} \sum_{\substack{k \neq i \\ k > j}} \gamma_{jk} = \sum_{j \neq i} \sum_{\substack{k \neq i \\ k > j}} \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

Centrality Values



Centrality	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Degree	4	3	2	4	4	1	2	2
Eccentricity	0.5	0.33	0.33	0.33	0.5	0.25	0.25	0.33
$e(v_i)$	2	3	3	3	2	4	4	3
Closeness	0.100	0.083	0.071	0.091	0.100	0.056	0.067	0.071
$\sum_j d(v_i, v_j)$	10	12	14	11	10	18	15	14
Betweenness	4.5	6	0	5	6.5	0	0.83	1.17

Prestige or Eigenvector Centrality

Let $p(u)$ be a positive real number, called the *prestige* score for node u. Intuitively the more (prestigious) the links that point to a given node, the higher its prestige.

$$p(v) = \sum_u \boldsymbol{A}(u,v) \cdot p(u)$$
$$= \sum_u \boldsymbol{A}^T(v,u) \cdot p(u)$$

Across all the nodes, we have

$$\boldsymbol{p}' = \boldsymbol{A}^T \boldsymbol{p}$$

where $\boldsymbol{p}$ is an n-dimensional prestige vector.
By recursive expansion, we see that

$$\boldsymbol{p}_k = \boldsymbol{A}^T \boldsymbol{p}_{k-1} = \left(\boldsymbol{A}^T\right)^2 \boldsymbol{p}_{k-2} = \cdots = \left(\boldsymbol{A}^T\right)^k \boldsymbol{p}_0$$

where $\boldsymbol{p}_0$ is the initial prestige vector. It is well known that the vector $\boldsymbol{p}_k$ converges to the dominant eigenvector of $\boldsymbol{A}^T$.

Computing Dominant Eigenvector: Power Iteration

The dominant eigenvector of $\boldsymbol{A}^T$ and the corresponding eigenvalue can be computed using the *power iteration* method.

It starts with an initial vector $\boldsymbol{p}_0$, and in each iteration, it multiplies on the left by $\boldsymbol{A}^T$, and scales the intermediate $\boldsymbol{p}_k$ vector by dividing it by the maximum entry $\boldsymbol{p}_k[i]$ in $\boldsymbol{p}_k$ to prevent numeric overflow.

The ratio of the maximum entry in iteration k to that in $k-1$, given as $\lambda = \frac{\boldsymbol{p}_k[i]}{\boldsymbol{p}_{k-1}[i]}$, yields an estimate for the eigenvalue.

The iterations continue until the difference between successive eigenvector estimates falls below some threshold $\epsilon > 0$.

PowerIteration $(\boldsymbol{A}, \epsilon)$**:**

$k \leftarrow 0$ // iteration

$\boldsymbol{p}_0 \leftarrow 1 \in \mathbb{R}^n$ // initial vector

repeat

  $k \leftarrow k+1$ $\boldsymbol{p}_k \leftarrow \boldsymbol{A}^T \boldsymbol{p}_{k-1}$ // eigenvector estimate

  $i \leftarrow \arg\max_j \{\boldsymbol{p}_k[j]\}$ // maximum value index

  $\lambda \leftarrow \boldsymbol{p}_k[i] / \boldsymbol{p}_{k-1}[i]$ // eigenvalue estimate

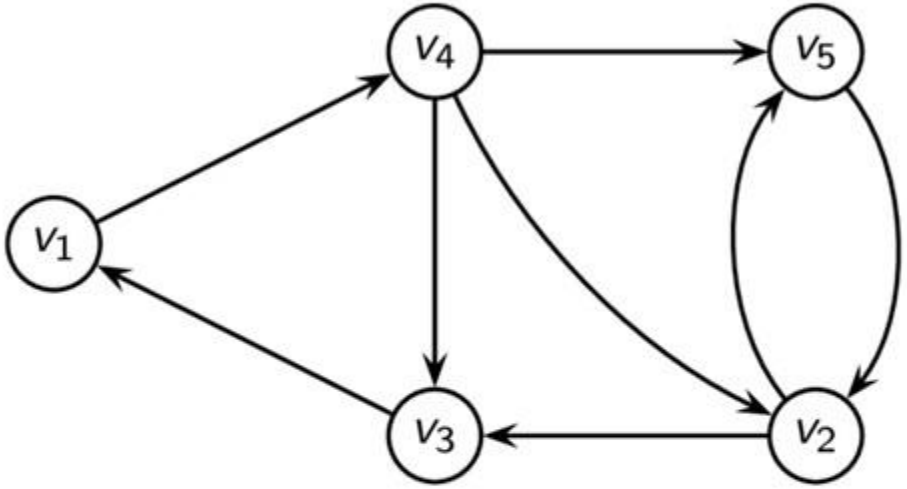
  $\boldsymbol{p}_k \leftarrow \frac{1}{\boldsymbol{p}_k[i]} \boldsymbol{p}_k$ // scale vector

until $\left\| \boldsymbol{p}_k - \boldsymbol{p}_{k-1} \right\| \leq \epsilon$

$\boldsymbol{p} \leftarrow \frac{1}{\|\boldsymbol{p}_k\|} \boldsymbol{p}_k$ // normalize eigenvector

return $\boldsymbol{p}, \lambda$

Power Iteration for Eigenvector Centrality: Example



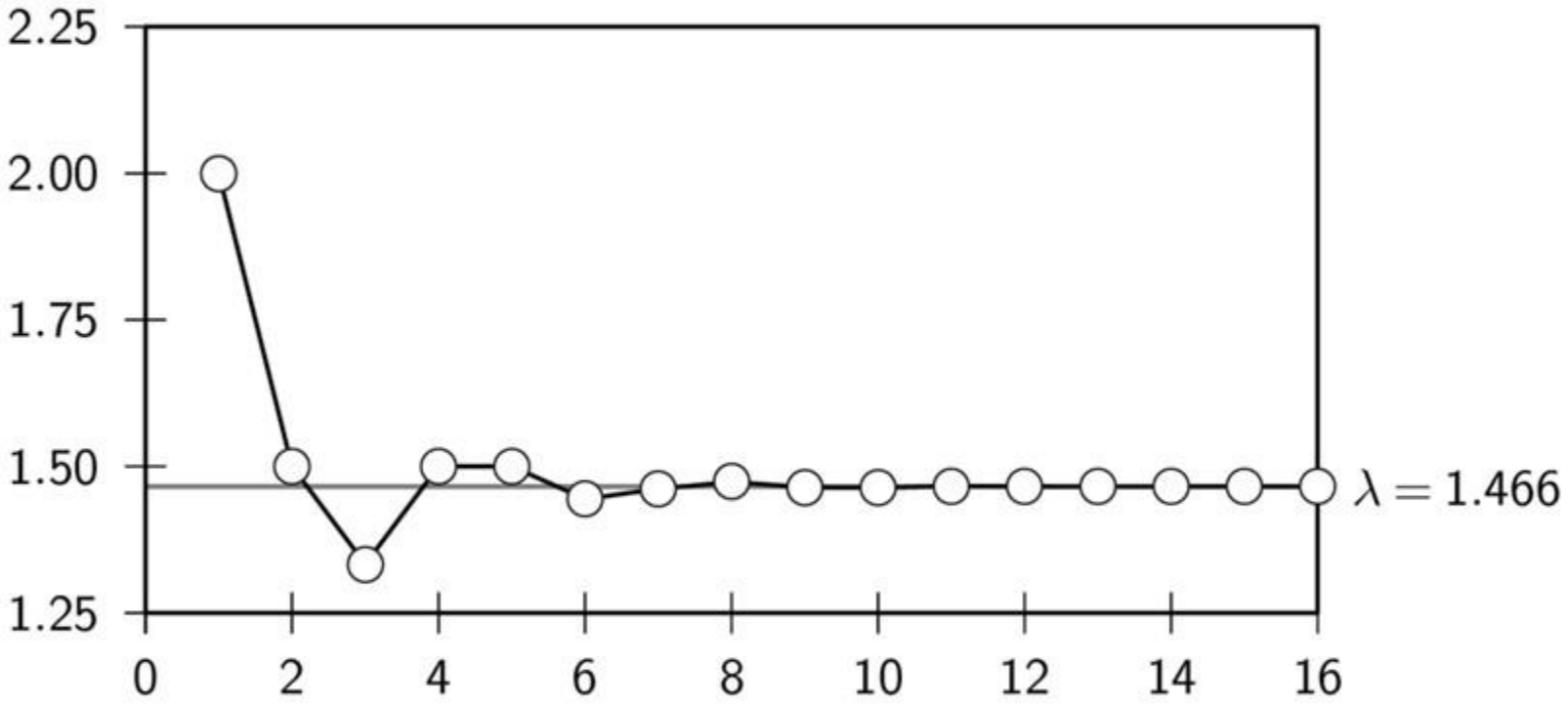
$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad \boldsymbol{A}^T = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Power Method via Scaling

$\boldsymbol{p}_0$	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	$\boldsymbol{p}_3$
$\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}$	$\begin{pmatrix}1\\2\\2\\1\\2\end{pmatrix} \to \begin{pmatrix}0.5\\1\\1\\0.5\\1\end{pmatrix}$	$\begin{pmatrix}1\\1.5\\1.5\\0.5\\1.5\end{pmatrix} \to \begin{pmatrix}0.67\\1\\1\\0.33\\1\end{pmatrix}$	$\begin{pmatrix}1\\1.33\\1.33\\0.67\\1.33\end{pmatrix} \to \begin{pmatrix}0.75\\1\\1\\0.5\\1\end{pmatrix}$
λ	2	1.5	1.33

$\boldsymbol{p}_4$	$\boldsymbol{p}_5$	$\boldsymbol{p}_6$	$\boldsymbol{p}_7$
$\begin{pmatrix}1\\1.5\\1.5\\0.75\\1.5\end{pmatrix} \to \begin{pmatrix}0.67\\1\\1\\0.5\\1\end{pmatrix}$	$\begin{pmatrix}1\\1.5\\1.5\\0.67\\1.5\end{pmatrix} \to \begin{pmatrix}0.67\\1\\1\\0.44\\1\end{pmatrix}$	$\begin{pmatrix}1\\1.44\\1.44\\0.67\\1.44\end{pmatrix} \to \begin{pmatrix}0.69\\1\\1\\0.46\\1\end{pmatrix}$	$\begin{pmatrix}1\\1.46\\1.46\\0.69\\1.46\end{pmatrix} \to \begin{pmatrix}0.68\\1\\1\\0.47\\1\end{pmatrix}$
1.5	1.5	1.444	1.462

Convergence of the Ratio to Dominant Eigenvalue



PageRank

PageRank is based on (normalized) prestige combined with a *random jump* assumption. The PageRank of a node v recursively depends on the PageRank of other nodes that point to it.

Normalized Prestige: Define $\boldsymbol{N}$ as the *normalized adjacency matrix*

$$\boldsymbol{N}(u,v) = \begin{cases} \frac{1}{od(u)} & \text{if } (u,v) \in E \\ 0 & \text{if } (u,v) \notin E \end{cases}$$

where $od(u)$ is the out-degree of node u.
Normalized prestige is given as

$$p(v) = \sum_u \boldsymbol{N}^T(v,u) \cdot p(u)$$

Random Jumps: In the random surfing approach, there is a small probability of jumping from one node to any of the other nodes in the graph. The normalized adjacency matrix for a fully connected graph is

$$\boldsymbol{N}_r = \frac{1}{n} 1_{n \times n}$$

where $1_{n \times n}$ is the $n \times n$ matrix of all ones.

PageRank: Normalized Prestige and Random Jumps

The PageRank vector is recursively defined as

$$\begin{aligned} \boldsymbol{p}' &= (1-\alpha)\boldsymbol{N}^T\boldsymbol{p} + \alpha\boldsymbol{N}_r^T\boldsymbol{p} \\ &= \left((1-\alpha)\boldsymbol{N}^T + \alpha\boldsymbol{N}_r^T\right)\boldsymbol{p} \\ &= \boldsymbol{M}^T\boldsymbol{p} \end{aligned}$$

α denotes the probability of random jumps. The solution is the dominant eigenvector of $\boldsymbol{M}^T$, where $\boldsymbol{M} = (1-\alpha)\boldsymbol{N} + \alpha\boldsymbol{N}_r$ is the combined normalized adjacency matrix.

Sink Nodes: If $od(u) = 0$, then only random jumps from u are allowed. The modified $\boldsymbol{M}$ matrix is given as

$$\boldsymbol{M}_u = \begin{cases} \boldsymbol{M}_u & \text{if } od(u) > 0 \\ \frac{1}{n}1_n^T & \text{if } od(u) = 0 \end{cases}$$

where 1_n is the n-dimensional vector of all ones.

Hub and Authority Scores (HITS)

The *authority score* of a page is analogous to PageRank or prestige, and it depends on how many "good" pages point to it. The *hub score* of a page is based on how many "good" pages it points to. In other words, a page with high authority has many hub pages pointing to it, and a page with high hub score points to many pages that have high authority.

Let $a(u)$ be the authority score and $h(u)$ the hub score of node u. We have:

$$a(v) = \sum_u \boldsymbol{A}^T(v,u) \cdot h(u)$$

$$h(v) = \sum_u \boldsymbol{A}(v,u) \cdot a(u)$$

In matrix notation, we obtain

$$\boldsymbol{a}' = \boldsymbol{A}^T \boldsymbol{h} \qquad\qquad \boldsymbol{h}' = \boldsymbol{A}\boldsymbol{a}$$

Recursively, we have:

$$\boldsymbol{a}_k = \boldsymbol{A}^T \boldsymbol{h}_{k-1} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{a}_{k-1}) = (\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{a}_{k-1}$$

$$\boldsymbol{h}_k = \boldsymbol{A}\boldsymbol{a}_{k-1} = \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{h}_{k-1}) = (\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{h}_{k-1}$$

The authority score converges to the dominant eigenvector of $\boldsymbol{A}^T\boldsymbol{A}$, whereas the hub score converges to the dominant eigenvector of $\boldsymbol{A}\boldsymbol{A}^T$.

Small World Property

Real-world graphs exhibit the *small-world* property that there is a short path between any pair of nodes. A graph G exhibits small-world behavior if the average path length μ_L scales logarithmically with the number of nodes in the graph, that is, if

$$\mu_L \propto \log n$$

where n is the number of nodes in the graph.

A graph is said to have *ultra-small-world* property if the average path length is much smaller than $\log n$, that is, if $\mu_L \ll \log n$.

Scale-free Property

In many real-world graphs it has been observed that the empirical degree distribution $f(k)$ exhibits a *scale-free* behavior captured by a power-law relationship with k, that is, the probability that a node has degree k satisfies the condition

$$f(k) \propto k^{-\gamma}$$

Taking the logarithm on both sides gives

$$\log f(k) = \log(\alpha k^{-\gamma})$$
$$\text{or } \log f(k) = -\gamma \log k + \log \alpha$$

which is the equation of a straight line in the log-log plot of k versus $f(k)$, with $-\gamma$ giving the slope of the line.

A power-law relationship leads to a scale-free or scale invariant behavior because scaling the argument by some constant c does not change the proportionality.

Clustering Effect

Real-world graphs often also exhibit a *clustering effect*, that is, two nodes are more likely to be connected if they share a common neighbor. The clustering effect is captured by a high clustering coefficient for the graph G.

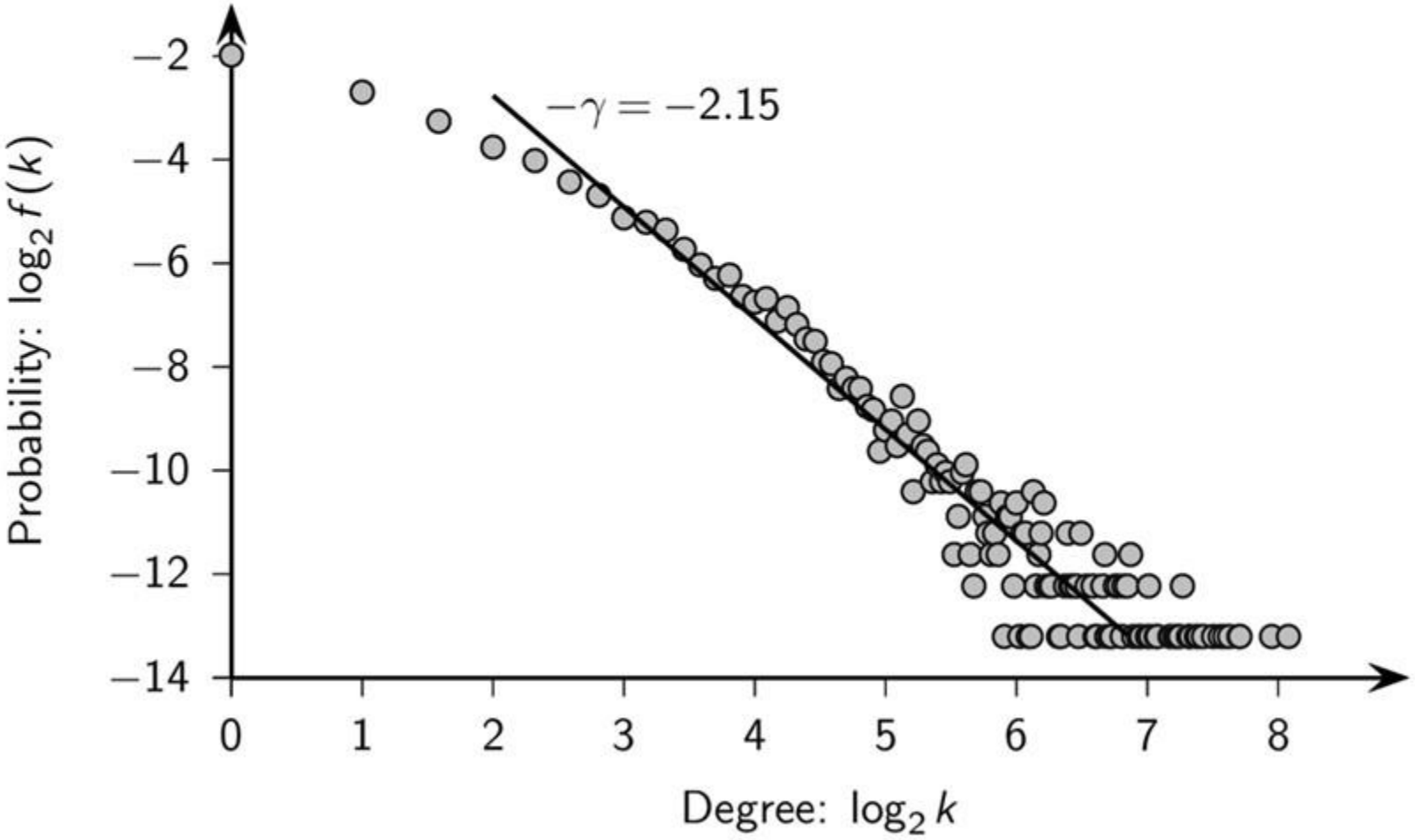
Let $C(k)$ denote the average clustering coefficient for all nodes with degree k; then the clustering effect also manifests itself as a power-law relationship between $C(k)$ and k:

$$C(k) \propto k^{-\gamma}$$

In other words, a log-log plot of k versus $C(k)$ exhibits a straight line behavior with negative slope $-\gamma$.

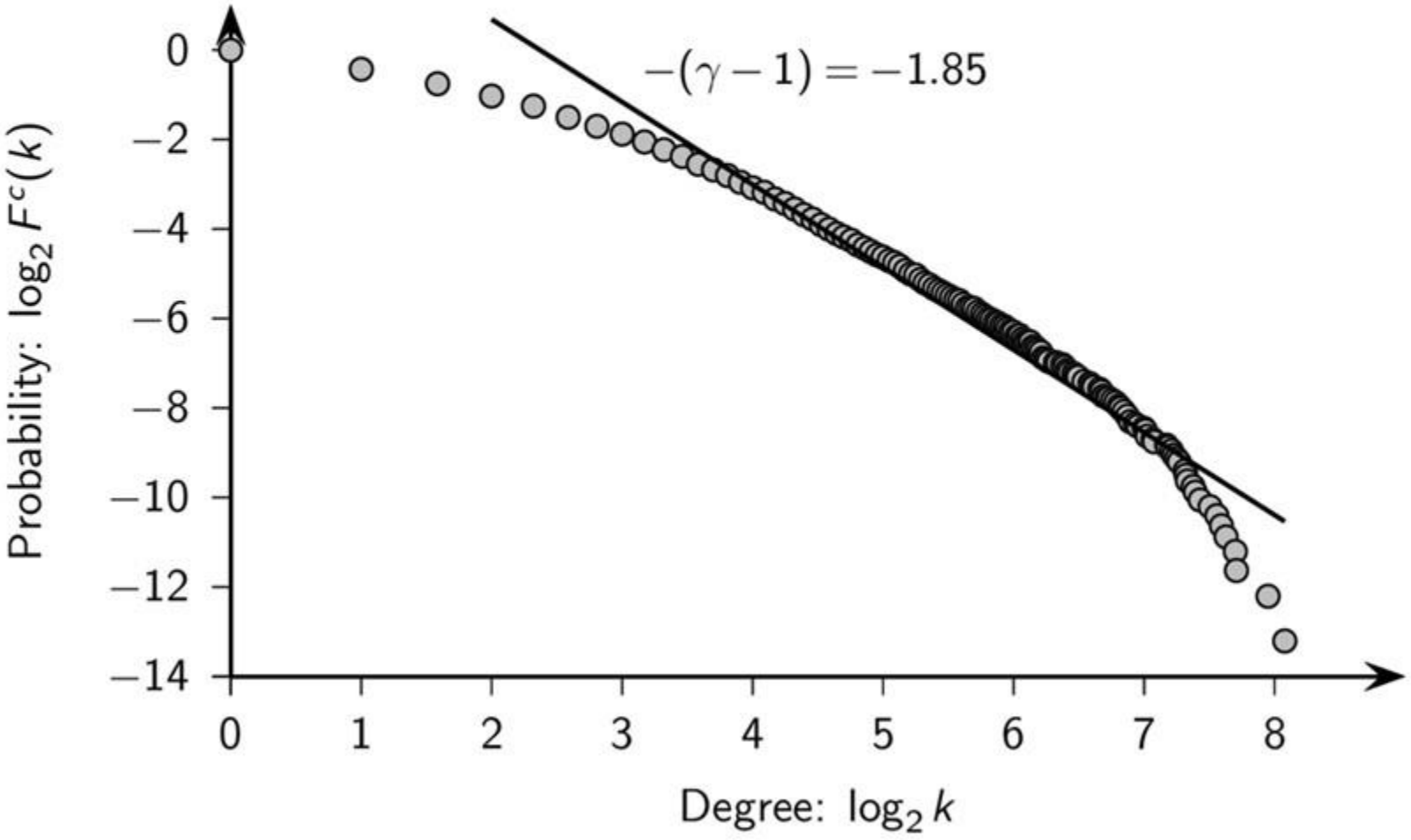
Degree Distribution: Human Protein Interaction Network

$|V| = n = 9521$, $|E| = m = 37060$

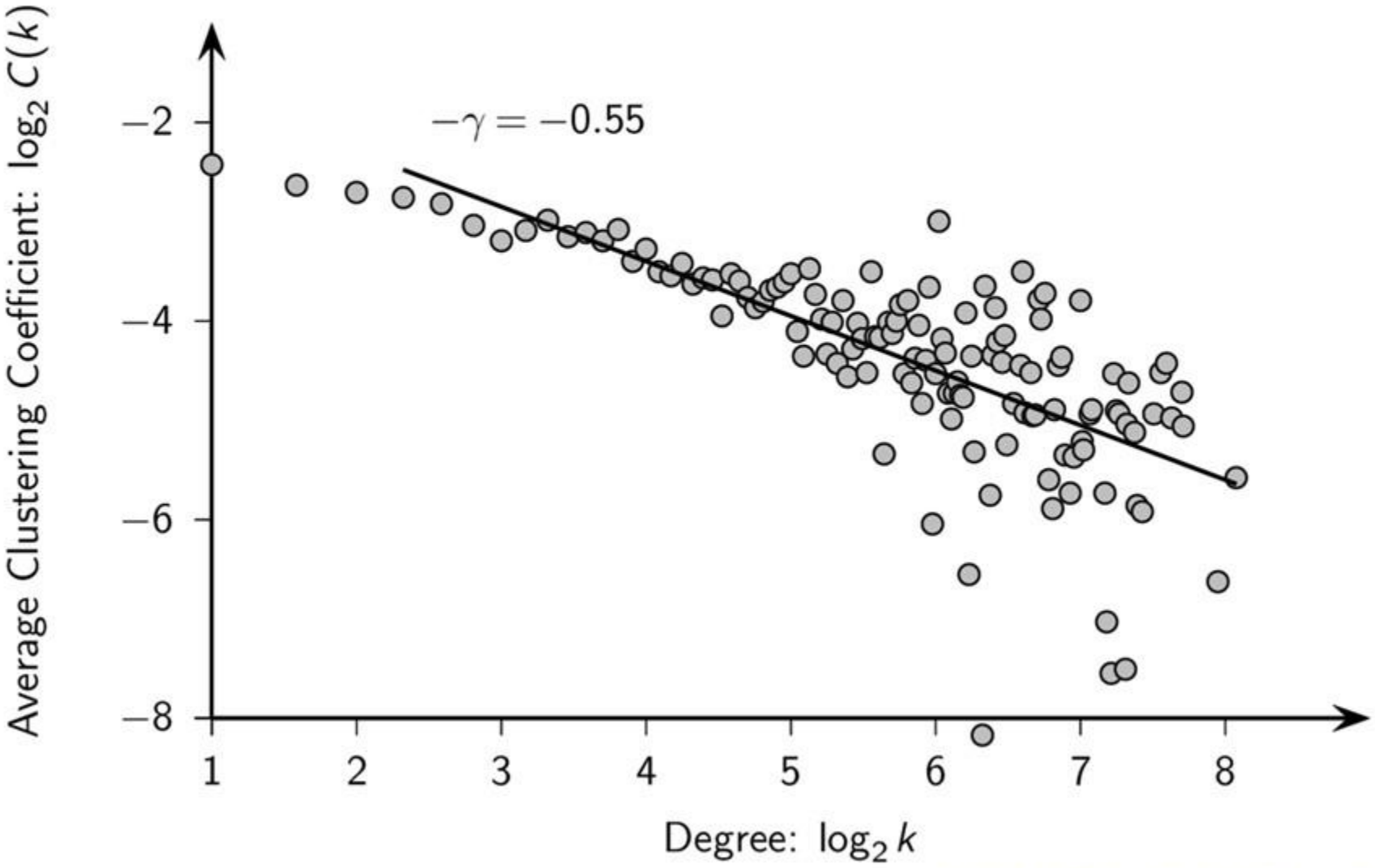


Cumulative Degree Distribution

$F^c(k) = 1 - F(k)$ where $F(k)$ is the CDF for $f(k)$



Average Clustering Coefficient



Erdös–Rényi Random Graph Model

The ER model specifies a collection of graphs $\mathcal{G}(n, m)$ with n nodes and m edges, such that each graph $G \in \mathcal{G}$ has equal probability of being selected:

$$P(G) = \frac{1}{\binom{M}{m}} = \binom{M}{m}^{-1}$$

where $M = \binom{n}{2} = \frac{n(n-1)}{2}$ and $\binom{M}{m}$ is the number of possible graphs with m edges (with n nodes).

Random Graph Generation: Randomly select two distinct vertices $v_i, v_j \in V$, and add an edge (v_i, v_j) to E, provided the edge is not already in the graph G. Repeat the process until exactly m edges have been added to the graph.

Let X be a random variable denoting the degree of a node for $G \in \mathcal{G}$. Let p denote the probability of an edge in G

$$p = \frac{m}{M} = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}$$

Random Graphs: Average Degree

Degree of a node follows a binomial distribution with probability of success p, given as

$$f(k) = P(X = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

since a node can be connected to $n-1$ other vertices.

The average degree μ_d is then given as the expected value of X:

$$\mu_d = E[X] = (n-1)p$$

The variance of the degree is

$$\sigma_d^2 = var(X) = (n-1)p(1-p)$$

Random Graphs: Degree Distribution

As $n \to \infty$ and $p \to 0$ the expected value and variance of X can be rewritten as

$$E[X] = (n-1)p \simeq np \text{ as } n \to \infty$$
$$var(X) = (n-1)p(1-p) \simeq np \text{ as } n \to \infty \text{ and } p \to 0$$

The binomial distribution can be approximated by a Poisson distribution with parameter λ, given as

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda = np$ represents both the expected value and variance of the distribution.

Thus, ER random graphs do not exhibit power law degree distribution.

Random Graphs: Clustering Coefficient and Diameter

Clustering Coefficient: Consider a node v_i with degree k. Since p is the probability of an edge, the expected number of edges m_i among the neighbors of a node v_i is simply

$$m_i = \frac{pk(k-1)}{2}$$

The clustering coefficient is

$$C(v_i) = \frac{2m_i}{k(k-1)} = p$$

which implies that $C(G) = \frac{1}{n}\sum_i C(v_i) = p$. Since for sparse graphs we have $p \to 0$, this means that ER random graphs do not show clustering effect.

Diameter: Expected degree of a node is $\mu_d = \lambda$, so in one hop a node can reach λ nodes. Coarsely, in k hops it can reach λ^k nodes. Thus, we have

$$\sum_{k=1}^{t} \lambda^k \leq n, \text{ which implies that } t = \log_\lambda n$$

It follows that the diameter of the graph is

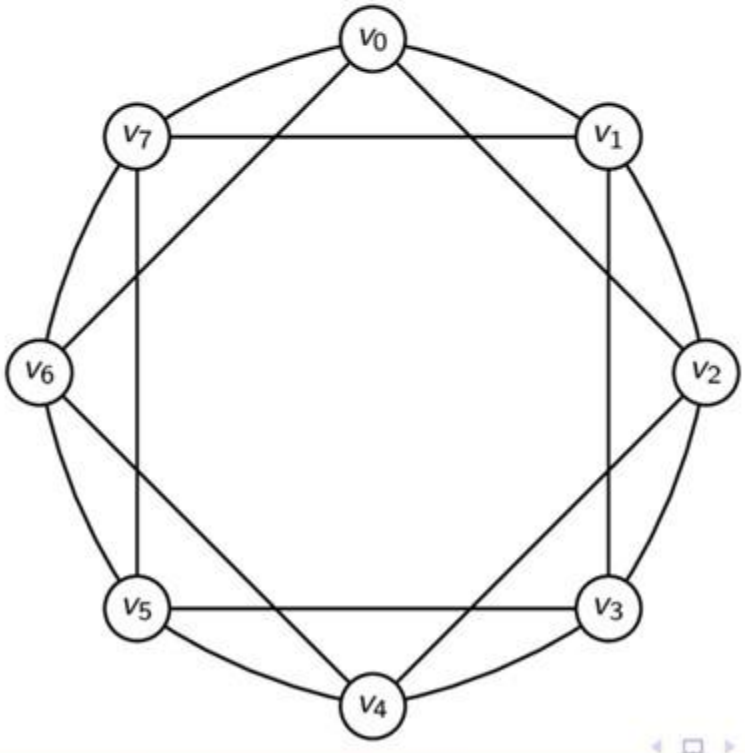
$$d(G) \propto \log n$$

Thus, ER random graphs are small-world.

Watts–Strogatz Small-world Graph Model

The Watts–Strogatz (WS) model starts with a regular graph of degree $2k$, having n nodes arranged in a circular layout, with each node having edges to its k neighbors on the right and left.
The regular graph has high local clustering. Adding a small amount of randomness leads to the emergence of the small-world phenomena.

Watts–Strogatz Regular Graph: $n = 8,\ k = 2$



WS Regular Graph: Clustering Coefficient and Diameter

The clustering coefficient of a node v is given as

$$C(v) = \frac{m_v}{M_v} = \frac{3k-3}{4k-2}$$

As k increases, the clustering coefficient approaches $\frac{3}{4}$ because $C(G) = C(v) \to \frac{3}{4}$ as $k \to \infty$. The WS regular graph thus has a high clustering coefficient.

The diameter of a regular WS graph is given as

$$d(G) = \begin{cases} \left\lceil \frac{n}{2k} \right\rceil & \text{if } n \text{ is even} \\ \left\lceil \frac{n-1}{2k} \right\rceil & \text{if } n \text{ is odd} \end{cases}$$

The regular graph has a diameter that scales linearly in the number of nodes, and thus it is not small-world.

Random Perturbation of Regular Graph

Edge Rewiring: For each edge (u, v) in the graph, with probability r, replace v with another randomly chosen node avoiding loops and duplicate edges.

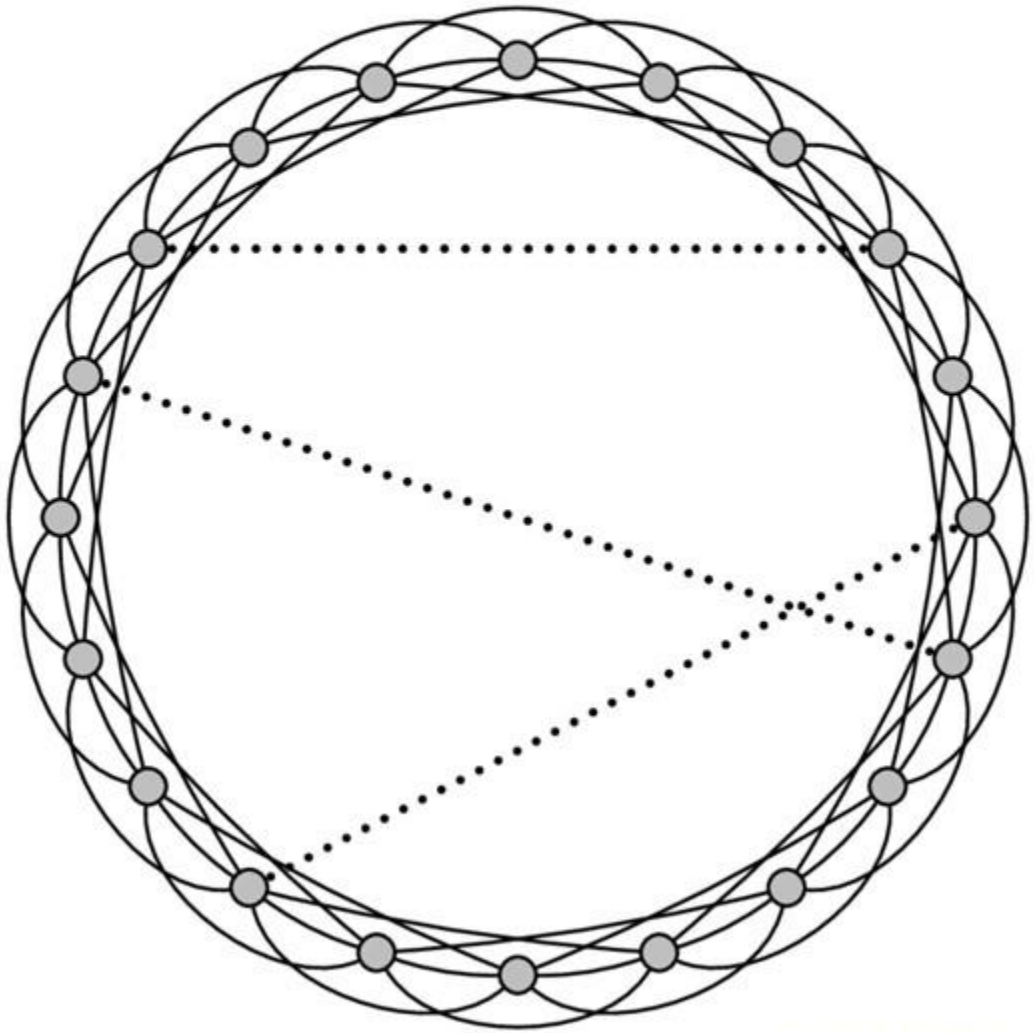
The WS regular graph has $m = kn$ total edges, so after rewiring, rm of the edges are random, and $(1 - r)m$ are regular.

Edge Shortcuts: Add a few *shortcut* edges between random pairs of nodes, with r being the probability, per edge, of adding a shortcut edge.

The total number of random shortcut edges added to the network is $mr = knr$. The total number of edges in the graph is $m + mr = (1 + r)m = (1 + r)kn$.

Watts–Strogatz Graph: Shortcut Edges

$n = 20,\ k = 3$



Properties of Watts–Strogatz Graphs

Degree Distribution: Let X denote the random variable denoting the number of shortcuts for each node. Then the probability of a node with j shortcut edges is given as

$$f(j) = P(X = j) = \binom{n'}{j} p^j (1-p)^{n'-j}$$

with $E[X] = n'p = 2kr$ and $p = \frac{2kr}{n-2k-1} = \frac{2kr}{n'}$.

The expected degree of each node in the network is therefore $2k + E[X] = 2k(1+r)$. The degree distribution is not a power law.

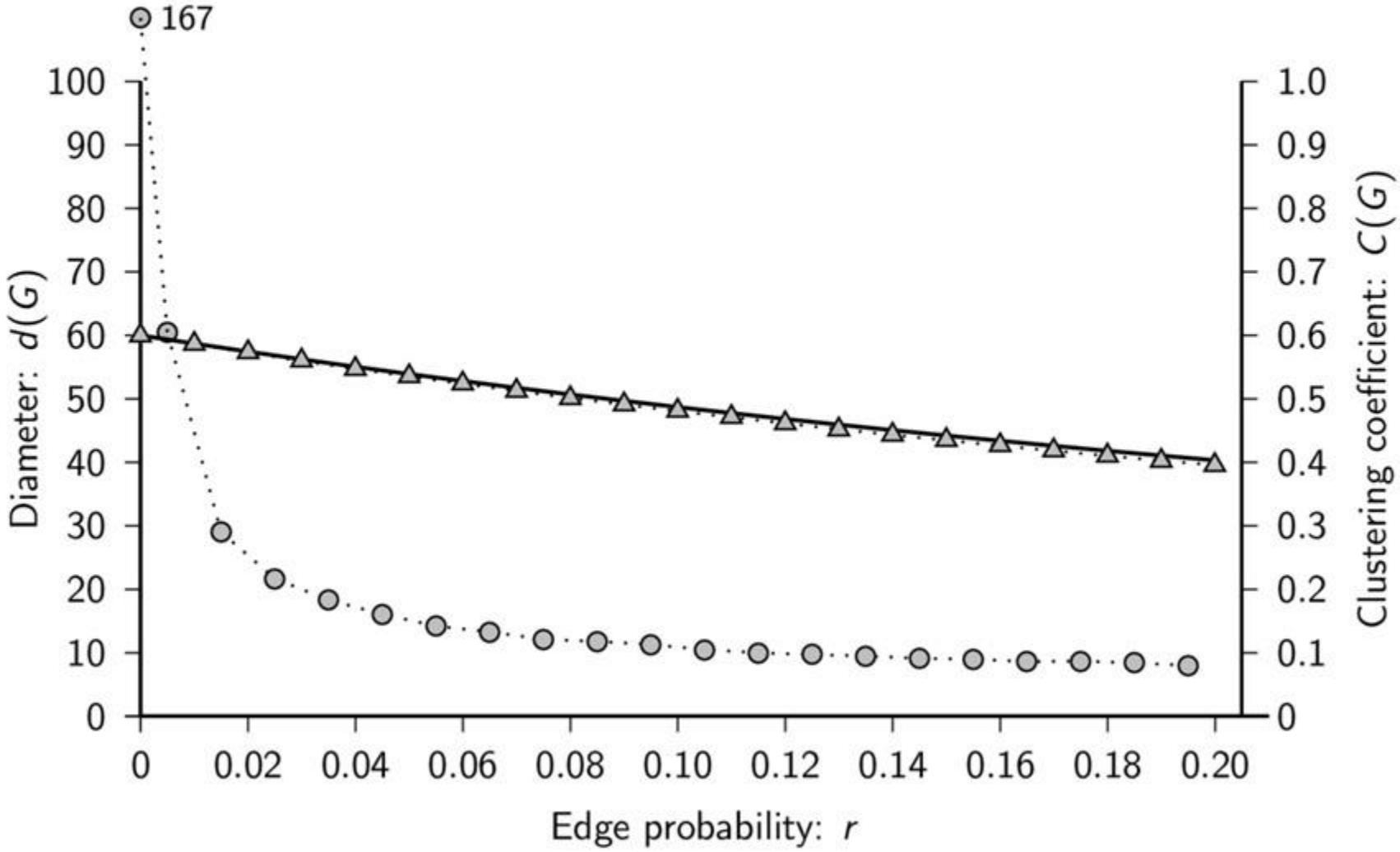
Clustering Coefficient: The clustering coefficient is

$$C(v) \simeq \frac{3(k-1)}{(1+r)(4kr+2(2k-1))} = \frac{3k-3}{4k-2+2r(2kr+4k-1)}$$

Thus, for small values of r the clustering coefficient remains high.

Diameter: Small values of shortcut edge probability r are enough to reduce the diameter from $O(n)$ to $O(\log n)$.

Watts-Strogatz Model: Diameter (circles) and Clustering Coefficient (triangles)



Barabási–Albert Scale-free Model

The Barabási–Albert (BA) yields a scale-free degree distribution based on *preferential attachment*; that is, edges from the new vertex are more likely to link to nodes with higher degrees.

Let G_t denote the graph at time t, and let n_t denote the number of nodes, and m_t the number of edges in G_t.

Initialization: The BA model starts with G_0, with each node connected to its left and right neighbors in a circular layout. Thus $m_0 = n_0$.

Growth and Preferential Attachment: The BA model derives a new graph G_{t+1} from G_t by adding exactly one new node u and adding $q \leq n_0$ new edges from u to q distinct nodes $v_j \in G_t$, where node v_j is chosen with probability $\pi_t(v_j)$ proportional to its degree in G_t, given as

$$\pi_t(v_j) = \frac{d_j}{\sum_{v_i \in G_t} d_i}$$

Barabási–Albert Graph

$n_0 = 3,\ q = 2,\ t = 12$

At $t = 0$, start with 3 vertices v_0, v_1, and v_2 fully connected (shown in gray).

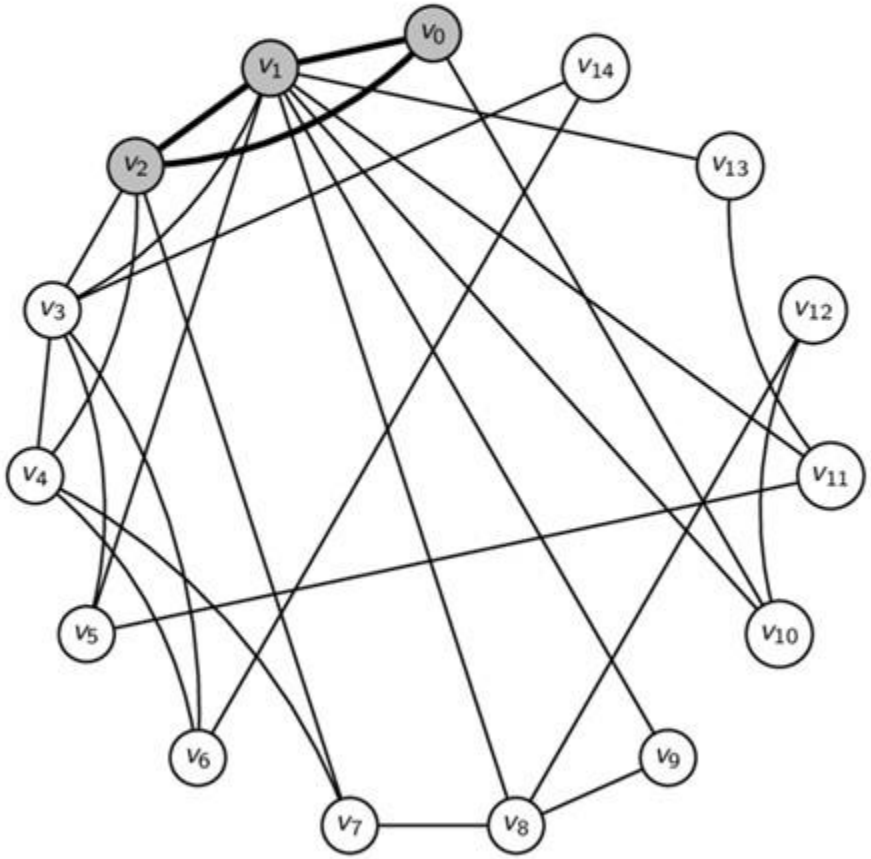
At $t = 1$, vertex v_3 is added, with edges to v_1 and v_2, chosen according to the distribution

$$\pi_0(v_i) = 1/3 \text{ for } i = 0, 1, 2$$

At $t = 2$, v_4 is added. Nodes v_2 and v_3 are preferentially chosen according to the probability distribution

$$\pi_1(v_0) = \pi_1(v_3) = \frac{2}{10} = 0.2$$

$$\pi_1(v_1) = \pi_1(v_2) = \frac{3}{10} = 0.3$$



Properties of the BA Graphs

Degree Distribution: The degree distribution for BA graphs is given as

$$f(k) = \frac{(q+2)(q+1)q}{(k+2)(k+1)k} \cdot \frac{2}{(q+2)} = \frac{2q(q+1)}{k(k+1)(k+2)}$$

For constant q and large k, the degree distribution scales as

$$f(k) \propto k^{-3}$$

The BA model yields a power-law degree distribution with $\gamma = 3$, especially for large degrees.

Diameter: The diameter of BA graphs scales as

$$d(G_t) = O\left(\frac{\log n_t}{\log \log n_t}\right)$$

suggesting that they exhibit *ultra-small-world* behavior, when $q > 1$.

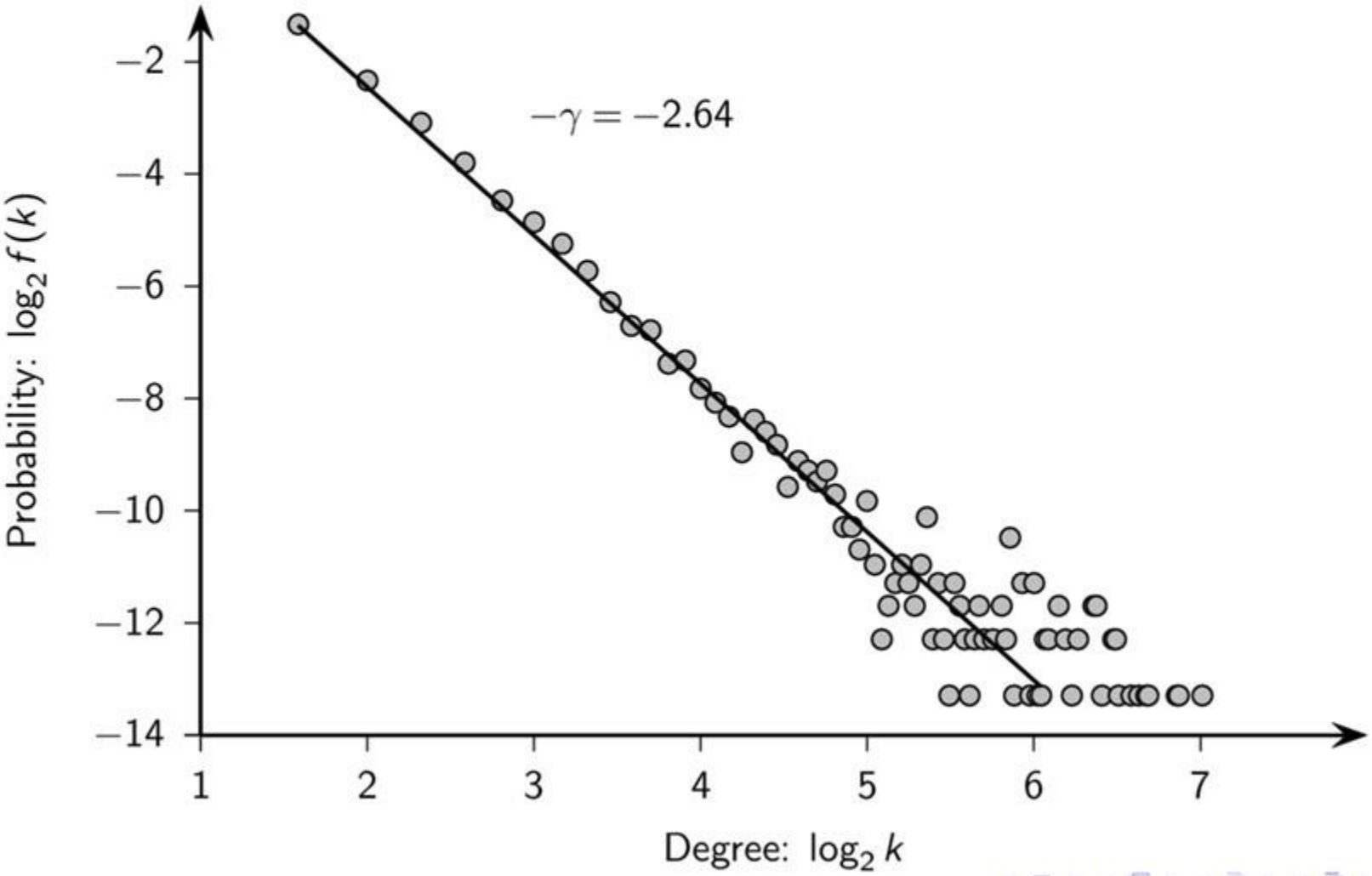
Clustering Coefficient: The expected clustering coefficient of the BA graphs scales as

$$E[C(G_t)] = O\left(\frac{(\log n_t)^2}{n_t}\right)$$

which is only slightly better than for random graphs.

Barabási–Albert Model: Degree Distribution

$n_0 = 3, t = 997, q = 3$



Data Mining and Machine Learning: 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 4: Graph Data