Data Mining and Machine Learning: Fundamental Concepts and Algorithms

datamininingbook.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 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:

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

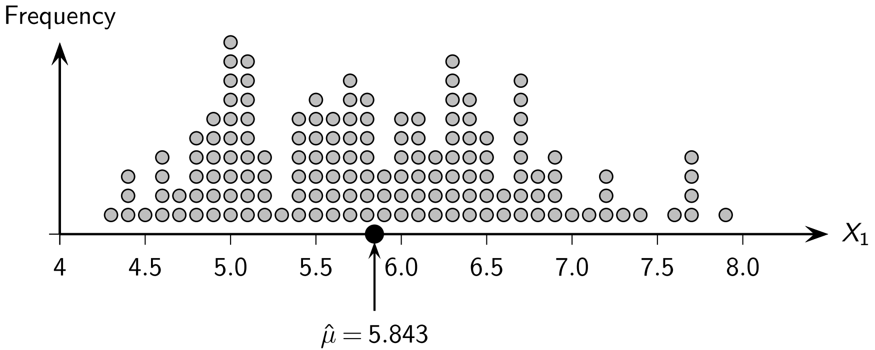
It serves as an estimator for the unknown mean value μ of X.

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.

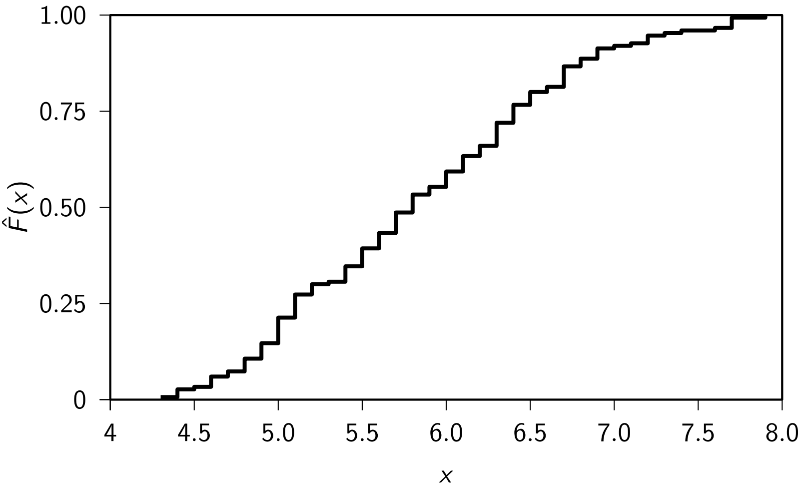
Mode

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.

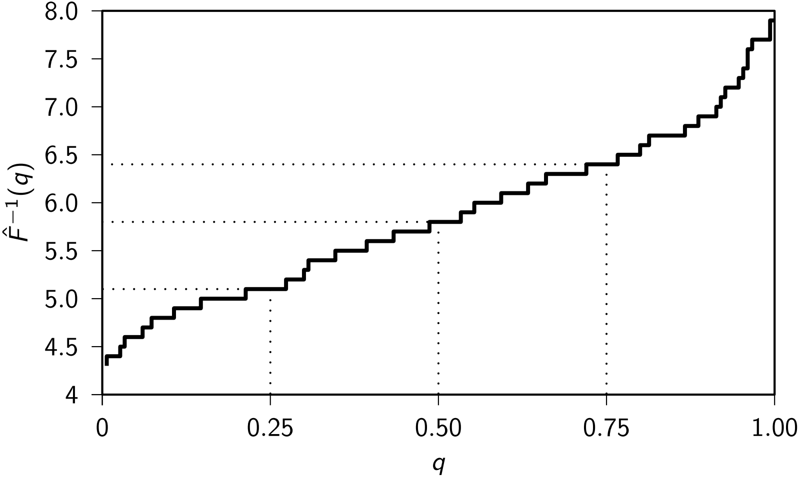
The *sample mode* is a value for which the empirical probability mass function attains its maximum, given as

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

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

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

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

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

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

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

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 $\mathbb{R}^n$:

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

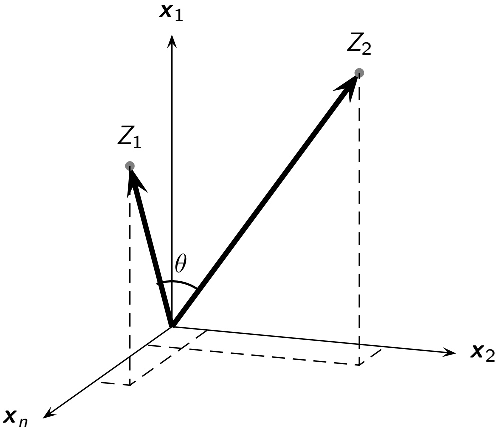
The sample covariance and the sample correlation are given as

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

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

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

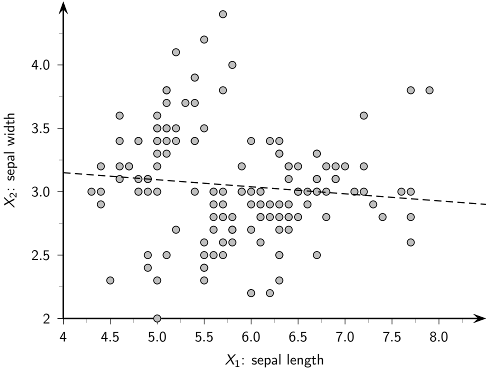
$$var(\boldsymbol{D}) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2$$

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

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

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 $X_1, X_2, \ldots, X_d$.

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

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, $\boldsymbol{X} = (X_1, X_2, \dots, X_d)^T$, 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

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

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

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

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

Sample Covariance Matrix: Inner and Outer Product

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

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

Inner product and outer product form for sample covariance matrix:

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

i.e., $\widehat{\Sigma}$ 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

$\mathbf{x}_i$	Age (X_1)	Income (X_2)
$\mathbf{x}_1$	12	300
$\mathbf{x}_2$	14	500
$\mathbf{x}_3$	18	1000
$\mathbf{x}_4$	23	2000
$\mathbf{x}_5$	27	3500
$\mathbf{x}_6$	28	4000
$\mathbf{x}_7$	34	4300
$\mathbf{x}_8$	37	6000
$\mathbf{x}_9$	39	2500
$\mathbf{x}_{10}$	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 $\mathbf{x}_2 = (14, 500)$ is scaled to

$$\mathbf{x}_2' = \left(\frac{14-12}{28}, \frac{500-300}{5700}\right) = (0.071, 0.035)$$

For z-normalization, we have

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

Thus, $\mathbf{x}_2 = (14, 500)$ is scaled to

$$\mathbf{x}_2' = \left(\frac{14-27.2}{9.77}, \frac{500-2680}{1726.15}\right) = (-1.35, -1.26)$$

Univariate Normal Distribution

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

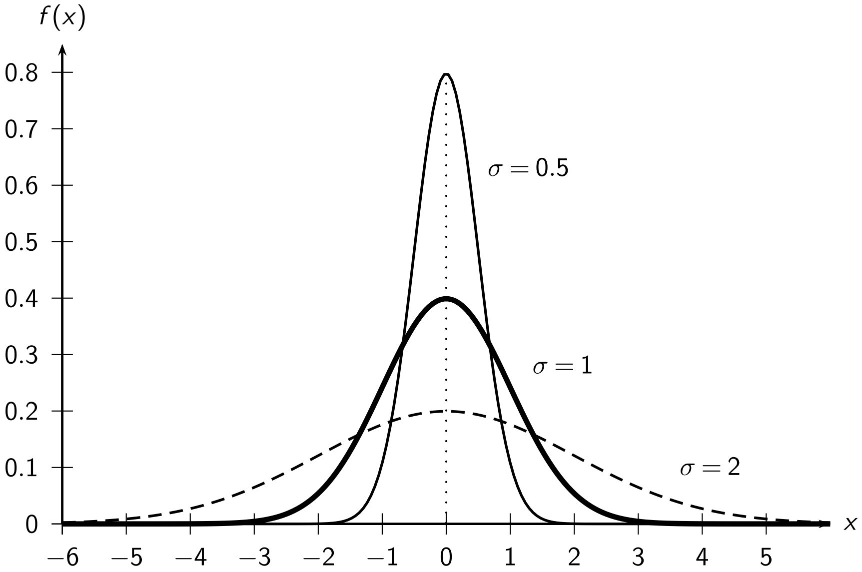
A random variable X has a normal distribution, with the parameters mean μ and variance σ^2, if the probability density function of X is given as follows:

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

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:

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

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

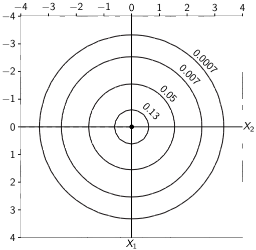
The term

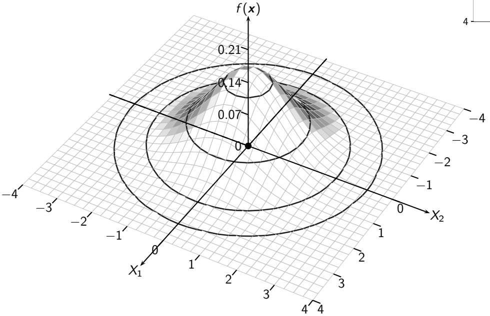
$$(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

measures the distance, called the *Mahalanobis distance* of the point $\mathbf{x}$ 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

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

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

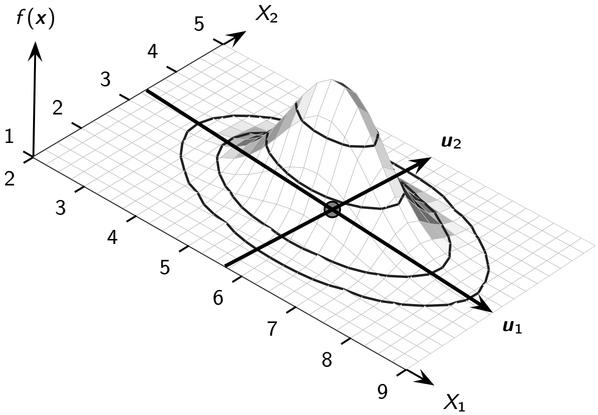
We have

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

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

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

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: Fundamental Concepts and Algorithms

datamininingbook.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 2: Numeric Attributes