Lecture 2: review basic mathematical knowledge

Dr. Huiping Cao

Outline

- Information from linear algebra
- Information from calculus
- Information from probability

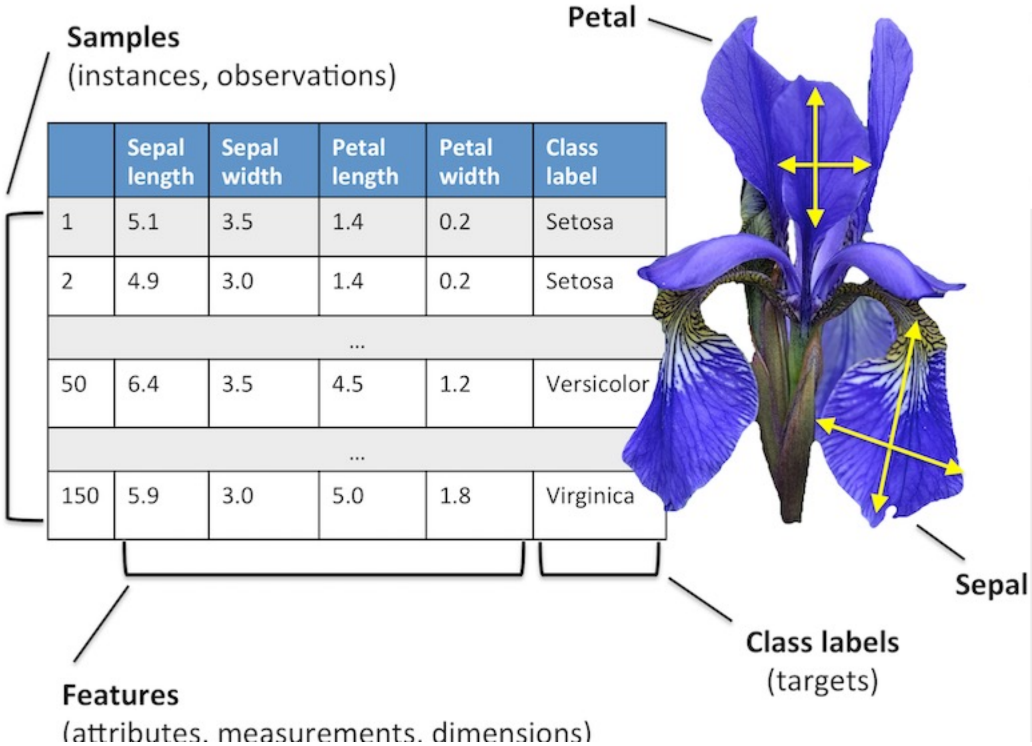
Scalar, Vector

- A **scalar value** refers to a single value. Representation: lower case letter. E.g., *x*
- A **vector** is is formally defined as an element of a vector space.
- A vector can be represented as lower case, bold-face letter. E.g., $\boldsymbol{x}$
- A vector is given by *n* coordinates and can be specified as **x** = (x_1, x_2,..., x_n).
  - A 2-dimensional vector (x_1, x_2) is often called a two-vector
  - An n-dimensional vector is often called an n-vector, and so on.
- The vector **norm** (or **magnitude, length**) of **x** = (x_1, x_2,..., x_n) is defined to be $|\boldsymbol{x}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$
- A vector with unit length is called a **unit vector**.

Matrix

- A **matrix** is is a rectangular array.
  - Example a 2×3 matrix
  - $\mathrm{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$
- Representation: upper case, bold-face letter. E.g., $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}}$
- Mathworld definition: A matrix is a concise and useful way of uniquely representing and working with **linear transformations**. In particular, every linear transformation can be represented by a matrix, and every matrix corresponds to a unique linear transformation.
- A vector is a special type of matrix.
- The individual items in an m×n matrix A, often denoted by $\mathbf{a}_{\mathbf{j}}^{\mathbf{(i)}}$, where i and j usually vary from 1 to m and n, respectively, are called its **elements** or **entries**.

Iris dataset



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Samples
(instances, observations)
Petal
Sepal
Class labels
(targets)
Features
(attributes, measurements, dimensions)

Matrix $\boldsymbol{X} \in \mathbb{R}^{\mathbf{150\times 4}}$

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \dots & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \dots & x_4^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & \dots & x_4^{(150)} \end{bmatrix}$$

$x_j^{(i)}: the\ jth\ value\ of\ the\ ith\ instance$

The i-th instance

$$\boldsymbol{x}^{(\boldsymbol{i})} \in \mathbb{R}^{\mathbf{1\times 4}} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & x_4^{(i)} \end{bmatrix}$$

The j-th feature

$$\boldsymbol{x_j} \in \mathbb{R}^{\mathbf{150\times 1}} = \begin{bmatrix} x_j^{(1)} \\ \dots \\ x_j^{(150)} \end{bmatrix}$$

Iris dataset representation

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \dots & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \dots & x_4^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & \dots & x_4^{(150)} \end{bmatrix}$$

The response variable (class labels) can be represented as a 150-dimensional column vector.

$$\boldsymbol{x}^{(\boldsymbol{i})} \in \mathbb{R}^{\mathbf{1\times 4}} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & x_4^{(i)} \end{bmatrix}$$

$$\boldsymbol{x_j} \in \mathbb{R}^{\mathbf{150\times 1}} = \begin{bmatrix} x_j^{(1)} \\ \dots \\ x_j^{(150)} \end{bmatrix}$$

$$\boldsymbol{y} \in \mathbb{R}^{\mathbf{150\times 1}} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(150)} \end{bmatrix}$$

Matrix operations

- Matrix addition
  - $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 3 & 3 & 3 \\ 4 & 4 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 6 \\ 8 & 9 & 10 \end{bmatrix}$
  - $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix} =?$

- Matrix subtraction
  - $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} - \begin{bmatrix} 3 & 3 & 3 \\ 4 & 4 & 4 \end{bmatrix} = \begin{bmatrix} -2 & -1 & 0 \\ 0 & 1 & 2 \end{bmatrix}$

- Matrix multiplication
  - $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 1\times 7 + 2\times 8 + 3\times 9 \\ 4\times 7 + 5\times 8 + 6\times 9 \end{bmatrix} = \begin{bmatrix} 50 \\ 122 \end{bmatrix}$

Matrix – linear transformation

- Linear transformation

$$y^{(1)} = x_1^{(1)}w^{(1)} + x_2^{(1)}w^{(2)} + x_3^{(1)}w^{(3)} + x_4^{(1)}w^{(4)}$$
$$y^{(2)} = x_1^{(2)}w^{(1)} + x_2^{(2)}w^{(2)} + x_3^{(2)}w^{(3)} + x_4^{(2)}w^{(4)}$$
$$\dots\ \dots$$
$$y^{(n)} = x_1^{(n)}w^{(1)} + x_2^{(n)}w^{(2)} + x_3^{(n)}w^{(3)} + x_4^{(n)}w^{(4)}$$

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ & & \dots & \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & x_4^{(n)} \end{bmatrix} . \begin{bmatrix} w^{(1)} \\ w^{(2)} \\ w^{(3)} \\ w^{(4)} \end{bmatrix}$$

Matrix transpose

The transpose of an m × n matrix **X** is the n × m matrix $\mathbf{X}^{\mathrm{T}}$

$$\mathrm{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \qquad X^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

$$x = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}, \qquad x^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

Identity matrix

- Identity matrix (unit matrix)
  - Square matrix
  - Ones on the main diagonal and zeros elsewhere
  - Denoted: I_n

$$I_1 = [1],\ I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ \cdots,\ I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

- Matrix multiplication: $X\ I_n = I_n X = X$

Matrix inverse

- An n×n matrix A is called **invertible** (also nonsingular or nondegenerate) if there exists an n×n matrix B such that A B = B A = I_n
- Matrix B is determined by A and is called the **inverse** of A, denoted as $\mathbf{A}^{-1}$

Derivative

- The **derivative** of a function of a real variable measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value).
- Derivatives are a fundamental tool of calculus.
- Example: the derivative of the position of a moving object with respect to time is the object's velocity: this measures how quickly the position of the object changes when time advances.
- A derivative can be defined on functions of a single variable.
- It is **scalar-valued**.

Gradient and partial derivative

- A **partial derivative** of a function of several variables is its derivative with respect to one of those variables, with the others held constant (as opposed to the total derivative, in which all variables are allowed to vary). Partial derivatives, e.g.:
  - $\frac{\partial(x^3+y^2+1)}{\partial x} = 3x^2, \frac{\partial(x^3+y^2+1)}{\partial y} = 2y$
- For functions of several variables, the gradient takes the place of derivative. The gradient is a vector-valued function.
- The **gradient** is a multi-variable generalization of the derivative.

Random variables, mean

- A **random variable** is described informally as a variable whose values depend on outcomes of a random phenomenon.
  - Discrete, continuous
- The **expected value (or mean)** of a discrete random variable is the probability-weighted average of all its possible values.
- The mean value of a random variable whose cumulative distribution function admits a density f(x)

$$E[X] = \int_{\mathbb{R}} xf(x)dx$$

Standard deviation

- The **standard deviation (SD)** is a measure of the amount of variation or dispersion of a set of values.
- A **low** standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set.
- A **high** standard deviation indicates that the values are spread out over a wider range.

Covariance

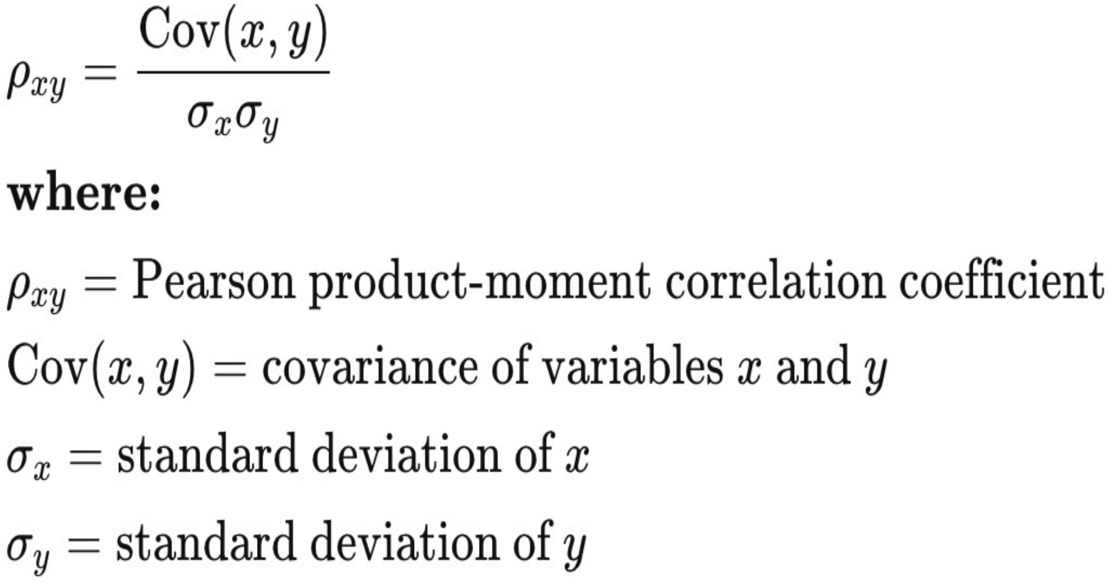
- A **covariance** matrix (also known as auto-covariance matrix, dispersion matrix, variance matrix, or variance–covariance matrix) is a square matrix giving the covariance between each pair of elements of a given random vector.
  - In the matrix diagonal there are variances, i.e., the covariance of each element with itself.
- The covariance matrix of a random vector X is typically **denoted** by $\mathbf{K}_{\mathbf{XX}}$ or **∑**.
- The (i,j) entry in the covariance matrix is defined to be

$$\mathrm{K}_{X_i X_j} = \mathrm{cov}[X_i, X_j] = \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])]$$

Correlation

- The correlation coefficient describes how one variable moves in relation to another.

- A positive correlation indicates that the two move in the same direction, with a +1.0 correlation when they move in tandem.
- A negative correlation coefficient tells you that they instead move in opposite directions. A correlation of zero suggests no correlation at all.



$$\rho_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

where:

ρ_{xy} = Pearson product-moment correlation coefficient

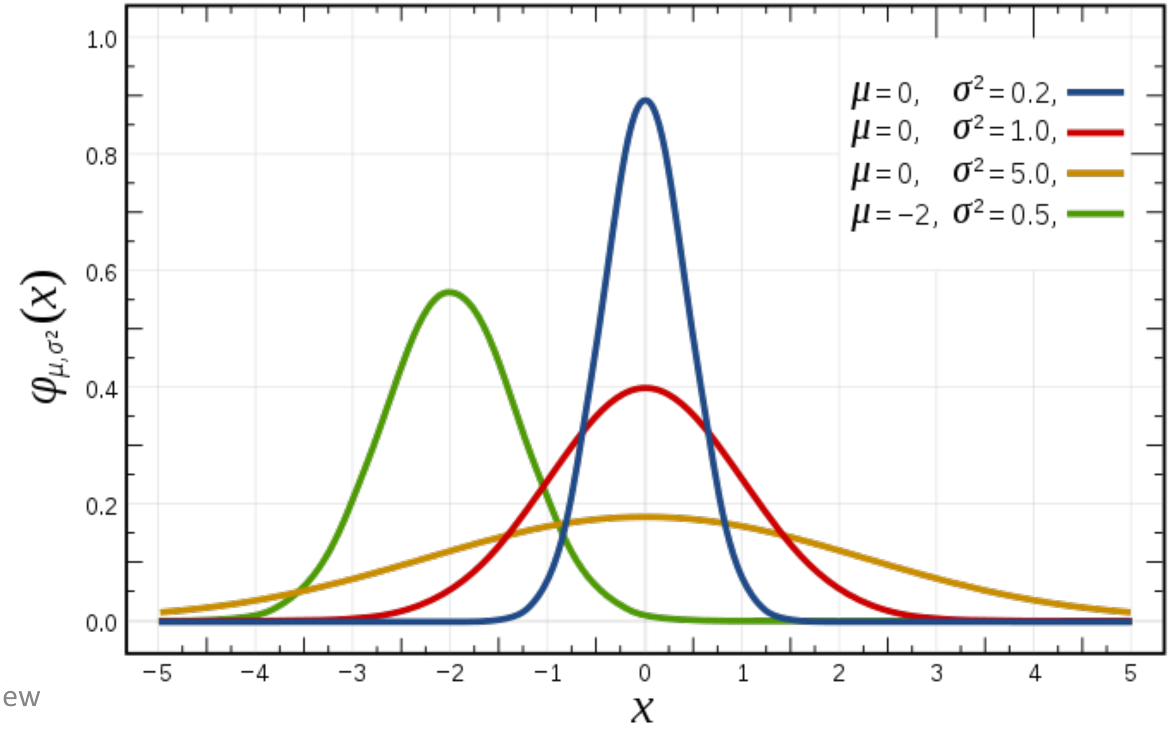
$\mathrm{Cov}(x, y)$ = covariance of variables x and y

σ_x = standard deviation of x

σ_y = standard deviation of y

Normal distribution

- A normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable.
- The general form of its probability density function is
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- μ is mean
- σ is standard deviation.
- $X \sim N(\mu, \sigma^2)$



Standard Normal Distribution

- The probability density of the standard Gaussian distribution (standard normal distribution) has zero mean and unit variance.
- $X \sim N(0,1)$
- z-score normalization: $z = \frac{x-\mu}{\sigma}$

Others

- Factorial: $n! = n \cdot (n-1) \cdot (n-2) \cdots \cdot 3 \cdot 2 \cdot 1$
    - E.g., $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$
- **Binomial coefficient** refers to the number of ways we can choose subsets of k-unordered elements from a set of size n. It is often called “n choose k”. The binomial coefficient is also referred to as **combination** or **combinatorial number** because the order of elements does not matter. It is written as follows

$$\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle = \frac{n!}{(n-k)!\,k!}$$