Basic methods for data analysis

Philipp Renz

January 3, 2022

Contents

1 Introduction

2 Summarizing univariate data

3 Linear algebra: Basics

In data analysis we will often represent our data as matrices. Consequently many algorithms are described using the concept of matrices and their properties which could be broadly called the study of linear algebra. Here I don't want to give an axiomatic introduction but rather provide some basic results with the prerequisite that the reader already knows some basics about vectors and matrices. We will denote vectors as bold lowercase letters

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \tag{1}$$

where a_i is the ith entry of the vector $\boldsymbol{a}$.
Matrices are denoted by bold uppercase letters

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}. \tag{2}$$

Depending on the context it is useful to view a matrix as a concetation of either row or column vectors.

$$\boldsymbol{A} = \begin{pmatrix} — & \boldsymbol{a}_1^r & — \\ & \vdots & \\ — & \boldsymbol{a}_m^r & — \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{a}_1^c & \dots & \boldsymbol{a}_n^c \\ | & & | \end{pmatrix}. \tag{3}$$

We made an explicit distinction between row- and column vectors using superscripts here but these will often be left out for better readability.

3.1 Matrix multiplication

Given two matrices

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in R^{m \times n}, \qquad \boldsymbol{B} = \begin{pmatrix} b_{11} & \dots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{np} \end{pmatrix} \in R^{n \times p} \tag{4}$$

their product $\boldsymbol{C} = \boldsymbol{AB}$ is of shape $(m \times p)$ and it's entries are

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} \tag{5}$$

4 Summarizing multivariate data

4.1 Bivariate data

4.2 More features

4.3 Principal component analysis

5 Linear algebra: Singular value decomposition

5.1 Another look at PCA

6 Multiple linear regression

7 Probability theory