A

Linear Algebra

This appendix provides a brief introduction to linear algebra with a focus on topics that are relevant to the material in this text. We begin by describing vectors, which can be used to represent both data objects and attributes. We then discuss matrices, which can be used both to represent data sets and to describe transformations on them.

A.1 Vectors

A.1.1 Definition

In Euclidean space, such as the ordinary two- and three-dimensional space with which we are familiar, a **vector** is a quantity that has **magnitude** and **direction**. It is traditionally represented as an arrow that has a length equal to its magnitude and an orientation given by its direction. Figure A.1(a) shows two vectors: vector $\mathbf{u}$, which has a length of 1 and is parallel to the y axis, and vector $\mathbf{v}$, which has a length of 2 and a direction of 45° with respect to the x axis. (We shall use lowercase bold letters, such as $\mathbf{u}$ and $\mathbf{v}$, to represent vectors. They are often also represented by italic lowercase letters, such as u and v.) Since a point can be regarded as a displacement from the origin in a particular direction, it can be represented by a vector from the origin to the point.

A.1.2 Vector Addition and Multiplication by a Scalar

Various operations can be performed on vectors. (In what follows, we assume that the vectors are all from the same space, i.e., have the same dimensionality.) For instance, vectors can be added and subtracted. This is best illustrated

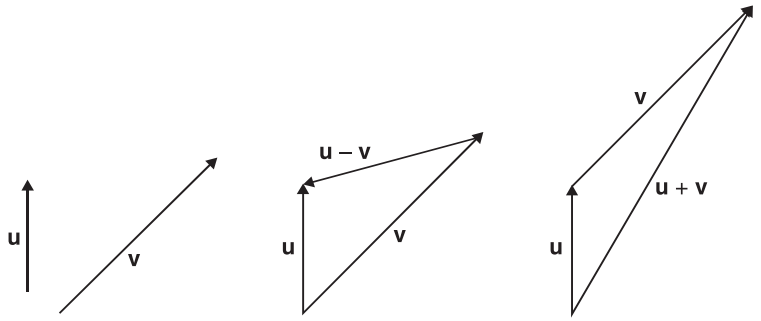

Figure A.1. Two vectors and their sum and difference.

graphically, and vector subtraction and addition are shown in Figures A.1(b) and A.1(c), respectively. Like the addition of numbers, vector addition has some familiar properties. If $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are three vectors, then these properties can be described as follows:

- **Commutativity of vector addition.** The order of addition does not matter. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.
- **Associativity of vector addition.** The grouping of vectors during addition does not matter. $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$.
- **Existence of an identity element for vector addition.** There exists a **zero vector**, simply denoted as $\mathbf{0}$, which is the identity element. For any vector $\mathbf{u}$, $\mathbf{u} + \mathbf{0} = \mathbf{u}$.
- **Existence of additive inverses for vector addition.** For every vector $\mathbf{u}$, there is an inverse vector $-\mathbf{u}$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.

Another important operation is the multiplication of a vector by a number, which, in the context of linear algebra, is typically called a **scalar**. Scalar multiplication changes the magnitude of the vector; the direction is unchanged if the scalar is positive and is reversed if the scalar is negative. If $\mathbf{u}$ and $\mathbf{v}$ are vectors and α and β are scalars (numbers), then the properties of the scalar multiplication of vectors can be described as follows:

- **Associativity of scalar multiplication.** The order of multiplication by two scalars does not matter. $\alpha(\beta u) = (\alpha\beta)u$.
- **Distributivity of scalar addition over multiplication of a scalar by a vector.** Adding two scalars and then multiplying the resulting sum by a vector is the same as multiplying each scalar by the vector and then adding the two resultant vectors. $(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}$.
- **Distributivity of scalar multiplication over vector addition.** Adding two vectors and then multiplying the sum by a scalar is the same as multiplying each vector by the scalar and then adding. $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$.
- **Existence of scalar identity.** If $\alpha = 1$, then for any vector $\mathbf{u}$, $\alpha\mathbf{u} = \mathbf{u}$.

A.1.3 Vector Spaces

A **vector space** is a set of vectors, along with an associated set of scalars (e.g., the real numbers) that satisfies the properties given above and that is closed under vector addition and multiplication by a scalar. (By closed, we mean that every result of vector addition and/or scalar multiplication results in a vector in the original set.) Vector spaces have the property that any vector can be represented as a **linear combination** of a small set of vectors, which are known as a **basis**. More specifically, if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are the basis vectors, then we can find a set of n scalars $\{\alpha_1, \ldots, \alpha_n\}$ for any vector $\mathbf{v}$, so that $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{u}_i$. We say that the basis vectors **span** the vector space. The **dimension** of a vector space is the minimum number of vectors that are necessary to form a basis. Typically, the basis vectors are taken to have unit length.

The basis vectors are usually **orthogonal**. The orthogonality of vectors is an extension of the two-dimensional notion of perpendicular lines and will be defined more precisely later on. Conceptually, orthogonal vectors are unrelated or independent. If basis vectors are mutually orthogonal, then expressing a vector as a linear combination of basis vectors effectively decomposes the vector into a number of **independent components**.

Thus, a vector in an n-dimensional space can be considered to be an n-tuple of scalars (numbers). To provide a concrete illustration, consider two-dimensional Euclidean space, where each point is associated with a vector that represents the displacement of the point from the origin. The displacement vector to any point can be written as the sum of a displacement in the x

direction and a displacement in the y direction, which are, respectively, the x and y coordinates of the point.

We will refer to the components of a vector $\mathbf{v}$ by using the notation $\mathbf{v} = (v_1, v_2, \ldots, v_{n-1}, v_n)$. (With reference to the equation, $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{u}_i$, $v_i = \alpha_i$.) Note that v_i is a component of $\mathbf{v}$, while $\mathbf{v}_i$ is one of a set of vectors.

With a component view of vectors, the addition of vectors becomes simple to understand; to add two vectors, we simply add corresponding components. For example, (2,3) + (4,2) = (6,5). To multiply a vector by a scalar, we multiply each component by the scalar, e.g., 3 * (2,3) = (6,9).

A.1.4 The Dot Product, Orthogonality, and Orthogonal Projections

We now define what it means for two vectors to be orthogonal. For simplicity, we restrict ourselves to Euclidean vector spaces, although the definitions and results are easily generalized. We begin by defining the **dot product** of two vectors.

Definition A.1 (Dot Product)**.** The dot product $\mathbf{u} \cdot \mathbf{v}$ of two vectors, $\mathbf{u}$ and $\mathbf{v}$, is given by the following equation:

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i. \tag{A.1}$$

In words, the dot product of two vectors is computed by multiplying corresponding components of a vector and then adding the resulting products. For instance, $(2, 3) \cdot (4, 1) = 2 * 4 + 3 * 1 = 11$.

In Euclidean space it can be shown that the dot product of two (non-zero) vectors is 0 if and only if they are perpendicular. Geometrically, two vectors define a plane, and their dot product is 0 if and only if the angle (in the plane) between the two vectors is 90°. We say that such vectors are **orthogonal**.

The dot product can also be used to compute the length of a vector in Euclidean space, namely, length($\mathbf{u}$) = $\sqrt{\mathbf{u} \cdot \mathbf{u}}$. The length of a vector is also known as its L_2 **norm** and is written as $||\mathbf{u}||$. Given a vector $\mathbf{u}$, we can find a vector that is pointing in the same direction as $\mathbf{u}$, but is of unit length, by dividing each component of $\mathbf{u}$ by its length; i.e., by computing $\mathbf{u}/||\mathbf{u}||$. We say that we have normalized the vector to have an L_2 norm of 1.

Given the notation for the norm of a vector, the dot product of a vector can be written as

$$\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}|| \; ||\mathbf{v}|| \; cos(\theta), \tag{A.2}$$

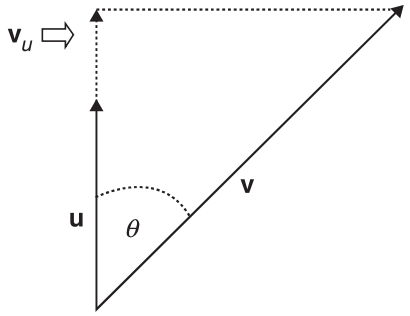

Figure A.2. Orthogonal projection of vector $\mathbf{v}$ in the direction of vector $\mathbf{u}$.

where θ is the angle between the two vectors. By grouping terms and reordering, this can be rewritten as

$$\mathbf{u} \cdot \mathbf{v} = (||\mathbf{v}||\ cos(\theta))\ ||\mathbf{u}|| = \mathbf{v}_u\ ||\mathbf{u}||, \tag{A.3}$$

where $\mathbf{v}_u = ||\mathbf{v}||\ cos(\theta)$ represents the length of $\mathbf{v}$ in the direction of $\mathbf{u}$ as illustrated in Figure A.2. If $\mathbf{u}$ is a unit vector, then the dot product is the component of $\mathbf{v}$ in the direction of $\mathbf{u}$. We refer to this as the **orthogonal projection** of $\mathbf{v}$ onto $\mathbf{u}$. Of course, it is also true that if $\mathbf{v}$ is a unit vector, then the dot product is the projection of $\mathbf{u}$ in the direction of $\mathbf{v}$.

An important consequence of this is that, given a set of orthogonal vectors of norm 1 that form a basis of a vector space, we can find the components of any vector with respect to that basis by taking the dot product of the vector with each basis vector.

A concept that is closely related to that of orthogonality is the notion of **linear independence**.

Definition A.2 (Linear Independence)**.** A set of vectors is linearly independent if no vector in the set can be written as a linear combination of the other vectors in another set.

If a set of vectors is not linearly independent, then they are **linearly dependent**. Note that we want our basis to consist of a set of vectors such that no vector is linearly dependent with respect to the remaining basis vectors, because if this were so, then we could eliminate that vector and still have a

set of vectors that span the entire vector space. If we choose our basis vectors to be mutually orthogonal (independent), then we automatically obtain a linearly independent set since any two vectors that are orthogonal are linearly independent.

A.1.5 Vectors and Data Analysis

Although vectors were originally introduced to deal with quantities such as force, velocity, and acceleration, they have proven useful for representing and understanding many other kinds of data. In particular, we can often regard a data object or an attribute as a vector. For example, Chapter 2 described a data set that consisted of 150 Iris flowers that were characterized by four attributes: sepal length, sepal width, petal length, and petal width. Each flower can be regarded as a four dimensional vector, and each attribute can be regarded as a 150 dimensional vector. As another example, a document can be represented as a vector, where each component corresponds to a term (word) and the value of each component is the number of times the term appears in the document. This yields a very sparse, high-dimensional vector, where by sparse, we mean that most of the entries of the vector are 0.

Once we have represented our data objects as vectors, we can perform various operations on the data that derive from a vector viewpoint. For example, using various vector operations, we can compute the similarity or distance of two vectors. In particular, the cosine similarity of two vectors is defined as

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}}{||\mathbf{u}||} \cdot \frac{\mathbf{v}}{||\mathbf{v}||}. \tag{A.4}$$

This similarity measure does not take into account the magnitude (length) of the vectors, but is only concerned with the degree to which two vectors point in the same direction. In terms of documents, this means that two documents are the same if they contain the same terms in the same proportion. Terms that do not appear in both documents play no role in computing similarity.

We can also simply define the distance between two vectors (points). If $\mathbf{u}$ and $\mathbf{v}$ are vectors, then the Euclidean distance between the two vectors (points) is simply

$$dist(\mathbf{u}, \mathbf{v}) = \sqrt{(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})}. \tag{A.5}$$

This type of measure is more appropriate for the Iris data, since the magnitude of the various components of the vectors does make a difference in whether they are considered to be similar.

Also, for vector data, it is meaningful to compute the mean of the set of vectors, which is accomplished by computing the mean of each component. Indeed, some clustering approaches, such as K-means (Chapter 7) work by dividing the data objects into groups (clusters) and characterizing each cluster by the mean of the data objects (data vectors). The idea is that a good cluster is one in which the data objects in the cluster are close to the mean, where closeness is measured by Euclidean distance for data like the Iris data and by cosine similarity for data like document data.

Other common operations that are performed on data can also be thought of as operations on vectors. Consider dimensionality reduction. In the simplest approach, some of the components of the data vector are eliminated, while leaving the others unchanged. Other dimensionality reduction techniques produce a new set of components (attributes) for the data vector that are linear combinations of the previous components. Still other methods change the vectors in more complicated ways. Dimensionality reduction is discussed further in Appendix B.

For certain areas of data analysis, such as statistics, the analysis techniques are expressed mathematically in terms of operations on data vectors and the data matrices that contain these data vectors. Thus, a vector representation brings with it powerful mathematical tools that can be used to represent, transform, and analyze the data.

In the remainder of this appendix, we will complete the story, by discussing matrices.

A.2 Matrices

A.2.1 Matrices: Definitions

A **matrix** is a tabular representation of a set of numbers as a collection of rows and columns. We will use uppercase bold letters, such as $\mathbf{A}$, to represent matrices. (Uppercase italic letters, such as A, are also used.) The term "m by n matrix" is commonly used to refer to a matrix with m rows and n columns. For example, the matrix $\mathbf{A}$, shown below, is a 2 by 3 matrix. If $m = n$, we say that the matrix is a **square matrix**. The transpose of $\mathbf{A}$ is written as $\mathbf{A}^T$ and is produced by interchanging the rows and columns of $\mathbf{A}$.

$$\mathbf{A} = \begin{bmatrix} 2 & 6 & 1 \\ 7 & 5 & 2 \end{bmatrix} \qquad \mathbf{A}^T = \begin{bmatrix} 2 & 7 \\ 6 & 5 \\ 1 & 2 \end{bmatrix}$$

The **matrix entries** are represented by subscripted, lowercase letters. For matrix $\mathbf{A}$, for example, a_{ij} is the entry in the i^{th} row and j^{th} column. Rows are numbered from top to bottom and columns from left to right. As a specific illustration, $a_{21} = 7$ is the entry in the second row and first column of $\mathbf{A}$.

Each row or column of a matrix defines a vector. For a matrix $\mathbf{A}$, the i^{th} **row vector** can be represented using the notation $\mathbf{a}_{i*}$ and the j^{th} **column vector** using the notation $\mathbf{a}_{*j}$. Using the previous example, $\mathbf{a}_{2*} = [7\ 5\ 2]$, while $\mathbf{a}_{*3} = [1\ 2]^T$. Notice that row and column vectors are matrices and must be distinguished; i.e., a row vector and column vector that have the same number of entries and the same values represent different matrices.

A.2.2 Matrices: Addition and Multiplication by a Scalar

Like vectors, matrices can be added by adding their corresponding entries (components). (Here we are assuming that the matrices have the same number of rows and columns.) More specifically, if $\mathbf{A}$ and $\mathbf{B}$ are two matrices having dimensions m by n, then the sum of $\mathbf{A}$ and $\mathbf{B}$ is defined as follows:

Definition A.3 (Matrix Addition)**.** The sum of two m by n matrices, $\mathbf{A}$ and $\mathbf{B}$, is an m by n matrix C, whose entries are given by the following equation:

$$c_{ij} = a_{ij} + b_{ij}. \tag{A.6}$$

For example,

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 5 & 4 \\ 2 & 9 \end{bmatrix} = \begin{bmatrix} 8 & 5 \\ 3 & 11 \end{bmatrix}.$$

Matrix addition has the following properties:

- **Commutativity of matrix addition.** The order of addition does not matter. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
- **Associativity of matrix addition.** The grouping of matrices during addition does not matter. $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$.
- **Existence of an identity element for matrix addition.** There exists a **zero matrix**, having all 0 entries and simply denoted as $\mathbf{0}$, which is the identity element. For any matrix $\mathbf{A}$, $\mathbf{A} + \mathbf{0} = \mathbf{A}$.
- **Existence of additive inverses for matrix addition.** For every matrix $\mathbf{A}$ there is a matrix $-\mathbf{A}$ such that $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$. The entries of $-\mathbf{A}$ are $-a_{ij}$.

As with vectors, we can also multiply a matrix by a scalar.

Definition A.4 (Scalar Multiplication of a Matrix)**.** The product of a scalar α and a matrix $\mathbf{A}$ is the matrix $B = \alpha A$, whose entries are given by the following equation.

$$b_{ij} = \alpha\, a_{ij} \tag{A.7}$$

Scalar multiplication has properties that are very similar to those of multiplying a vector by a scalar.

- **Associativity of scalar multiplication.** The order of multiplication by two scalars does not matter. $\alpha(\beta\mathbf{A}) = (\alpha\beta)\mathbf{A}$.
- **Distributivity of scalar addition over multiplication of a scalar by a matrix.** Adding two scalars and then multiplying the sum by a matrix is the same as multiplying each scalar times the matrix and then adding the two resultant matrices. $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$.
- **Distributivity of scalar multiplication over matrix addition.** Adding two matrices and then multiplying the sum by a scalar is the same as multiplying each matrix by the scalar and then adding. $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$.
- **Existence of scalar identity.** If $\alpha = 1$, then for any matrix $\mathbf{A}$, $\alpha\mathbf{A} = \mathbf{A}$.

None of the previous properties should be surprising since we can think of a matrix as being composed of row or column vectors, and hence, matrix addition or the multiplication by a scalar amounts to adding corresponding row or column vectors or multiplying them by a scalar.

A.2.3 Matrices: Multiplication

We can define a multiplication operation for matrices. We begin by defining multiplication between a matrix and a vector.

Definition A.5 (Multiplication of a Matrix by a Column Vector)**.** The product of an m by n matrix $\mathbf{A}$ and an n by 1 column matrix $\mathbf{u}$ is the m by 1 column matrix $\mathbf{v} = \mathbf{A}\mathbf{u}$, whose entries are given by the following equation.

$$v_i = \mathbf{a}_{i*} \cdot \mathbf{u} \tag{A.8}$$

In other words, we take the dot product of the transpose of $\mathbf{u}$ with each row vector of $\mathbf{A}$. In the following example, notice that the number of rows in $\mathbf{u}$ must be the same as the number of columns of $\mathbf{A}$.

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 17 \\ 9 \end{bmatrix}$$

We can similarly define the multiplication of a matrix by a row vector on the left side.

Definition A.6 (Multiplication of a Matrix by a Row Vector)**.** The product of a 1 by m row matrix $\mathbf{u}$ and an m by n matrix $\mathbf{A}$ is the 1 by n row matrix $\mathbf{v} = \mathbf{uA}$, whose entries are given by the following equation.

$$v_i = \mathbf{u} \cdot (\mathbf{a}_{*j})^T \tag{A.9}$$

In other words, we take the dot product of the row vector with the transpose of each column vector of $\mathbf{A}$. An example is given below.

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 2 & 9 \end{bmatrix} = \begin{bmatrix} 9 & 22 \end{bmatrix}$$

We define the product of two matrices as an extension to the above idea.

Definition A.7. The product of an m by n matrix $\mathbf{A}$ and an n by p matrix $\mathbf{B}$ is the m by p matrix $\mathbf{C} = \mathbf{AB}$, whose entries are given by the equation

$$c_{ij} = \mathbf{a}_{i*} \cdot (\mathbf{b}_{*j})^T \tag{A.10}$$

In words, the ij^{th} entry of $\mathbf{C}$ is the dot product of the i^{th} row vector of $\mathbf{A}$ and the transpose of the j^{th} column vector of $\mathbf{B}$.

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 2 & 9 \end{bmatrix} = \begin{bmatrix} 17 & 21 \\ 9 & 22 \end{bmatrix}$$

Matrix multiplication has the following properties:

- **Associativity of matrix multiplication.** The order of multiplication of matrices does not matter. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.
- **Distributivity of matrix multiplication.** Matrix multiplication is distributive with respect to matrix addition. $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ and $(\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{BA} + \mathbf{CA}$.

- **Existence of an identity element for matrix multiplication.** If $\mathbf{I}_p$ is the p by p matrix with 1's only on the diagonal and 0 elsewhere, then for any m by n matrix $\mathbf{A}$, $\mathbf{AI}_n = \mathbf{A}$ and $\mathbf{I}_m\mathbf{A} = \mathbf{A}$. (Note that the identity matrix is an example of a **diagonal matrix**, which is a matrix whose off diagonal entries are all 0, i.e., $a_{ij} = 0$, if $i \neq j$.)

In general, matrix multiplication is not commutative, i.e., $\mathbf{AB} \neq \mathbf{BA}$.

A.2.4 Linear Transformations and Inverse Matrices

If we have an n by 1 column vector $\mathbf{u}$, then we can view the multiplication of an m by n matrix $\mathbf{A}$ by this vector on the right as a transformation of $\mathbf{u}$ into an m-dimensional column vector $\mathbf{v} = \mathbf{Au}$. Similarly, if we multiply $\mathbf{A}$ by a (row) vector $\mathbf{u} = [u_1, \ldots, u_m]$ on the left, then we can view this as a transformation of $\mathbf{u}$ into an n-dimensional row vector $\mathbf{v} = \mathbf{uA}$. Thus, we can view any m by n matrix $\mathbf{A}$ as a function that maps one vector space onto another.

In many cases, the transformation (matrix) can be described in easily understood terms.

- A **scaling matrix** leaves the direction of the vector unchanged, but changes its length. This is equivalent to multiplying by a matrix that is the identity matrix multiplied by a scalar.

- A **rotation matrix** changes the direction of a vector, but leaves the magnitude of the vector unchanged. This amounts to a change of coordinate system.

- A **reflection matrix** reflects a vector across one or more coordinate axes. This would be equivalent to multiplying some of the entries of the vector by -1, while leaving the other entries unchanged.

- A **projection** matrix takes vectors into a lower dimensional subspace. The simplest example is the modified identity matrix where one or more of the 1's on the diagonal have been changed into 0's. Such a matrix eliminates the vector components corresponding to those zero entries, while preserving all others.

Of course, a single matrix can do two kinds of transformations at once, e.g., scaling and rotation.

Following are a few properties of matrices when viewed as functions that map vectors from one vector space to another.

- Matrices are **linear transformations**, i.e., $\mathbf{A}(\alpha\mathbf{u}+\beta\mathbf{v}) = \alpha\mathbf{A}\mathbf{u}+\beta\mathbf{A}\mathbf{v}$ and $(\alpha\mathbf{u}+\beta\mathbf{v})\mathbf{A} = \alpha\mathbf{u}\mathbf{A}+\beta\mathbf{v}\mathbf{A}$.

- The set of all transformed row vectors of a matrix $\mathbf{A}$ is called the **row space** of $\mathbf{A}$ because the row vectors of the matrix, or some subset of them, form a basis for the subspace of transformed row vectors. This is evident from the following equation, which expresses the product of a 1 by m row vector $\mathbf{u} = [u_1, \ldots, u_m]$ and an m by n matrix $\mathbf{A}$ as a linear combination of the rows of the matrix.

$$\mathbf{v} = \mathbf{u}\mathbf{A} = \sum_{i=1}^{m} u_i \mathbf{a}_{i*} \tag{A.11}$$

  The dimension of the row space tells us the number of linearly independent rows of $\mathbf{A}$.

- The set of all transformed column vectors is called the **column space** of $\mathbf{A}$. The column vectors of the matrix, or some subset of them, form a basis for the subspace of transformed column vectors. This is clear from the following equation, which expresses the product of an n by 1 column vector $\mathbf{u} = [u_1, \ldots, u_n]^T$ and an m by n matrix $\mathbf{A}$ as a linear combination of the columns of the matrix.

$$\mathbf{v} = \mathbf{A}\mathbf{u} = \sum_{j=1}^{n} u_j \mathbf{a}_{*j} \tag{A.12}$$

  The dimension of the column space tells us the number of linearly independent columns of $\mathbf{A}$.

- The **left nullspace** is the set of row vectors that the matrix maps to 0.

- The **right nullspace** (or more commonly, just nullspace) is the set of column vectors that the matrix maps to 0.

Note that the **rank of a matrix** is the minimum of the dimensionality of the row space and column space and is often used to characterize a matrix. For instance, if we take a single 1 by n row vector and duplicate it m times to create an m by n matrix, we would only have a matrix of rank 1.

A question of practical and theoretical importance is whether matrices, like real numbers, have multiplicative inverses. First, we note that because of the nature of matrix multiplication (i.e., dimensions have to match), a matrix

must be square if it is to have an **inverse matrix**. Thus, for an m by m matrix $\mathbf{A}$, we are asking if we can find a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_m$. The answer is that some square matrices have inverses and some do not.

More abstractly, an m by m matrix has an inverse only if both of its null spaces contain only the 0 vector, or if, equivalently, the row and column spaces are both of dimension m. (This is equivalent to the rank of the matrix being m.) Conceptually, an m by m matrix has an inverse if and only if it uniquely maps every non-zero m-dimensional row (column) vector onto a unique, non-zero m-dimensional row (column) vector.

The existence of an inverse matrix is important when solving various matrix equations.

A.2.5 Eigenvalue and Singular Value Decomposition

We now discuss a very important area of linear algebra: eigenvalues and eigenvectors. Eigenvalues and eigenvectors, along with the related concept of singular values and singular vectors, capture the structure of matrices by allowing us to factor or decompose matrices and express them in a standard format. For that reason, these concepts are useful in the solution of mathematical equations and for dimensionality and noise reduction. We begin with the definition of eigenvalues and eigenvectors.

Definition A.8 (Eigenvectors and Eigenvalues)**.** The eigenvalues and eigenvectors of an n by n matrix $\mathbf{A}$ are, respectively, the scalar values λ and the vectors $\mathbf{u}$ that are solutions to the following equation.

$$A\mathbf{u} = \lambda\mathbf{u} \tag{A.13}$$

In other words, **eigenvectors** are the vectors that are unchanged, except for magnitude, when multiplied by $\mathbf{A}$. The **eigenvalues** are the scaling factors. This equation can also be written as $(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}$.

For square matrices, it is possible to decompose the matrix using eigenvalues and eigenvectors.

Theorem A.1. *Assume that* $\mathbf{A}$ *is an* n *by* n *matrix with* n *independent (orthogonal) eigenvectors,* $u_1, \ldots, u_n$ *and* n *corresponding eigenvalues,* $\lambda_1, \ldots, \lambda_n$*. Let* $\mathbf{U}$ *be the matrix whose columns are these eigenvectors, i.e.,* $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ *and let* $\mathbf{\Lambda}$ *be a diagonal matrix, whose diagonal entries are the* λ_i*,* $1 \leq i \leq n$*. Then* $\mathbf{A}$ *can be expressed as*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}. \tag{A.14}$$

Thus, $\mathbf{A}$ can be decomposed into a product of three matrices. $\mathbf{u}$ is known as the **eigenvector matrix** and $\mathbf{\Lambda}$ as the **eigenvalue matrix**.

More generally, an arbitrary matrix can be decomposed in a similar way. Specifically, any m by n matrix $\mathbf{A}$ can be factored into the product of three matrices as described by the following theorem.

Theorem A.2. *Assume that* $\mathbf{A}$ *is an* m *by* n *matrix. Then* $\mathbf{A}$ *can be expressed as follows*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{A.15}$$

Where $\mathbf{U}$ *is* m *by* m, Σ *is* m *by* n, *and* $\mathbf{V}$ *is* n *by* n. $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal matrices, i.e., their columns are of unit length and are mutually orthogonal. Thus,* $\mathbf{U}\mathbf{U}^T = \mathbf{I}_m$ *and* $\mathbf{V}\mathbf{V}^T = \mathbf{I}_n$. $\mathbf{\Sigma}$ *is a diagonal matrix whose diagonal entries are non-negative and are sorted so that the larger entries appear first, i.e.,* $\sigma_{i,i} \geq \sigma_{i+1,i+1}$

The column vectors of $\mathbf{V}$, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the **right singular vectors**, while the columns of $\mathbf{U}$ are the **left singular vectors**. The diagonal elements of $\mathbf{\Sigma}$, the **singular value matrix**, are typically written as $\sigma_1, \ldots, \sigma_n$ and are called the **singular values** of $\mathbf{A}$. (This use of σ should not be confused with the use of σ to represent the standard deviation of a variable.) There are at most $rank(A) \leq \min(m, n)$ non-zero singular values.

It can be shown that the eigenvectors of $\mathbf{A}^T\mathbf{A}$ are the right singular vectors (i.e., the columns of $\mathbf{V}$), while the eigenvectors of $\mathbf{A}\mathbf{A}^T$ are the left singular vectors (i.e., the columns of $\mathbf{U}$). The non-zero eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are the σ_i^2, i.e., the squares of the singular values. Indeed, the eigenvalue decomposition of a square matrix can be regarded as a special case of singular value decomposition.

The singular value decomposition (SVD) of a matrix can also be expressed with the following equation. Note that while $\mathbf{u}_i\mathbf{v}_i^T$ might look like a dot product, it is not, and the result is a rank 1 m by n matrix.

$$\mathbf{A} = \sum_{i=1}^{rank(\mathbf{A})} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{A.16}$$

The importance of the above representation is that every matrix can be expressed as a sum of rank 1 matrices that are weighted by singular values. Since singular values, which are sorted in non-increasing order, often decline rapidly in magnitude, it is possible to obtain a good approximation of a matrix by using only a few singular values and singular vectors. This is useful for dimensionality reduction and will be discussed further in Appendix B.

A.2.6 Matrices and Data Analysis

We can represent a data set as a data matrix, where each row is a data object and each column is an attribute. (We can, with equal validity, have attributes as rows and objects as columns.) Matrix representation provides a compact and well-structured representation for our data and permits the easy manipulation of the objects or attributes of the data through various matrix operations.

Systems of linear equations are one very common example of the usefulness of the matrix representation of data. A system of linear equations can be written as the matrix equation $\mathbf{Ax} = \mathbf{b}$ and solved using matrix operations.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

In particular, if $\mathbf{A}$ has an inverse, the system of equations has a solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. If not, then the system of equations has either no solution or an infinite number of solutions. Note that in this case, our rows (data objects) were equations and our columns were variables (attributes).

For many statistical and data analysis problems, we want to solve linear systems of equations, but these equations cannot be solved in the manner just described. For example, we may have a data matrix where the rows represent patients and the columns represent characteristics of the patients—height, weight, and age—and their response to a particular medication, e.g., a change in blood pressure. We want to express blood pressure (the independent variable) as a linear function of the other (dependent) variables, and we can write a matrix equation in much the same way as above. However, if we have more patients than variables—the usual case—the inverse of the matrix does not exist.

In this case, we still want to find the best solution for the set of equations. This means that we want to find the best linear combination of the independent variables for predicting the dependent variable. Using linear algebra terminology, we want to find the vector $\mathbf{Ax}$ that is as close to $\mathbf{B}$ as possible; in other words, we want to minimize $||\mathbf{b} - \mathbf{Ax}||$, which is the length of the vector $\mathbf{b} - \mathbf{Ax}$. This is known as the **least squares** problem. Many statistical techniques (e.g., **linear regression**, which is discussed in Appendix D) require

the solution of a least squares problem. It can be shown that the least squares solution of the equation $\mathbf{Ax} = \mathbf{b}$ is $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.

Singular value and eigenvalue decomposition are also very useful in analyzing data, particularly in the area of dimensionality reduction, which is discussed in Appendix B. Note that noise reduction can also occur as a side effect of dimensionality reduction.

While we have given a few examples of the application of linear algebra, we have omitted many more. Examples of other areas where linear algebra is important in the formulation and solution of problems include solving systems of differential equations, optimization problems (such as linear programming), and graph partitioning.

A.3 Bibliographic Notes

There are many books that provide good coverage of linear algebra, including those by Demmel [758], Golub and Van Loan [759], and Strang [760].

Bibliography

[758] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM Press, September 1997.

[759] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, November 1996.

[760] G. Strang. *Linear Algebra and Its Applications*. Harcourt Brace & Company, Orlando, FL, 3rd edition, 1986.