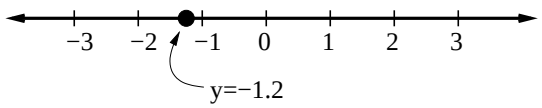
15.062
Data Mining: Algorithms and Applications
Matrix Math Review

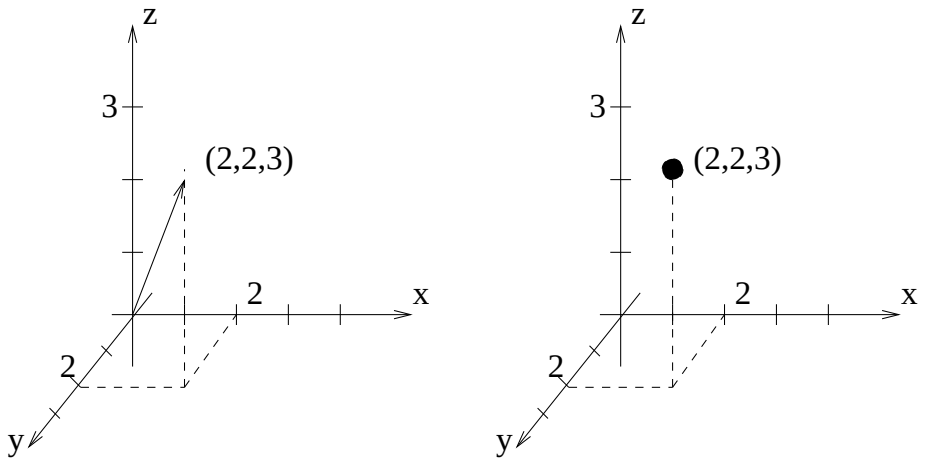
The purpose of this document is to give a brief review of selected linear algebra concepts that will be useful for the course and to develop some intuition for these concepts.

1 Introduction and Definitions

A **scalar** is simply a value or number, such as 7, -1.2, or 1266. We typically refer to scalars using lower case variables (e.g. $x = 7$, $y = -1.2$, $z = 1266$), and we can visualize them as lying on the real number line:



A **vector** is an array of scalars, such as $(-1/2, 3/2)$, $(2, 2, 3)$, or $(-3, 0, 3, 0, 0, 4, 2, 1)$. We typically refer to vectors using bold lower case variables (e.g. $\mathbf{w} = (2, 2, 3)$). For vectors of dimension 2 or 3, we can visualize them as points in a coordinate space, or as "arrows" in a coordinate space pointing away from the origin point (0,0):



A **matrix** is a table of numbers. Examples include:

$$A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}, \quad C = \begin{bmatrix} 1/2 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/2 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/2 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/2 \end{bmatrix}$$

We typically refer to matrices using upper case variables. To refer to an element of a matrix, we use subscripts. For instance, M_{12} refers to the element in the first row and second column of the matrix M. In the matrix A, we have $A_{12} = 2$. The **dimensions** of a matrix are the number of rows and columns. A $m \times n$ matrix has m rows and n columns. The matrix A above has dimensions 2×2, B has dimensions 2×4, and C has dimensions 4×4. Note that we can think of a vector as a special case of a matrix with one of the dimensions equal to 1. That is, the vector $\mathbf{w} = [2\ 2\ 3]$ can be thought of as a matrix of dimensions 1×3.

A matrix with the same number of rows as columns is called a **square** matrix. A square matrix M for which $M_{ij} = M_{ji}$ is called a **symmetric** matrix, and a square matrix with nonzero entries only on the diagonal entries is called a **diagonal** matrix (Note that the matrix C above is an example of a matrix that is symmetric but not diagonal.). The **transpose** of a matrix M is denoted by M' and is formed by switching the rows and columns of M. That is, for any row i and column j, $M_{ij} = M'_{ji}$. For example, the transpose of the matrix A given above is:

$$A' = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix}$$

2 Matrix Operations

2.1 Matrix Addition and Subtraction

We can add (or subtract) matrices that have the same dimensions by simply adding (or subtracting) on an element-by-element basis. For instance:

$$\begin{bmatrix} 1 & 1 & 1 \\ 3 & 4 & -1 \end{bmatrix} + \begin{bmatrix} 0 & -1 & 7 \\ 2 & 4 & -5 \end{bmatrix} = \begin{bmatrix} 1 & -0 & 8 \\ 5 & 8 & -6 \end{bmatrix}$$

$$\begin{bmatrix} 1.0 & 0.5 \\ 1.5 & 4.5 \end{bmatrix} - \begin{bmatrix} 0.0 & 1.0 \\ 4.0 & -5.0 \end{bmatrix} = \begin{bmatrix} 1.0 & -0.5 \\ -2.5 & 9.5 \end{bmatrix}$$

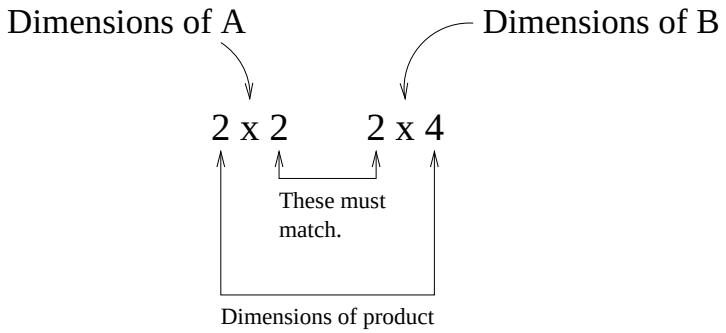
2.2 Scalar Multiplication

We can also multiply a matrix of any dimensions by a scalar by performing the multiplication on an element-by-element basis. For instance:

$$2 * \begin{bmatrix} 1 & 1 & 1 \\ 3 & 4 & -1 \\ 2 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 \\ 6 & 8 & -2 \\ 4 & 4 & 0 \end{bmatrix}$$

2.3 Matrix Multiplication

Mutliplication of two matrices is somewhat more complicated. Matrix multiplication is only defined for matrices of certain dimensions. To multiply two matrices, the number of columns in the first matrix must equal the number of rows in the second matrix. The resulting product matrix will then have the same number of rows as the first matrix and the same number of columns as the second matrix. For instance, suppose we wanted to multiply a 2×2 matrix A by a 2×4 matrix B. We can perform this multiplication, since the number of columns of A equals the number of rows of B (2). The resulting matrix will have dimensions 2×4. A diagram:



We cannot, however, multiply a 2×2 matrix A by a 4×4 matrix C, since the number of columns of A (2), is different from the number of columns of C (4).

To see how to multiply two matrices, let's look first at the special case of multiplying a 1×3 matrix by a 3×1 matrix. Consider:

$$\mathbf{a} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix}$$

We denote the matrix product of **a** and **b** as **ab**, and it is computed by multiplying the corresponding elements of **a** and **b**, then summing the results. So,

$$\begin{aligned} \mathbf{ab} &= \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix} \\ &= [(1 \times 10) + (2 \times 20) + (3 \times 30)] \\ &= [10 + 40 + 90] \\ &= [140] \end{aligned}$$

We can also multiply matrices with multiple rows and columns. Consider:

$$A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$$

Notice that A has 2 columns and D has 2 rows, so we can multiply AD. Element $1,1$ of the result is the matrix product of the first row of A and the first column of D. Element $1,2$ is the first row of A times the second column of D, and so forth. To illustrate:

$$\begin{aligned} AD = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix} &= \begin{bmatrix} \begin{bmatrix} -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} & \begin{bmatrix} -1 & 2 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} & \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} (-1 \times 3) + (2 \times 5) & (-1 \times 4) + (2 \times 6) \\ (0 \times 3) + (1 \times 5) & (0 \times 4) + (1 \times 6) \end{bmatrix} \\ &= \begin{bmatrix} 7 & 8 \\ 5 & 6 \end{bmatrix} \end{aligned}$$

Matrix multiplication is **associative**, meaning that if we would like to compute ABC, we can first compute AB then multiply by C, or we can multiply A times BC. In symbols, this means that $(AB)C = A(BC)$. It is also **distributive**, meaning that as long as the dimensions are appropriate for matrix multiplication, we have $A(B + C) = AB + AC$. Matrix multiplication, however, is not **commutative**, meaning that in general $AB \neq BA$.

2.3.1 Linear Combinations

Suppose we have a set of k vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}$ (all of the same dimension), and a set of k scalars $a_1, a_2, \ldots, a_k$. Then a sum of the following form,

$$a_1\mathbf{x_1} + a_2\mathbf{x_2} + \cdots + a_k\mathbf{x_k}$$

is called a **linear combination** of the vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}$. Note that if we use matrix multiplication to multiply two vectors, then we are really just computing a linear combination of the second vector using the first vector for scalars (or we can view it as a linear combination of the first vector using the second vector for scalars). Similarly, we can view a general matrix multiplication as computing sets of linear combinations. In this sense, multiplying by a matrix is sometimes also called a **linear transform**.

2.3.2 Matrices as a Notational Device

Matrices and vectors allow us to significantly simplify notation. Consider, for instance, a linear system of 4 equations and 4 unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= b_3 \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= b_4 \end{aligned}$$

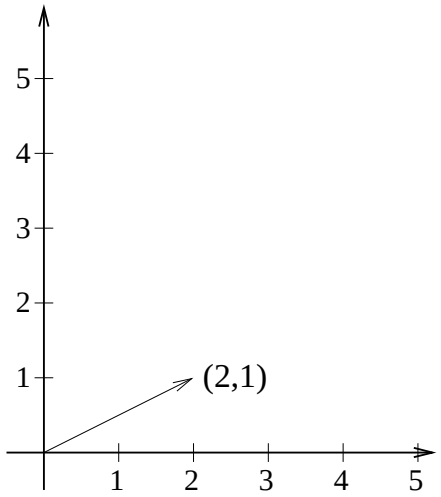
If we define:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

We can use matrix and vector notation to write the entire system of equations simply as $A\mathbf{x} = \mathbf{b}$.

2.3.3 Graphical Interpretation of Matrix Multiplication

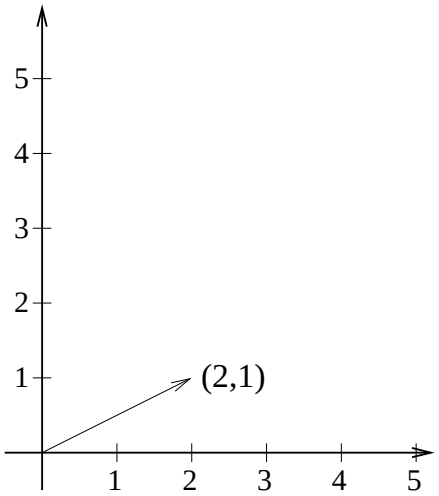
We can interpret matrix multiplication as a transformation of vectors in the coordinate space. For illustration, let's consider the case where we multiply a 2×1 matrix (vector) by a 2×2 matrix. The following is a plot of the vector $\mathbf{x} = (2, 1)$:



Observe the effect of multiplying the vector by the matrix T for various choices:

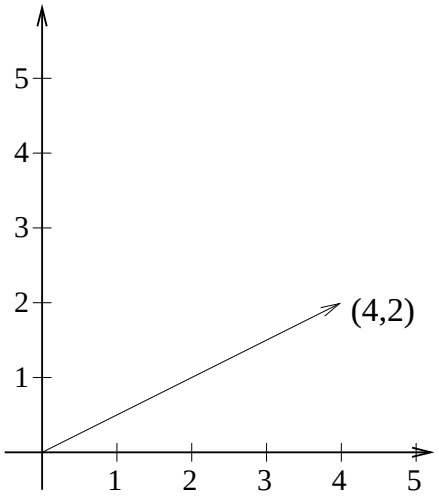
- $T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. This matrix is called the identity matrix and has the special property that it leaves the original vector unchanged.

$$T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$



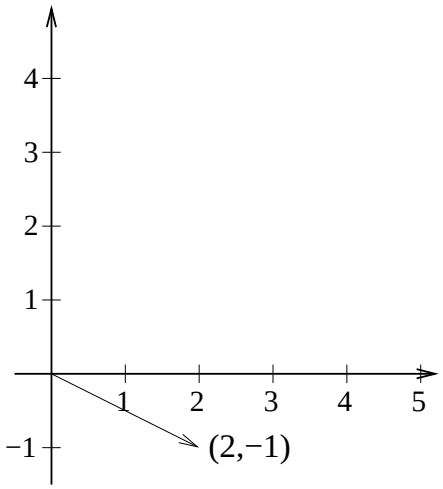
- $T = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. This matrix leaves the vector direction unchanged, but doubles its length.

$$T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$



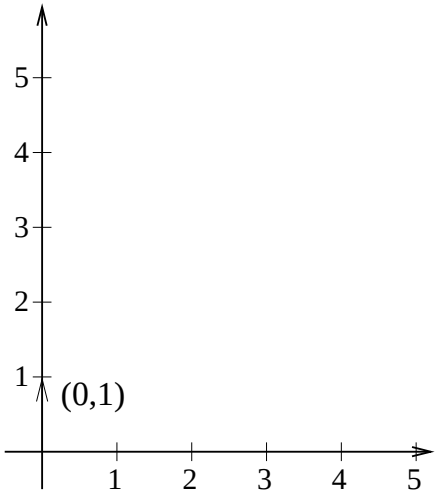
- $T = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. This matrix leaves the vector length unchanged, but flips the vector over the horizontal axis.

$$T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

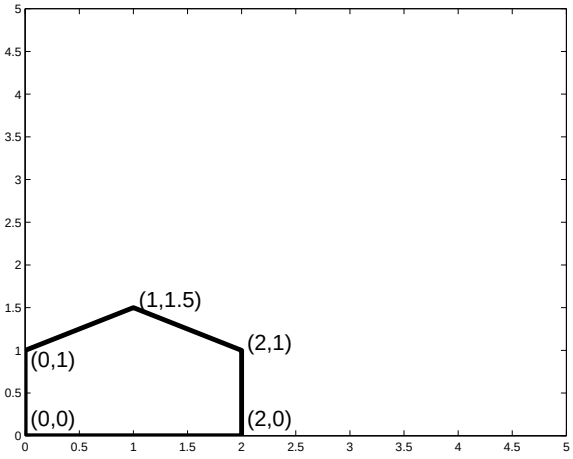


- $T = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$. In general, matrix multiplication will have the effect of changing both the length and orientation of the vector:

$$T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



Let's now consider how matrix multiplication affects sets of vectors. The following plot shows 5 vectors plotted in a coordinate space. I also connect the points with lines to form a pentagonal shape:



Suppose we multiply each of the vectors by the matrix $T = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. We perform the multiplications below:

$$T \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

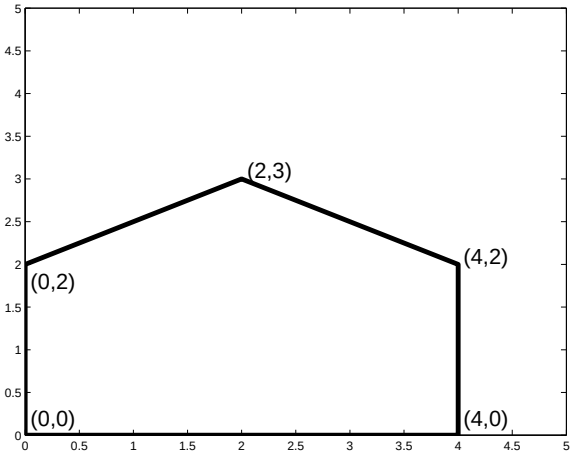
$$T \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

$$T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

$$T \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

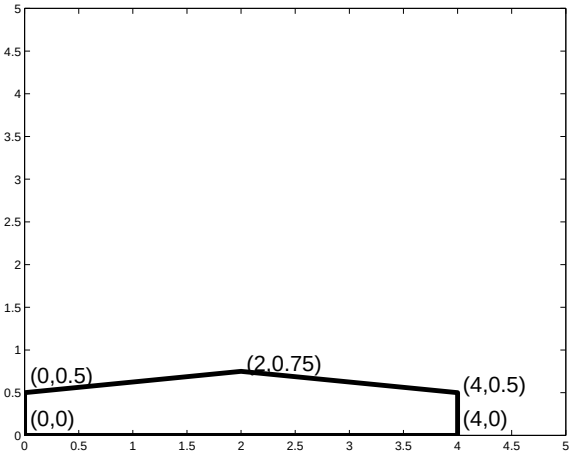
$$T \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

The "transformed" vectors are plotted below. We notice that multiplying by the matrix T has the effect of stretching the shape by a factor of two in each coordinate direction.



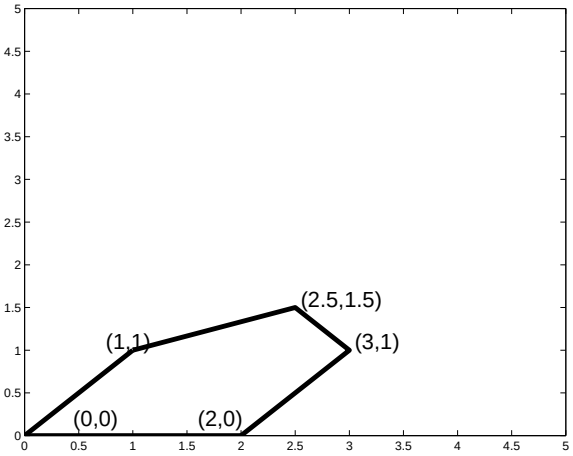
We repeat the exercise for various choices of T. Note the effect of the matrix multiplication on the size and orientation of the shape. For the following choice of T, for instance, the shape is stretched in one coordinate direction and shrunk in the other:

$$T = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix} \longrightarrow$$



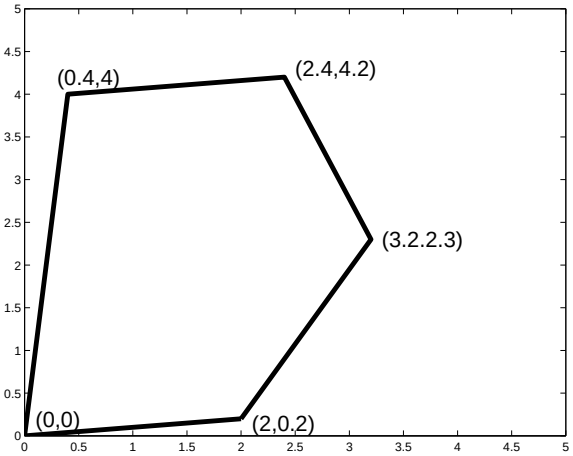
Multiplication by this T tilts the shape:

$$T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \longrightarrow$$



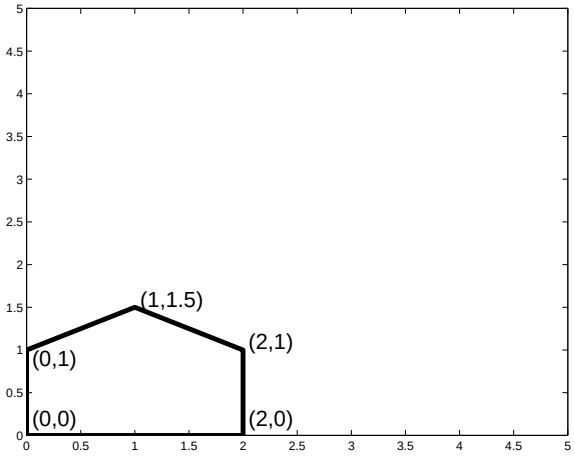
Multiplication by this T has the effect of stretching, tilting, and flipping the shape:

$$T = \begin{bmatrix} 0.2 & 2 \\ 2 & 0.2 \end{bmatrix} \longrightarrow$$



Note that this final choice of T has no effect on the size or orientation of the shape. This is the only 2×2 matrix with this property, and is a special matrix called the identity matrix. This is the subject of the next section.

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \longrightarrow$$



2.4 Identity Matrix and Matrix Inverses

A square matrix which has ones on its diagonal and zeros elsewhere is called an **identity** matrix and is typically designated I. For example, the 3×3 identity matrix is as follows:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Identity matrices have the special property that any matrix times the identity matrix is the matrix itself. That is for any matrix M and the appropriate identity matrix I, we have $MI = M$. Note that for the special case of the 1×1 identity matrix, [1], this property simply reduces to the simple fact that any scalar times 1 gives the scalar itself.

Division is not defined with respect to matrices. However, for some square matrices there is a notion of a matrix inverse. The **inverse** of a matrix M is denoted M^{-1} and has the property that $MM^{-1} = I$. The details of computing matrix inverses is beyond the scope of this document (For details, see the references given at the end of the document). Only square matrices can have inverses, and inverses do not even exist for all square matrices. A square matrix that has an inverse is called an **invertible** matrix.

The matrix $A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$ used in earlier sections has the unusual property that it is its own inverse. One can easily verify using matrix multiplication that $AA = I$. Thus A is invertible and $A^{-1} = A$. (This is not typically the case!)

Note briefly the analogy with the scalar case. Recall that the inverse of a scalar x is simply $1/x$. If we multiply a scalar x times its inverse $1/x$, we get the scalar identity, which is 1. Thus the notions of matrix identity and matrix inverses is a generalization of well-known notions in arithmetic.

3 Eigenvalues and Eigenvectors

A somewhat more advanced topic related to matrices is that of eigenvalues and eigenvectors of a square matrix. The definition is as follows: suppose we have a square matrix M, and that we can find a vector $\mathbf{x}$ and a scalar λ such that the following equation is true:

$$M\mathbf{x} = \lambda\mathbf{x}$$

Then the vector $\mathbf{x}$ is called an **eigenvector** of the matrix M and the scalar λ is the corresponding **eigenvalue**. We will not explain how to compute eigenvalues and eigenvectors here (see the references for details), but we will briefly discuss some properties of them.

As an example, consider the simple matrix $F = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$. First, see that the vector $\mathbf{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and the scalar $\lambda = 7$ are an eigenvector and an eigenvalue for F because:

$$\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 7 \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In fact, the zero vector is an eigenvector for every square matrix. Furthermore, any choice of scalar is a corresponding eigenvalue.

We can also find more interesting eigenvector/eigenvalue pairs. The vector $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and the scalar $\lambda = 5$ are an eigenvector and eigenvalue for F because:

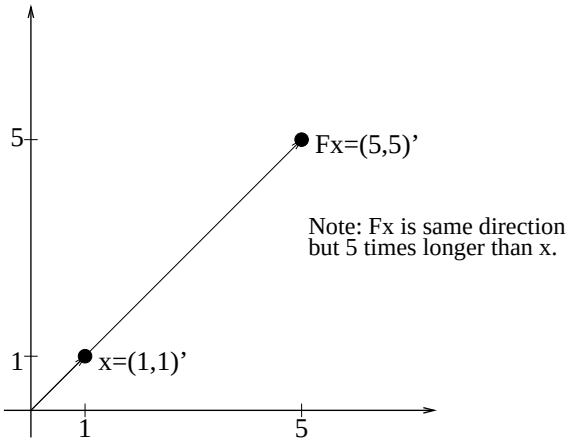
$$\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

We can also derive another eigenvector and eigenvalue by multiplying the previous ones by 2. That is, we can easily show that $\mathbf{x} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\lambda = 10$ are also an eigenvector/eigenvalue pair for F. In general, if we have an eigenvector/eigenvalue pair $\mathbf{x}$ and λ for a matrix, then any multiple of $\mathbf{x}$ and λ will also be an eigenvector/eigenvalue pair.

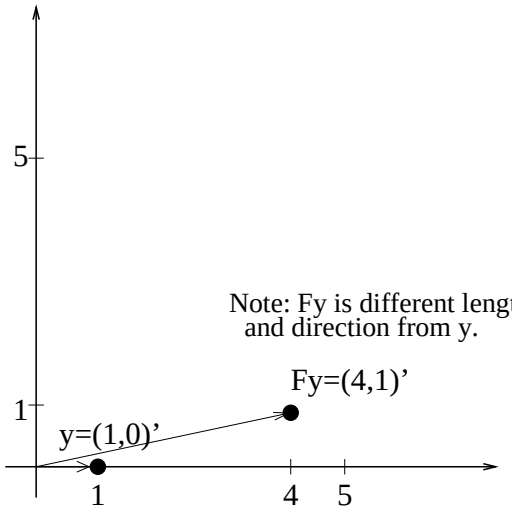
There is one more set of eigenvectors and eigenvalues of F that are not multiples of $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\lambda = 5$. These are the eigenvector $\mathbf{x} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and the eigenvalue $\lambda = 3$ (and all multiples of these). How many eigenvectors and eigenvalues does a matrix have? In general, a $n \times n$ square matrix has at most n linear independent eigenvector/eigenvalue pairs (not counting the zero eigenvector), meaning that there are at most n eigenvectors that cannot be formed by adding and scalar multiplying the other eigenvectors.

What is the intuition behind the eigenvectors and eigenvalues? As we did in section 2.3.3, let's think of matrix multiplication as a transformation of vectors. We saw in the plots of section 2.3.3 that in general, when we multiply a matrix times a vector the vector is transformed. If we think of vectors as arrows pointing away from the origin point $(0, 0)$, then this transformation will in general change the length and direction of the vector. Eigenvectors are special vectors whose length changes, but who remain pointing in the same direction. The eigenvalue is the factor by which the length changes in the process of the matrix multiplication.

To illustrate, consider again the matrix $F = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$, and recall that $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\lambda = 5$ are an eigenvector and eigenvalue for F. Let's examine the plot of $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $F\mathbf{x} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$ and note the relationship between $\mathbf{x}$ and $F\mathbf{x}$:



On the other hand, the vector $\mathbf{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is not an eigenvector of F. Note that $\mathbf{y}$ and $F\mathbf{y}$ in the plot below do not exhibit this special relationship:



Eigenvectors and eigenvalues serve as useful summaries of matrices, and there is a deep theory of eigenvectors and eigenvalues in linear algebra. For a more detailed explanation of eigenvalues and eigenvalues and their properties, please see the references.

4 Covariance Matrices and Positive Definiteness

Recall that if we have two discrete random variables X and Y with means μ_X and μ_Y, we define covariance of X and Y as:

$$cov(X, Y) = \sum_{i=1}^{n} p_i(x_i - \mu_X)(y_i - \mu_Y)$$

We define the **variance-covariance** matrix (often called just the **covariance** matrix) of X and Y as follows:

$$\begin{bmatrix} var(X) & cov(X,Y) \\ cov(Y,X) & var(Y) \end{bmatrix}$$

The covariance matrix is clearly square, and since $cov(X,Y) = cov(Y,X)$, we have that the covariance matrix is also symmetric. A property that is less obvious is that the covariance matrix has all positive eigenvalues. A square symmetric matrix with this property is refered to as a **positive definite** matrix.

We can also define the so-called **correlation** matrix of X and Y as follows:

$$\begin{bmatrix} 1 & corr(X,Y) \\ corr(Y,X) & 1 \end{bmatrix}$$

Correlation matrices are also square, symmetric, and positive definite.

We can also generalize the notion of the covariance and correlation matrices to the case with more than two random variables. These matrices will also be square, symmetric, and positive definite. As an example, the following is the covariance matrix for three random variables X, Y, and Z:

$$\begin{bmatrix} var(X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & var(Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & var(Z) \end{bmatrix}$$

5 Concepts of Distance

Distance is a key measure of proximity in data mining and in classical statistics. In classification, if observation $\boldsymbol{x}$ is "close" to $\boldsymbol{z}$, one may assume that those two points belong to the same class. However, how do we measure closeness or distance?

Before elaborating on the different distance measures, let's review some basic properties of distances, including symmetry, distinguishability and triangular inequality.

1. *Symmetry*: Given two observations, $\boldsymbol{x}$ and $\boldsymbol{z}$, the distance $D(\boldsymbol{x}, \boldsymbol{z})$, between the observations $\boldsymbol{x}$ and $\boldsymbol{z}$ is equal to the distance between observations $\boldsymbol{z}$ and $\boldsymbol{x}$ and should be greater than zero. i.e, $D(\boldsymbol{x}, \boldsymbol{z}) = D(\boldsymbol{z}, \boldsymbol{x}) \geq 0$.

2. *Distinguishability*: Given two observations, $\boldsymbol{x}$ and $\boldsymbol{z}$, if $D(\boldsymbol{x}, \boldsymbol{z}) = 0$, then $\boldsymbol{x}$ and $\boldsymbol{z}$ are the same observations . If $D(\boldsymbol{x}, \boldsymbol{z}) \neq 0$, then $\boldsymbol{x}$ and $\boldsymbol{z}$ are not the same.

3. *Triangular Inequality*: Given three observations, $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{w}$, $D(\boldsymbol{x}, \boldsymbol{z}) \leq D(\boldsymbol{x}, \boldsymbol{w}) + D(\boldsymbol{w}, \boldsymbol{z})$. This property says that the lengths of any given side of a triangle is less than the sum of the lengths of the remaining two sides. Note: sometimes we may relax this condition.

A simple and common measure of distance is the *Euclidean distance*. Suppose vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ are d-dimensional vectors, where x_i and z_i is the i^{th} element of vector $\boldsymbol{x}$ and $\boldsymbol{z}$, respectively. Then the Euclidean distance is defined as follows:

$$ED(\boldsymbol{x}, \boldsymbol{z}) = \sqrt{\sum_{i=1}^{d} (x_i - z_i)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{z})'(\boldsymbol{x} - \boldsymbol{z})} \tag{1}$$

Often, we may simply deal with the squared Euclidean distance ($ED(\boldsymbol{x}, \boldsymbol{z})^2 = \sum_{i=1}^{d} (x_i - z_i)^2$) for notational simplicity (note that the square root function in monotonic with its argument, thus $ED(\boldsymbol{x}, \boldsymbol{z}) > ED(\boldsymbol{x}, \boldsymbol{w})$ implies $ED(\boldsymbol{x}, \boldsymbol{z})^2 > ED(\boldsymbol{x}, \boldsymbol{w})^2$. Since we often care only about the relative pairwise distances, it would not matter if we used distances versus squared distances).

However, Euclidean distance is not independent of scale. For example, suppose the first dimension is the "height" on an individual and the second dimension is "weight". The relative distances between

individuals, $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{w}$, may change if we use "feet" to measure height instead of "inches", or "pounds" versus "kilograms" for weight. One set of units may say person $\boldsymbol{x}$ is closer (more similar) to person $\boldsymbol{z}$ than to person $\boldsymbol{w}$, but different units of measurements may change this relation (although the physical characteristics of each individuals clearly remain the same).

We deal with this problem by scaling the variables with is standard deviation. Suppose σ_j is the standard deviation of dimension j. Then the squared *statistical distance* between points $\boldsymbol{x}$ and $\boldsymbol{z}$ is as follows:

$$SD(\boldsymbol{x}, \boldsymbol{z})^2 = \sum_{i=1}^{d} \left(\frac{x_i - z_i}{\sigma_j} \right)^2 \tag{2}$$

The statistical distance is the same as taking the Euclidean distance of data after the data is standardized (each dimension is divided by its standard deviation). Thus, whether height is measured with inches or feet, the statistical distances should be the same. Thus, it is invariant with respect to a change in scale.

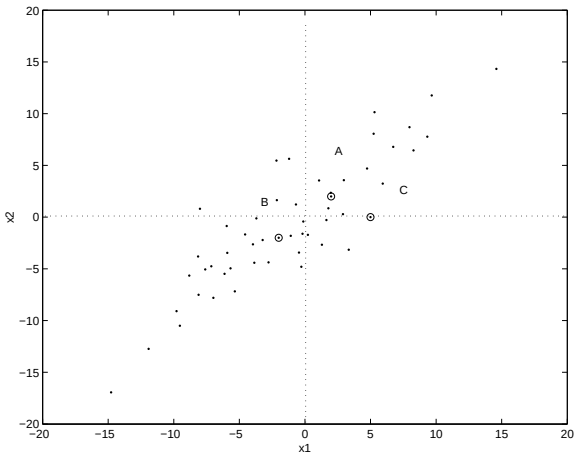
However, statistical distances still do not take into account correlation between the variables. For example, Figure 1 illustrates a data set where there exist a positive correlation between the two variables. It seems apparent that we need to consider this correlation when measuring "closeness" between points. *Mahalanobis distance* is a measure of distance that does just that. The squared Mahalanobis distance is defined as follows:

$$MD(\boldsymbol{x}, \boldsymbol{z})^2 = (\boldsymbol{x} - \boldsymbol{z})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{z}) \tag{3}$$

where Σ is the d by d dimensional covariance matrix of the variables. Clearly, if there are no correlation between the variables (i.e., Σ is a diagonal matrix), then the Mahalanobis distance is the same as the statistical distance. Further, if the variables are standardized (i.e. Σ is the identity matrix), then the Mahalanobis distance is the same as the Euclidean distance.

Figure 1 illustrates the advantage of Mahalanobis distance. It plots 53 sample points from a data set with mean (0,0) and the covariance matrix $\Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ (i.e, the variance of variable 1 is 5, the variance of variable 2 is 5 and the covariances of variables 1 and 2 is equal to 3).

Figure 1:



The three points, A, B, and C, have the following coordinates: A = (2,2), B = (-2,-2), C = (0, 5). Without doing any calculations, which of points B and C seem "closer" to A from the picture? The following table illustrates the pairwise distances:

Both Euclidean and statistical distances imply that point C is closer to A than B. However, when considering the positive covariance between variable 1 and 2, point C seems further away from A than B, with respect to the other data points. We used this concept in discriminant analysis and will revisit it in clustering analysis, where distance measure play a central role.

Table 1:

Pair of points	ED^2	SD^2	MD^2
A and B	32	6.4	4
A and C	13	2.6	6.3125

References

[1] Steven J. Leon. *Linear Algebra with Applications*. New York: Macmillan Publishing Company, 1990.

[2] Thomas P. Ryan. *Modern Regression Methods*. New York: John Wiley & Sons, 1997, pp. 101-106.

[3] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley, MA: Cambridge Press, 1998.