Vectors, Matrices, and Data Mining

Yuzhen Ye

1 Vectors

1.1 What's vector

A vector is a quantity that has magnitude and direction. Vectors are often represented by lowercase bold letters (such as **u** and **v**) or italic lowercase letters (u and v). A point in Euclidean space can be represented by a vector from the origin to the point.

1.2 Vector addition and multiplication by a scalar

Vectors can be added and subtracted, as shown in Figure 1.

- Commutativity of vector addition: $u + v = v + u$.
- Association of vector addition: $(u + v) + w = u + (v + w)$.
- Existence of a zero vector: $u + 0 = u$.
- Existence of additive inverses for vector addition. $u + (-u) = 0$.

Scalar multiplication changes the magnitude of a vector, but the direction is unchanged if the scalar is positive and is reversed if the scale is negative. Assume u and v are vectors, and α and β are scalars,

- Associativity of scalar multiplication: $\alpha(\beta u) = (\alpha\beta)u$.
- Distributivity of scalar addition over multiplication of a scalar by a vector: $(\alpha + \beta)u = \alpha u + \beta u$.
- Distributivity of scalar multiplication over vector addition: $\alpha(u + v) = \alpha u + \beta v$.
- Existence of scalar identity. If $\alpha = 1$, then $\alpha u = u$.

1.3 Vector Spaces

A vector space is a set of vectors, along with an associated set of scalars that satisfies the above properties and that is closed under vector addition and multiplication by a scalar.

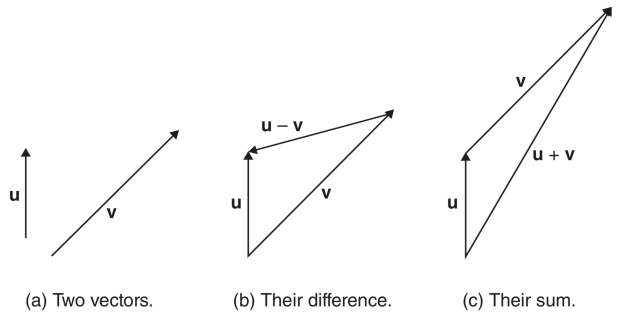

Figure 1: Vectors, their sum and difference. (Figure from Tan)

1.4 Vector dot product

The dot product of two vectors $\boldsymbol{u}$ and v: $\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i$.

For instance, $(3,4) \cdot (2,5) = 3 \times 2 + 4 \times 5 = 26$.

The dot product of two non-zero vectors is 0 if and only if they are perpendicular (i.e., the two vectors are orthogonal).

1.5 Vector in DM

An item, or an attribute can be represented as a vector. Items are row vectors, and attributes are column vectors.

The length of a vector in Euclidean space can be computed using the dot product: $length(\mathbf{u}) = \sqrt{\mathbf{u} \cdot \mathbf{u}}$. The length of the vector is also known as its $L2$ norm ($\|\mathbf{u}\|$). A vector can be normalized to have an $L2$ norm of 1 as following: $\mathbf{u}/\|\mathbf{u}\|$.

The dot product of two vectors can be written as: $\mathrm{u} \cdot \mathrm{v} = \|\mathrm{u}\|\|\mathrm{v}\| \cos(\theta)$, where θ is the angle between the two vectors.

Cosine similarity between two items can be computed using dot product: $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}}{\|\mathbf{u}\|} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|}$. This similarity doesn't consider the length (magnitude) of the vectors, but is only concerned with the degree to which two vectors point in the same direction. In terms of documents, their cos similarity will be 1 if they contain the same terms in the same proportion (and terms that don't appear in both documents play no role in computing similarity).

The Euclidean distance between two vectors is: $d(\mathbf{u}, \mathbf{v}) = \sqrt{(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})}$.

2 Matrices

2.1 Basics

A matrix is a tabular representation of a set of numbers as a collection of rows and columns. Matrices are often represented by uppercase bold letters (e.g., **A**, **M**) or uppercase italic letters (e.g., A, M). A table with m rows and n columns is a "m by n matrix". A matrix with the same number of rows and columns ($m = n$) is a square matrix. The transpose of a matrix A is written as A^T and is produced by interchanging the rows and columns of A. The entry in the i^{th} row and j^{th} column is a_{ij}, and each row or column defines a vector: $\mathbf{a}_{i*}$ is row i, and $\mathbf{a}_{*j}$ is column j.

2.2 Matrix operations

Addition of matrices (of the same dimensions) is done by adding individual entries. Assume $C = A + B$, $c_{ij} = a_{ij} + b_{ij}$. Matrix addition has the following properties:

- Commutativity of matrix addition: $A + B = B + A$.
- Associativity of matrix addition: $(A + B) + C = A + (B + C)$.
- Existence of an identity element for matrix addition: $A + 0 = A$.
- Existence of additive inverses for matrix addition, $A + (-A) = 0$.

The product of a scalar α and a matrix A is the matrix $B = \alpha A$, where $b_{ij} = \alpha a_{ij}$. Scalar multiplication has the following properties.

- Associativity of scalar multiplication. $\alpha(\beta A) = (\alpha\beta)A$.
- Distribution of scalar addition over multiplication of a scalar by a matrix. $(\alpha + \beta)A = \alpha A + \beta A$.
- Distribution of scalar multiplication over matrix addition. $\alpha(A + B) = \alpha A + \beta A$.

Matrix multiplication is done as following (assuming $C = AB$): $c_{ij} = \mathbf{a}_{i*} \cdot \mathbf{b}_{*j}^T$, where A is m by n, B is n by p, and C is m by p. In other words, the ij_{th} entry of matrix C is the dot product of the i_{th} row vector of A and j_{th} column vector of B. For example, $\begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 2 & 9 \\ 8 & 1 \end{bmatrix} = \begin{bmatrix} 33 & 23 \\ 25 & 24 \end{bmatrix}$.

Matrix multiplication has the following properties:

- Associativity of matrix multiplication. $(AB)C = A(BC)$.
- Distribution of matrix multiplication. $A(B + C) = AB + AC$.

- Existence of an identity element for matrix multiplication. If 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 A, $AI_n = A$ and $I_m A = A$.

In generally, matrix multiplication is not commutative, i.e., $AB \neq BA$.

2.3 Determinants

In linear algebra, the determinant is a scalar value that can be computed from the elements of a square matrix and encodes certain properties of the linear transformation (see below) described by the matrix. Geometrically, it can be viewed as the volume scaling factor of the linear transformation described by the matrix. The determinant is positive or negative according to whether the linear mapping preserves or reverses the orientation of n-space.

The determinant of matrix A is represented as $det(A)$, det A, or $|A|$.

For a matrix of 2 by 2, $|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$.

For a matrix of 3 by 3,

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh.$$

The determinant of an identity matrix is 1: $\det(I_n) = 1$.

The determinant of a matrix of arbitrary size can be defined by the Leibniz formula or the Laplace formula. The Leibniz formula for the determinant of an n by n matrix A is

$$\det(A) = \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i} \right).$$

Here the sum is computed over all permutations of the set 1, 2, ..., n. A permutation is a function that reorders this set of integers. The value in the ith position after the reordering is denoted by i. For example, for n = 3, the original sequence 1, 2, 3 might be reordered to = [2, 3, 1], with 1 = 2, 2 = 3, and 3 = 1. The set of all such permutations (also known as the symmetric group on n elements) is denoted by Sn. For each permutation , sgn() denotes the signature of , a value that is $+1$ whenever the reordering given by can be achieved by successively interchanging two entries an even number of times, and 1 whenever it can be achieved by an odd number of such interchanges.

Laplace's formula expresses the determinant of a matrix in terms of its minors (an iterative approach). The minor Mi,j is defined to be the determinant of the $(n1) \times (n1)$-matrix that results from A by removing the i-th row and the j-th column. The expression $(1)^{i+j} M_{i,j}$ is known as a cofactor. The determinant of A is given by

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \text{(for a fixed } i\text{)} = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \text{(for a fixed } j\text{)}.$$

Let A be an arbitrary nn matrix of complex numbers with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then the determinant of A is the product of all eigenvalues,

$$\det(A) = \prod_{i=1}^{n} \lambda_i = \lambda_1 \lambda_2 \cdots \lambda_n.$$

The product of all non-zero eigenvalues is referred to as pseudo-determinant.

Conversely, determinants can be used to find the eigenvalues of the matrix A: they are the solutions of the characteristic equation, $\det(A - \lambda I) = 0$, where I is the identity matrix of the same dimension as A and λ is a (scalar) number which solves the equation (there are no more than n solutions, where n is the dimension of A).

2.4 Geometric meaning of a matrix

If an n by n real matrix A is written in terms of its column vectors $A = \left[\begin{array}{c|c|c|c} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array} \right]$, then

$$A \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{a}_1, \quad A \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{a}_2, \quad \ldots, \quad A \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{a}_n.$$

This means that A maps the unit n-cube to the n-dimensional parallelotope defined by the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, the region $P = \left\{ c_1 \mathbf{a}_1 + \cdots + c_n \mathbf{a}_n \mid 0 \leq c_i \leq 1 \ \forall i \right\}$.

The determinant gives the signed n-dimensional volume of this parallelotope, $\det(A) = \pm \mathrm{vol}(P)$, $\det(A) = \pm \mathrm{vol}(P)$, and hence describes more generally the n-dimensional volume scaling factor of the linear transformation produced by A. The sign shows whether the transformation preserves or reverses orientation. In particular, if the determinant is zero, then this parallelotope has volume zero and is not fully n-dimensional, which indicates that the dimension of the image of A is less than n. This means that A produces a linear transformation which is neither onto nor one-to-one, and so is not invertible.

2.5 Rank of matrix

The maximum number of linearly independent vectors in a matrix is equal to the number of non-zero rows in its row echelon matrix. Therefore, to find the rank of a matrix, we simply transform the matrix to its row echelon form and count the number of non-zero rows.

2.6 Echelon form

A matrix is in row echelon form (ref) when it satisfies the following conditions:

- The first non-zero element in each row (called the leading entry) is 1.
- Each leading entry is in a column to the right of the leading entry in the previous row.

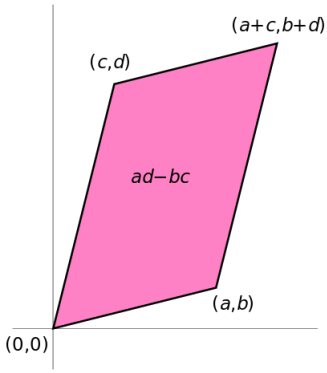

Figure 2: The area of the parallelogram is the absolute value of the determinant of the matrix formed by the vectors representing the parallelogram's sides. (Figure from wikipedia)

- Rows with all zero elements, if any, are below rows having a non-zero element.

A matrix is in reduced row echelon form (rref) if it satisfies the additional condition: The leading entry in each row is the *only* non-zero entry in its column.

A matrix can be changed into its echelon form using a series of elementary row operations (Gaussian elimination).

- Pivot the matrix
  - Find the pivot, the first non-zero entry in the first column of the matrix.
  - Interchange rows, moving the pivot row to the first row.
  - Multiply each element in the pivot row by the inverse of the pivot, so the pivot equals 1.
  - Add multiples of the pivot row to each of the lower rows, so every element in the pivot column of the lower rows equals 0.
- To get the matrix in row echelon form, repeat the pivot
  - Repeat the procedure from Step 1 above, ignoring previous pivot rows.
  - Continue until there are no more pivots to be processed.
- To get the matrix in reduced row echelon form, process non-zero entries above each pivot.

- Identify the last row having a pivot equal to 1, and let this be the pivot row.
- Add multiples of the pivot row to each of the upper rows, until every element above the pivot equals 0.

3 Linear Transformation

3.1 Fundamentals

We say we have a linear transformation $T\colon \mathbb{R}^n \to \mathbb{R}^m$, we have $T(\vec{x}) = A\vec{x}$ where A is a $n \times m$ matrix.

3.2 Examples

In two-dimensional space $\mathbb{R}^2$ linear maps are described by 2×2 real matrices. These are some examples:

1) rotation by 90 degrees counterclockwise: $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

2) reflection against the x axis: $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$

3) scaling by 2 in all directions: $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

3.3 Design transformation matrix

Start from an identity matrix $I_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$, the columns in this matrix can be called e_1, e_2, $\dots$, and e_n. Then the transformation T can be applied to individual columns: $A = [T(e_1), T(e_2), \dots, T(e_n)]$ (i.e., apply operations to each of the basis).

Assume we have a triangle in $\mathbb{R}^2$ composed of three vectors $\begin{bmatrix} -3 \\ 2 \end{bmatrix}$, $\begin{bmatrix} 3 \\ -2 \end{bmatrix}$, $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$ (see Figure 1). Here is the transformation: reflect around y, and stretch in

y direction by two times. We can write down the transformation as $T(\begin{bmatrix} x \\ y \end{bmatrix}) = \begin{bmatrix} -x \\ 2y \end{bmatrix}$. For example, after the transformation $\begin{bmatrix} -3 \\ 2 \end{bmatrix}$ becomes $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$.

How we define the matrix A for this transformation? We start with an identify matrix: $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and apply the transformation to each of its columns: $A = \begin{bmatrix} T(\begin{bmatrix} 1 \\ 0 \end{bmatrix}) T(\begin{bmatrix} 0 \\ 1 \end{bmatrix}) \end{bmatrix}$. So we have $A = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$.

Let's verify $T(\begin{bmatrix} x \\ y \end{bmatrix}) = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ using vector $\begin{bmatrix} -3 \\ 2 \end{bmatrix}$. $T(\begin{bmatrix} -3 \\ 2 \end{bmatrix}) = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} -3 \\ 2 \end{bmatrix} = \begin{bmatrix} (-1) \times (-3) + 0 \times 2 \\ 0 \times (-3) + 2 \times 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, which is what we get by applying the transformation directly to the vector.

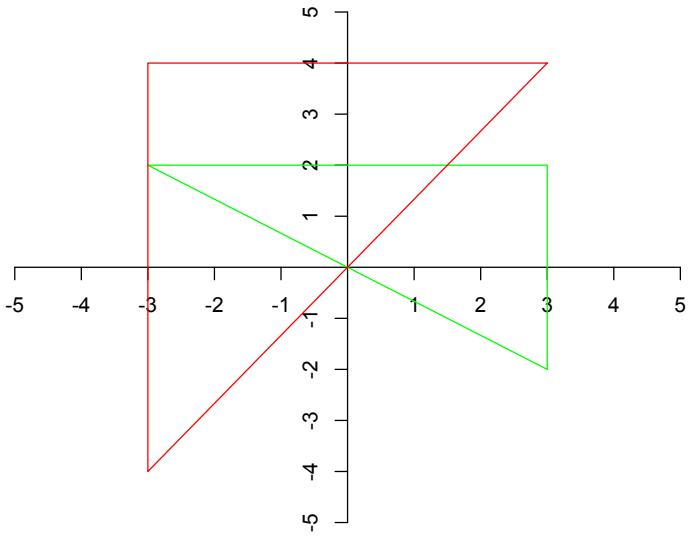

Figure 3: Example of a linear transformation. The original triangle is shown in green, and the transformed triangle is shown in red.

3.4 Linear Sketches

A linear sketch is a random linear projection: M: $\mathbb{R}^n \to \mathbb{R}^k$ (where $k << n$) that preserves properties of any $v \in \mathbb{R}^n$ with high probability.

Count-Min Sketch: In each column, place a 1 in a random row.

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \\ v5 \\ v6 \\ v7 \\ v8 \end{bmatrix} = \begin{bmatrix} v3+v4 \\ v1+v5+v7 \\ v2+v6+v8 \end{bmatrix}.$$

This example uses v2+v6+v8 as estimate for v2 (also v6 and v8).

Count-Sketch: Like Count-Min but non-zero entries $\in R-1, 1$:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \\ v5 \\ v6 \\ v7 \\ v8 \end{bmatrix} = \begin{bmatrix} v3-v4 \\ -v1+v5-v7 \\ v2-v6+v8 \end{bmatrix}.$$

This example uses v2-v6+v8 as estimate for v2.

4 Link Analysis and PageRank

4.1 Power iteration approach to PageRank

Given a graph of webpages, a transition matrix **M** can be derived. The transition matrix for the toy Web shown in Figure 1:

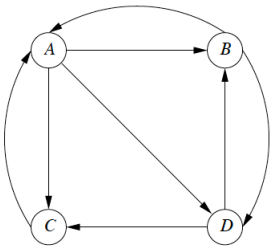

Figure 4: A toy example of Web.

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}.$$

In this matrix, the first column expresses that a surfer at page A has a $1/3$ probability of next being at page B, C, and D (but not A); the second column shows the probability of surfer starting at B and ending at page A ($1/2$), B (0), C (0) and D ($1/2$). Note, in this case, the columns each add to one (i.e., M is *stochastic*).

Suppose a surfer starts at any of the pages of the Web with equal probability. Denote the probabilities as a vector $\mathbf{v} = [1/4, 1/4, 1/4, 1/4]^T$. What's the probability of the surfer ending at page A, B, C and D after one step of crawling? The distribution of the surfer will be $M\mathbf{v}$.

$$M\mathbf{v} = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

We see that the probability of surfer being at page A is $9/24$, which is $1/4 * 0 + 1/4 * 1/2 + 1/4 * 1 + 1/4 * 0 = 9/24$.

The distribution of the surfer approaches a limiting distribution $\mathbf{v}$ that satisfies $\mathbf{v} = M\mathbf{v}$, provided two conditions are met: the graph is strongly connected (i.e., it is possible to get from any node to any other node); and there is no dead ends (i.e., nodes that have no incoming edges). In fact, the limiting $\mathbf{v}$ is an eigenvector of M (an eigenvector of a matrix M is a vector $\mathbf{v}$ that satisfies $\mathbf{v} = \lambda M\mathbf{v}$). And because M is stochastic, $\mathbf{v}$ is the principal eigenvector (its associated eigenvalue is the largest of all eigenvalues, which is 1, again as M is stochastic).

For the above example, the eigenvector $\mathbf{v} = \begin{bmatrix} 3/9 & 2/9 & 2/9 & 2/9 \end{bmatrix}^T$.

5 Eigenvalues and Eigenvectors of matrices

5.1 Definition

Let M be a square matrix. Let λ be a constant and $\mathbf{e}$ a nonzero column vector with the same number of row as as M. Then λ is an eigenvalue of M and $\mathbf{e}$ is the corresponding eigenvector of M if $M\mathbf{e} = \lambda\mathbf{e}$. If $\mathbf{e}$ is an eigenvector and c is a constant, then $c\mathbf{e}$ is also an eigenvector of M. To avoid ambiguity, it is required that every eigenvector be a unit vector (of length 1), and its first component be nonzero (so no direction change).

For example, if $M = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$, one of its eigenvector is $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$ (with length of 1) and its corresponding eigenvalue is 7.

5.2 Computing eigenvalues and eigenvectors using determinants

In the PageRank algorithm, the importance of the web pages (the stationary distribution) can be computed using the power iteration approach: staring with any unit vector v (of the importance distribution) and compute $M^i\mathbf{v}$ (M is the transition matrix) iteratively until it converges. As M is a stochastic matrix, the limiting vector is the principal eigenvector (the eigenvector with the largest eigenvalue) and its corresponding eigenvalue is 1. This approach can be generalized to find all pairs of eigenvalue and eigenvector (O(n^3) algorithm).

Eigenpairs are defined by $M\mathbf{e} = \lambda\mathbf{e}$, i.e., $(M - \lambda I)\mathbf{e} = \mathbf{0}$. In order for $(M - \lambda I)\mathbf{e} = \mathbf{0}$ to hold for a vector $\mathbf{e} \neq \mathbf{0}$, the determinant of $M - \lambda I$ must be 0. The determinant of $(M - \lambda I)$, $det(M - \lambda I)$, is an nth-degree polynominal in λ, from which we can get the n values of λ (the roots of the characteristic equation $det(M - \lambda I) = 0$), that are the eigenvalues of M, and then for any such value, say c, we can solve the equation $M\mathbf{e} = c\mathbf{e}$ to derive the corresponding eigenvector.

For example, to compute the eigenpairs for $M = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$, we first get $M - \lambda I$, which is $\begin{bmatrix} 3-\lambda & 2 \\ 2 & 6-\lambda \end{bmatrix}$. Its determinant is $\lambda^2 - 9\lambda + 14$, which needs to be set to 0. This gives $\lambda = 7$ and $\lambda = 2$, the two eigenvalues. Let's use 7 as the example to show how to find its corresponding eigenvector.

$$\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 7 \begin{bmatrix} x \\ y \end{bmatrix}.$$

We have two equations: $3x + 2y = 7x$ and $2x + 6y = 7y$. Because there are no constant terms in any of these equations (basically the two equations say the same thing, $y = 2x$), there are no unique solutions. One possible solution is $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$, which is not a unit vector. This vector can be normalized by the length of the vector (which is $\sqrt{5}$) to derive the principal eigenvector, $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$. The second principal eigenvector can be computed using the same approach, and it is $\begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$

5.3 Computing eigenvalues and eigenvectors using power iteration

In the PageRank algorithm, the M is stochastic and the power iteraction approach results in eigenvector of length 1. To make the power iteration work

for general matrices, some modification needs to be made: $x_{k+1} := \frac{Mx_k}{\|Mxk\|}$ (instead of $x_{k+1} := Mx_k$), where $\|N\|$ is the *Frobenius norm*, i.e., the square root of the sum of the squares of the elements of N. Once $\mathbf{x}$ converges, it is the principal eigenvector of M. As $M\mathbf{x} = \lambda\mathbf{x}$, we can compute the eigenvalue $\lambda = \mathbf{x}^T M\mathbf{x}$.

Assume $M = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$, and $\mathbf{x}_0$ is set to $[1,1]^T$. After one iteration, we get $\begin{bmatrix} 5 \\ 8 \end{bmatrix}$. By normalizing this vector by its Frobenius norm (which is 9.434), we get $\mathbf{x}_1 = \begin{bmatrix} 0.530 \\ 0.848 \end{bmatrix}$. This process can be iterated until $\mathbf{x}$ converges to $\begin{bmatrix} 0.447 \\ 0.894 \end{bmatrix}$, which is the principal vector. $\lambda = \mathbf{x}^T M\mathbf{x} = 6.993$. This value is slightly different from the eigenvalue (7) computed using the determinant approach shown above. In this case, power iteration introduced small errors due to limited precision (in other cases, the errors could be introduced by early stopping of the iteration). To find the second eigenpair, we create a new matrix $M^* = M - \lambda_1 \mathbf{x}\mathbf{x}^T$, where λ and $\mathbf{x}$ are the eigenpair with the largetst eigenvalue. We can find the second eigenpair by processing this matrix as we did the original matrix M.

5.4 Matrix of eigenvectors

Suppose we have an n by n matrix M whose eigenvectors are $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$. Let E be the matrix whose ith column is $\mathbf{e}_i$. Then $EE^T = E^T E = I$. Why? The eigenvectors of a symmetric matrix are orthogonal unit vectors.

Any matrix of orthonormal vectors (unit vectors that are orthogonal to one another) represents a rotation and/or reflection of the axes of a Euclidean space.

6 Principal-component analysis

PCA is a technique for taking a dataset consisting of a set of points (D) in a high-dimensional space and finding the directions along which the data points (vectors) line up best. The main idea is to find the eigenvectors of $D^T D$ or DD^T. The matrix of the eigenvectors can be considered as a rigid rotation in a high-dimensional space. When this transformation is applied to the original data, the axis corresponding to the principal eigenvector is the one along which the points are most "spread out", i.e., the variance of the data is maximized.

Given an m by n data matrix D, whose m rows are data points, and n columns are the attributes, the covariance matrix of D is the matrix S, which has entries s_{ij} of the data, $s_{ij} = covariance(d_{*i}, d_{*j})$, i.e., s_{ij} is the covariance of the i^{th} and j^{th} attributes (columns) of the data. The covariance of two attributes

is a measure of how strongly the attributes vary together. If the data matrix D is preprocessed so that the mean of each attribute is 0, then $S = D^T D$.

A goal of PCA is to find a transformation of the data such that,

- Each pair of the new attributes has 0 covariance;
- The attributes are ordered with respect to how much of the variance of the data each attribute captures;
- The first attribute captures as much of the variance of the data as possible;
- Subject to the orthogonality requirement, each successive attribute captures as much of the remaining variance as possible.

A transformation of the data that has these properties can be obtained by using eigenvalue analysis of the covariance matrix S.

7 Singular-Value Decomposition (SVD)

PCA is equivalent to an SVD analysis of the data matrix, once the mean of each variable has been removed. However, it is not always desirable to remove the mean from the data, especially if the data is relatively sparse.

An m by n matrix A can be written as, $\mathbf{M} = \sum_{i=1}^{rank(M)} = \alpha_i \mathbf{u}_i \mathbf{v}_i^T = \mathbf{U\Sigma V}^T$, where $\mathbf{U}$ is an $m \times r$ column-orthonormal matrix (each column is a unit vector and the dot product of any two columns is o), $\mathbf{V}$ is an $n \times r$ column-orthonormal matrix ($\mathbf{V}$ is used in its transposed form, meaning that the rows of $\mathbf{V}^T$ that are orthonormal), and $\mathbf{\Sigma}$ is a diagonal matrix.

7.1 Computing the SVD of a Matrix

$\mathbf{M}^T = (\mathbf{U\Sigma V}^T)^T = (\mathbf{V}^T)^T \mathbf{\Sigma}^T \mathbf{U}^T = \mathbf{V\Sigma}^T \mathbf{U}^T = \mathbf{V\Sigma U}^T$ (transposing has no effect for diagonal matrix).

Now, $\mathbf{M}^T\mathbf{M} = \mathbf{V\Sigma U}^T\mathbf{U\Sigma V}^T$. As $\mathbf{U}$ is orthonormal, $\mathbf{U}^T\mathbf{U}$ is an identity matrix. $\mathbf{M}^T\mathbf{M} = \mathbf{V\Sigma}^2\mathbf{V}^T$. Multiple both sides of this equation on the right by $\mathbf{V}$, we get $\mathbf{M}^T\mathbf{MV} = \mathbf{V\Sigma}^2\mathbf{V}^T\mathbf{V}$. Again $\mathbf{V}^T\mathbf{V}$ is an identity matrix. Thus, $\mathbf{M}^T\mathbf{MV} = \mathbf{V\Sigma}^2$. It says that $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{M}^T\mathbf{M}$ and $\mathbf{\Sigma}^2$ is the diagonal matrix whose entries are the corresponding eigenvalues.

Thus the same algorithm that computes the eigenpairs for $\mathbf{M}^T\mathbf{M}$ gives us the matrix $\mathbf{V}$ for the SVD of $\mathbf{M}$ itself ($\mathbf{V}$ represents the patterns among columns).

Using similar manipulations, we get $\mathbf{MM}^T\mathbf{U} = \mathbf{U\Sigma}^2$. That is $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{MM}^T$ ($\mathbf{U}$ represents the patterns in rows).

To compute the singular values for this SVD, we just need to take the square roots of the eigenvalues for $\mathbf{M}^T\mathbf{M}$ (or $\mathbf{M}\mathbf{M}^T$) (the eigenvalues of $\mathbf{M}^T\mathbf{M}$ are the eigenvalues of $\mathbf{M}\mathbf{M}^T$ plus addition 0's, if the dimension of $\mathbf{M}\mathbf{M}^T$ is larger; or the opposite would be true).