CS395T Large-Scale Data Mining
Lecture 3: Linear Algebra Review

scribed by Ramgopal Mettu

1 Notation

We will refer to the set of real numbers with the symbol $\mathbf{R}$. In general, lower case Greek letters such as α and β will be used to denote real numbers. The set of complex numbers will be denoted $\mathbf{C}$. The set of all d-dimensional real vectors will be denoted $\mathbf{R}^d$ and the set of all $m \times n$ real matrices will be denoted $\mathbf{R}^{m \times n}$. Vectors in $\mathbf{R}^d$ will be column vectors, e.g., for x in $\mathbf{R}^d$, x is written

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}.$$

The corresponding row vector x^T is written

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}.$$

2 Vectors and Norms

A vector is described by its "size" and its "direction". The *norm* of a vector x, written $\|x\|$ is a way to measure the size of a vector. Every norm must satisfy three properties: (i) for all vectors x, $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$; (ii) for all vectors x, $\|\alpha \cdot x\| = |\alpha| \cdot \|x\|$; and (iii) for all vectors x and y, $\|x + y\| \leq \|x\| + \|y\|$. $\|x\|_2$ denotes the 2-*norm* or the *Euclidean norm* of a vector x in $\mathbf{R}^d$ and is equal to $\sqrt{\sum_{i=1}^d x_i^2}$. Examples of other norms include the L_1 norm, where $\|x\|_1 = \sum_{i=1}^d |x_i|$ and the L_∞ norm, where $\|x\|_\infty = \max_{i:1 \leq i \leq d} x_i$. In general, the L_p norm is defined as $\sqrt[p]{\sum_{i=1}^d x_i^p}$. In the rest of this discussion, we will write $\|x\|$ to denote the 2-norm of x.

For x, y in $\mathbf{R}^d$, the *inner product* or *dot product* of x and y is $\sum_{i=1}^d x_i y_i$ and is written $\langle x, y \rangle$ or $x^T y$. An important property of the inner product is that it satisfies the *Cauchy-Schwartz Inequality*, i.e., for two vectors x and y

$$|x^T y| \leq \sqrt{x^T x} \cdot \sqrt{y^T y} = \|x\|_2 \|y\|_2.$$

We get a strict equality only if x and y are in the same "direction", that is, $y = \alpha \cdot x$. For the general case, let θ be the angle between x and y. Then

$$\cos\theta = \frac{x^T y}{\|x\| \|y\|}.$$

This implies that

$$x^T y = \|x\| \|y\| \cos \theta.$$

For non-zero x and y, when $x^T y = 0$, x and y are said to be *orthogonal.*

3 Linear Independence

A set of n non-zero vectors $y_1, y_2, \ldots, y_n$ is *linearly independent* if $\sum_{i=1}^{d} \alpha_i \cdot y_i = 0$ implies that $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$. The *span* of a set of vectors $y_1, y_2, \ldots, y_n$ in $\mathbf{R}^d$ is $\{\sum_{i=1}^{d} \alpha_i y_i \mid \alpha_i \in \mathbf{R}\}$, i.e., the set of all linear combinations of $y_1, y_2, \ldots, y_n$, and is denoted $span(y_1, y_2, \ldots, y_n)$. A *basis* is the largest set of linearly independent vectors in a given linear subspace. The *dimension* of a given linear subspace is the number of vectors in its basis.

4 Matrices, Eigenvalues and Eigenvectors

An $m \times n$ *matrix* is an array of numbers; it can also be viewed as a map from $\mathbf{R}^n$ to $\mathbf{R}^m$. That is, a matrix A in $\mathbf{R}^{m \times n}$ maps any n-dimensional vector x to an m-dimensional vector Ax. A matrix is *symmetric* if $A = A^T$. We will write the determinant of A as $det(A)$. A matrix is defined to be *singular* if it has no inverse; it turns out that a matrix A is singular if and only if $det(A) = 0$. Also, just as we measured the "size" of a vector using a norm, we can also measure the "size" of a matrix using a *matrix norm.* A matrix norm measures how much a matrix can "magnify" a vector. One common matrix norm is the 2-norm, where

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

As with vectors, we can generalize this to a matrix p-norm by replacing the vector 2-norms with vector p-norms. The 2-norm of a matrix is equal to the maximum singular value of A (see Section 5). For example, the identity matrix I has a 2-norm of 1 (since for all x, $Ix = x$).

A real number λ is an *eigenvalue* and a vector q ($q \neq 0$) is the corresponding *eigenvector* of an $n \times n$ matrix A if $Aq = \lambda q$. By observing that $(A - \lambda I)q = 0$, we know that the $(A - \lambda I)$ is singular. Then we know that the rank of A is less than n and that $det(A - \lambda I) = 0$. Since the determinant is a polynomial of degree n, we know that it has n roots in $\mathbf{C}$. These n roots are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of A. Without loss of generality, we assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The eigenvector q_i corresponding to the eigenvalue λ_i is the vector q for which $Ax = \lambda_i q$. The *dominant eigenvector* of a matrix is the eigenvector corresponding to the largest eigenvalue.

If A is a symmetric and positive definite matrix, then we know that all of its eigenvalues are real and that a pair of eigenvectors are orthogonal if the corresponding eigenvalues are distinct. To see why eigenvectors q_i and q_j ($i \neq j$) are orthogonal, consider $0 = q_i^T A q_j - q_i^T A q_j = \lambda_j q_i^T q_j - \lambda_i q_i^T q_j = (\lambda_j - \lambda_i) q_i^T q_j$. If λ_i and λ_j are distinct, then q_i and q_j must be orthogonal.

Let Q be the matrix $[q_1 q_2 \ldots q_n]$, and let Λ be the matrix

$$\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Then, by the definition of eigenvectors we know that $AQ = Q\Lambda$. Now, since all of the eigenvectors are orthogonal, we know that $QQ^T = I$, so we can write $A = Q\Lambda Q^T$. This is called the *eigenvalue decomposition* of A. An equivalent way to write the eigenvalue decomposition of A is $\sum_{i=1} n\lambda_i q_i q_i^T$.

From the eigenvalue decomposition we can see that $A^k = Q\Lambda^k Q^T$, so the eigenvalues of A^k are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$, and the eigenvectors are still $q_1, q_2, \ldots, q_n$. Suppose that the first eigenvalue has greatest magnitude, i.e., $|\lambda_1| > |\lambda_n|$. Then as k approaches infinity, for a vector x, $A^k x$ converges to $\lambda_1 q_1 q_1^T x$ as long as $q_1^T x$ is non-zero.

In the discussion of hubs and authorities, we had a matrix A and an initial vectors a_0 and h_0. We were computing the vectors $a_k = (A^T A)^k h_0$ and $h_k = (AA^T)^k a_0$ so that we could determine which documents were good authorities and which documents were good hubs. Since the matrices $A^T A$ and AA^T are symmetric and positive definite, we can use the result above to infer that a_k and h_k converge to the dominant eigenvectors of AA^T and $A^T A$ respectively.

5 The Singular Value Decomposition

Eigenvalues and eigenvectors only exist for square matrices; there is another decomposition of a matrix that is more fundamental. For any matrix A in $\mathbf{R}^{m\times n}$ (without loss of generality assume that m/gen), the *singular value decomposition* of A is written $A = U\Sigma V^T$. U is an $m \times n$ matrix, where $UU^T = I$ (the rows of U, $u_1, u_2, \ldots, u_n$ are called the *left singular vectors*). Σ is a non-zero $n \times n$ diagonal matrix

$$\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix}$$

with the singular values $\sigma_1, \sigma_2, \ldots, \sigma_n$ of A along its diagonal, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. V is an $n \times n$ matrix such that $VV^T = I$ (the rows of V, $v_1, v_2, \ldots, v_n$ are called the *right singular vectors*). The *dominant left and right singular vectors* are defined analogous to dominant eigenvectors. We can observe that $A^T = V\Sigma U^T$ and that hence $AA^T = (V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2 V^T$ and similarly that $A^T A = U\Sigma^2 U^T$. But these are exactly eigenvalue decompositions of AA^T and $A^T A$. Thus the singular values $\sigma_i = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(AA^T)}$ for $1 \leq i \leq n$. The left singular vectors are eigenvectors of AA^T and the right singular vectors are eigenvectors of $A^T A$. We can also write A as $\sum_{i=1} n\sigma_i u_i v_i$. Now, we can see that the hub vector h_k converges to the dominant left singular vector u_1 and that the authority vector converges to the dominant right singular vector v_1.