CS 391D Data Mining: A Mathematical Perspective Fall 2009

Lecture 2

Lecturer: Inderjit Dhillon Date: Sept 3, 2009

Keywords: *Vectors, Matrices, Eigenvalue Decomposition, SVD*

1 Notation

We will denote the set of real numbers by $\mathbb{R}$. 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 by $\mathbb{C}$. The set of all d-dimensional real vectors will be denoted $\mathbb{R}^d$ and the set of all $m \times n$ real matrices will be denoted $\mathbb{R}^{m\times n}$. Vectors in $\mathbb{R}^d$ will be column vectors, e.g.,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}, \quad \mathbf{x} \in \mathbb{R}^d.$$

The corresponding row vector $\mathbf{x}^T$ is written

$$\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}.$$

2 Vector Norms

A vector is described by its "size" and its "direction". The *norm* of a vector $\mathbf{x}$, written $\|\mathbf{x}\|$, is a way to measure the size of a vector. Every norm must satisfy three properties for all vectors $\mathbf{x}$ and $\mathbf{y}$:

1. $\|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = 0$;
2. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$;
3. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

$\|\mathbf{x}\|_2$ denotes the 2-*norm* or the *Euclidean norm* of a vector $\mathbf{x} \in \mathbb{R}^d$ and is equal to $\sqrt{\sum_{i=1}^d |x_i|^2}$. Examples of other norms include the L_1 norm, where $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$ and the L_∞ norm, where $\|\mathbf{x}\|_\infty = \max_{1\leq i\leq d} |x_i|$. 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 $\|\mathbf{x}\|$ to denote $\|\mathbf{x}\|_2$.

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

$$|\mathbf{x}^T\mathbf{y}| \leq \sqrt{\mathbf{x}^T\mathbf{x}} \cdot \sqrt{\mathbf{y}^T\mathbf{y}} = \|\mathbf{x}\|\|\mathbf{y}\|, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

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

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

$$\text{so } \mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta.$$

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

3 Linear Independence

A set of n non-zero vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are *linearly independent* if $\sum_{i=1}^{d} \alpha_i \cdot \mathbf{y}_i = 0$ implies that $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$. The *span* of a set of vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ is $\{\sum_{i=1}^{d} \alpha_i\mathbf{y}_i \mid \alpha_i \in \mathbb{R}\}$, i.e., the set of all linear combinations of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, and is denoted $span(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$. A *basis* of such a linear subspace is a maximal set of linearly independent vectors in the 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 linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. In particular, a matrix $\mathbf{A}$ in $\mathbb{R}^{m\times n}$ maps any n-dimensional vector $\mathbf{x}$ to an m-dimensional vector $\mathbf{Ax}$. Matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^T$. We will write the determinant of $\mathbf{A}$ as $det(\mathbf{A})$. A matrix is said to be *singular* if it has no inverse; a matrix $\mathbf{A}$ is singular if and only if $det(\mathbf{A}) = 0$. 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 matrix 2-norm, where

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

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 turns out to be equal to the maximum singular value of $\mathbf{A}$ (see Section 5). For example, the identity matrix $\mathbf{I}$ has a 2-norm of 1 (since for all $\mathbf{x}$, $\mathbf{Ix} = \mathbf{x}$).

A number λ is an *eigenvalue* and a vector $\mathbf{q}$ ($\mathbf{q} \neq 0$) is the corresponding *eigenvector* of an $n \times n$ matrix $\mathbf{A}$ if $\mathbf{Aq} = \lambda\mathbf{q}$. By observing that $(\mathbf{A} - \lambda\mathbf{I})\mathbf{q} = 0$, we know that $(\mathbf{A} - \lambda\mathbf{I})$ is singular. Thus $det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Since the determinant is a polynomial of degree n, we know that it has n roots in $\mathbb{C}$. These n roots are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathbf{A}$. Without loss of generality, we will order the eigenvalues such that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. The eigenvector $\mathbf{q}_i$ corresponding to the eigenvalue λ_i satisfies $\mathbf{Aq}_i = \lambda_i\mathbf{q}_i$. The *dominant eigenvector* of a matrix is the eigenvector corresponding to the largest eigenvalue.

If $\mathbf{A}$ is real and symmetric, then it can be shown that all of its eigenvalues are real and that every pair of eigenvectors is orthogonal if the corresponding eigenvalues are distinct. To see why eigenvectors $\mathbf{q}_i$ and $\mathbf{q}_j$ $(i \neq j)$ are orthogonal, consider $0 = \mathbf{q}_i^T(\mathbf{Aq}_j) - (\mathbf{q}_i^T\mathbf{A})\mathbf{q}_j = \lambda_j\mathbf{q}_i^T\mathbf{q}_j - \lambda_i\mathbf{q}_i^T\mathbf{q}_j = (\lambda_j - \lambda_i)\mathbf{q}_i^T\mathbf{q}_j$. If λ_i and λ_j are distinct, then $\mathbf{q}_i$ and $\mathbf{q}_j$ must be orthogonal. Even when eigenvalues are equal, the eigenvectors of a real, symmetric matrix may be chosen to be orthogonal. Further, if $\mathbf{A}$ is positive definite, then all eigenvalues are positive.

Let $\mathbf{Q}$ be the matrix $[\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$, and let $\mathbf{\Lambda}$ be the diagonal matrix of eigenvalues:

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

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

From the eigenvalue decomposition we can see that $\mathbf{A}^k = \mathbf{Q\Lambda}^k\mathbf{Q}^T$, so the eigenvalues of $\mathbf{A}^k$ are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$, and the eigenvectors are still $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ (see Theorem 3 in section 6). Suppose that the first eigenvalue has greatest magnitude, i.e., $|\lambda_1| > \lambda_2 \geq \cdots \geq |\lambda_n|$. Then as k approaches infinity, then $\mathbf{A}^k\mathbf{x}/\|\mathbf{A}^k\mathbf{x}\|$ converges to $\mathbf{q}_1$ as long as $\mathbf{q}_1^T\mathbf{x}$ is non-zero.

5 Singular Value Decomposition

Unlike the eigenvalue decomposition which exists only for square matrices, the *singular value decomposition*(SVD) exists for all matrices — it is a more fundamental decomposition. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $m \geq n$, be a matrix. The *singular value decomposition* of $\mathbf{A}$ is the factorization $\mathbf{A} = \mathbf{U\Sigma V}^T$, where $\mathbf{U} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n$, and $\mathbf{\Sigma}$ is a diagonal matrix, $\text{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_i \geq 0$, $1 \leq i \leq n$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are referred to as the *left* and *right singular vectors*, respectively, and the *singular values* of $\mathbf{A}$ are the diagonal elements of $\mathbf{\Sigma}$.

Using the singular value decomposition of $\mathbf{A}$, we have $\mathbf{AA}^T = \mathbf{U\Sigma V}^T\mathbf{V\Sigma}^T\mathbf{U}^T = \mathbf{U\Sigma I\Sigma U}^T = \mathbf{U\Sigma}^2\mathbf{U}^T$. Thus, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{AA}^T$. Similarly, $\mathbf{A}^T\mathbf{A} = \mathbf{V\Sigma}^2\mathbf{V}^T$. Therefore, the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{A}^T\mathbf{A}$. The singular values of $\mathbf{A}$ are the non-negative square roots of the eigenvalues of $\mathbf{AA}^T$ or $\mathbf{A}^T\mathbf{A}$.

Let $\mathbf{u}_i \in \mathbb{R}^m$ and $\mathbf{v}_i \in \mathbb{R}^n$, $1 \leq i \leq n$, denote the i-th columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. Then,

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1\mathbf{v}_1^T \\ \sigma_2\mathbf{v}_2^T \\ \vdots \\ \sigma_n\mathbf{v}_n^T \end{bmatrix} \\
&= \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_n\mathbf{u}_n\mathbf{v}_n^T
\end{aligned}$$

Let $\mathbf{A}_k = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^T$, $k \leq n$, be the k-truncated SVD. Observe that $\mathbf{A}_k$ is a matrix of rank k. Among all matrices of rank k, $\mathbf{A}_k$ serves as the "best" approximation to $\mathbf{A}$ in the following sense. The 2-norm of the approximation error $\|\mathbf{A} - \mathbf{A}_k\|_2 \leq \|\mathbf{A} - \mathbf{M}_k\|_2$, for any matrix $\mathbf{M}_k$ of rank k. This is a classical result in linear algebra, for a proof see Theorem 2.5.3 on page 72 of [GL96].

6 Further Properties

Both the matrices $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are symmetric and positive semi-definite, that is, all eigenvalues are non-negative.

Theorem 1 *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have identical non-zero eigenvalues. Further, if $\mathbf{q}$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$ then $\mathbf{A}^T\mathbf{q}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$.*

Proof: Let λ be a non-zero eigenvalue of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{q}$ be the eigenvector corresponding to λ. Then $(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,\mathbf{q}$. Premultiplying both sides by $\mathbf{A}^T$, we have $\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = (\mathbf{A}^T\mathbf{A})\cdot(\mathbf{A}^T\mathbf{q}) = \lambda\cdot(\mathbf{A}^T\mathbf{q})$. Therefore λ is an eigenvalue of $\mathbf{A}^T\mathbf{A}$ with $\mathbf{A}^T\mathbf{q}$ as the corresponding eigenvector. ■

Theorem 2 *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\mathbf{A}\mathbf{A}^T$ is a positive semi-definite matrix.*

Proof: Let λ be an eigenvalue of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{q}$ be the eigenvector corresponding to λ. Then $(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,\mathbf{q}$. Premultiplying both sides by $\mathbf{q}^T$, we have $\mathbf{q}^T\mathbf{A}\mathbf{A}^T\mathbf{q} = \lambda\,\mathbf{q}^T\mathbf{q}$ which implies that $\lambda = \dfrac{\mathbf{q}^T\mathbf{A}\mathbf{A}^T\mathbf{q}}{\mathbf{q}^T\mathbf{q}} = \dfrac{\mathbf{z}^T\mathbf{z}}{\mathbf{q}^T\mathbf{q}}$, where $\mathbf{z} = \mathbf{A}^T\mathbf{q}$. Observe that $\mathbf{q}^T\mathbf{q} > 0$ and $\mathbf{z}^T\mathbf{z} \geq 0$. Therefore $\lambda \geq 0$ implying that all eigenvalues of $\mathbf{A}\mathbf{A}^T$ are non-negative. ■

We now prove that the eigenvalues of $(\mathbf{A}\mathbf{A}^T)^k$, $k \geq 1$, are related to the eigenvalues of $\mathbf{A}\mathbf{A}^T$. In particular, if λ is an eigenvalue of $\mathbf{A}\mathbf{A}^T$ then λ^k is an eigenvalue of $(\mathbf{A}\mathbf{A}^T)^k$. Moreover, $\mathbf{A}\mathbf{A}^T$ and $(\mathbf{A}\mathbf{A}^T)^k$ have identical eigenvectors.

Theorem 3 *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Further, let $\mathbf{q}$ an eigenvector of $\mathbf{A}\mathbf{A}^T$ corresponding to the eigenvalue λ. Then the matrix $\left(\mathbf{A}\mathbf{A}^T\right)^k$ has λ^k as an eigenvalue with $\mathbf{q}$ being the corresponding eigenvector.*

Proof: We know that $(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,\mathbf{q}$. Premultiplying both sides by $\mathbf{A}\mathbf{A}^T$, we have $(\mathbf{A}\mathbf{A}^T)^2\,\mathbf{q} = \lambda\,(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,(\lambda\,\mathbf{q}) = \lambda^2\,\mathbf{q}$. Thus, by mathematical induction, $(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{q} = \lambda^k\,\mathbf{q}$, $k \geq 1$. ■

Let λ_i and $\mathbf{q}_i$, $0 < i \leq n$, be the eigenvalues and the corresponding eigenvectors, respectively, of $\mathbf{A}\mathbf{A}^T$. Without loss of generality, assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. We can write

$$\begin{aligned}(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{h}_0 &= \lambda_1^k\mathbf{q}_1\mathbf{q}_1^T\mathbf{h}_0 + \lambda_2^k\mathbf{q}_2\mathbf{q}_2^T\mathbf{h}_0 + \cdots + \lambda_n^k\mathbf{q}_n\mathbf{q}_n^T\mathbf{h}_0,\\ \text{so } \mathbf{h}_k &= \frac{(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{h}_0}{\|(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{h}_0\|} \;\to\; \mathbf{q}_1 \text{ as } k\to\infty,\end{aligned}$$

if $\mathbf{q}_1^T\mathbf{h}_0 \neq 0$ and $\lambda_1 > \lambda_2$.

NOTE: Video lectures(Realvideo) from an MIT course on linear algebra are available at http://web.mit.edu/18.06/www/Video/video-fall-99.html. Gil Strang is also the author of some excellent undergraduate textbooks on linear algebra and applied mathematics.

References

[GL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations.* Johns Hopkins University Press, 3rd edition, 1996.