CS 683: Advanced Algorithms	Singular Value Decomposition and Application to Clustering and Dimension-Reduction
04/26/01	Scribe: Chek Beng Chua, Ranjithkumar Rajagopalan

1 Introduction

Suppose we represent documents written in English by vectors with each coordinate representing each word in a vocabulary of, say, 60,000 words. If we wish to classify the documents according to subject, then we would find that most of the words in the vocabulary are not required. This motivates the idea of dimension-reduction.

Suppose there are n documents and m words in our vocabulary. Let the vectors representing the documents be $a_1, \ldots, a_n$ and form the matrix $A = [a_1|\cdots|a_n]$. We would like to be able to say that the vectors $\{a_i\}$ approximately live in a lower dimensional space. When we say that the vectors $\{a_i\}$ live in a low-dimensional space, it is the same as saying that the columns of A span a low-dimensional space. We call the span of the columns of A the *range space* of A and denote it by $range(A)$. The dimension of $range(A)$ is called the *rank* of A and denoted by $rank(A)$. The *null space* of A defined to be $\{x : Ax = 0\}$ and denoted by $null(A)$. It is well known that $dim(range(A)) + dim(null(A)) = n$.

2 Singular Value Decomposition

How can we tell if $\{a_i\}$ approximately lives in a low-dimensional space? In view of the discussion above, we may ask the equivalent question : How can we tell if A approximately low rank? In order to answer this question, let us view A as a function from $\mathbb{R}^n$ into $\mathbb{R}^m$ defined by $x \mapsto Ax$. Consider A's action on the unit sphere S^{n-1} in $\mathbb{R}^n$. A will map S^{n-1} to some ellipsoid in $\mathbb{R}^m$. The length of the longest axis of this ellipsoid is given by

$$\max_{||x||=1} ||Ax|| \equiv \max_{x\neq 0} \frac{||Ax||}{||x||} =: ||A||_2,$$

which is called the *2-norm* of A.

Definition 2.1 *Q is an* orthogonal matrix *if Q is square and $Q^TQ = I$ (or equivalently, the columns (or rows) of Q are orthonormal).*

Here are some properties of an orthogonal matrix Q.

1. $||Q||_2 = 1$ since for all $x \in \mathbb{R}^n$, $||Qx||^2 = (Qx)^T(Qx) = x^TQ^TQx = x^Tx = ||x||^2$.

2. $||QA||_2 = ||A||$ since for all $x \in \mathbb{R}^n$, $||QAx||^2 = (QAx)^T(QAx) = x^TA^TQ^TQAx = (Ax)^T(Ax) = ||Ax||^2$.

3. $||AQ||_2 = ||A||$ since $||Qx|| = ||x||$.

Theorem 2.2 *(Existence of Singular Value Decomposition) For all matrices $A \in \mathbb{R}^{m\times n}$ ($n < m$), $U^T AV = \Sigma$ for some matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and $\Sigma \in \mathbb{R}^{m\times n}$, where U and V are both orthogonal and $\Sigma = Diag(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$. [$A = U^T \Sigma V$ is called a singular value decomposition (SVD) of A.]*

Proof. We will prove the theorem by induction on the size of A.

Let $\sigma = ||A||_2$. Choose unit vectors x and y so that $Ax = \sigma y$. Extend both x and y to orthonormal bases in their respective spaces. With these bases, form the orthogonal matrices $V = [x|V_1]$ and $U = [y|U_1]$. Now,

$$U^T AV = U^T[\sigma y|AV_1] = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix},$$

for some vector w and matrix B. We need to argue that $w = 0$. Let $A_1 = U^T AV$. Then, $||A_1||_2 = ||A||_2 = \sigma$. Let z be a unit vector in the direction of $[\sigma, w^T]$. Then,

$$||A_1 z|| \geq \frac{1}{\sigma^2 + w^T w}[\sigma, w]^T[\sigma, w] = (\sigma^2 + w^T w)^{1/2} \geq \sigma.$$

Since $||A_1||_2 = \sigma$, equality must hold in the above inequality. So, $w^T w = 0$, which implies that $w = 0$.

By inductive hypothesis, B has a SVD, i.e. there exist some orthogonal $\hat{U}$ and $\hat{V}$, and diagonal Σ such that $\hat{U}^T B\hat{V} = \Sigma$. Then,

$$\begin{bmatrix} 1 & 0 \\ 0 & \hat{U} \end{bmatrix}^T U^T AV \begin{bmatrix} 1 & 0 \\ 0 & \hat{V} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \hat{U} \end{bmatrix}^T \begin{bmatrix} \sigma & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \hat{V} \end{bmatrix} = \begin{bmatrix} \sigma & 0 \\ 0 & \hat{U}^T B\hat{V} \end{bmatrix} = \begin{bmatrix} \sigma & 0 \\ 0 & \Sigma \end{bmatrix},$$

gives a SVD for A. ■

Here are some facts about SVD.

- $||A||_2 = \sigma_1$.
- If r is such that $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_n$, then $rank(A) = r$.
- (Variational characterization of SVD) $\sigma_1 = \max_{||x||=1,||y||=1} y^T Ax$. Here's a brief justification of the equality. To maximize $y^T Ax$, we should choose x so that Ax is as long as possible and then choose y to be aligned with Ax. So, the choices for x and y would be the first columns of V and U respectively.

3 Application to Low-Dimensional Approximation

Now, we go back to the problem of approximating $\{a_i\}$ by some $\{a_i'\}$ that live in some k-dimensional subspace. This is equivalent to approximating A by some B which has rank $\leq k$. To measure closeness between matrices, we shall use the 2-norm.

Theorem 3.1 *(Eckart - Young) Suppose $UDiag(\sigma_1, \ldots, \sigma_n)V^T$ is a SVD of A. Let $A_k = UDiag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)V^T$. Then,*

$$||A - A_k||_2 = \min_{rank(B)=k} ||A - B||_2.$$

Proof. Let $\Sigma_k = U^T(A - A_k)V = Diag(\sigma_1, \ldots, \sigma_k, \sigma_{k+1}, \ldots, \sigma_n) - Diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) = Diag(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_n)$. Then, $||A - A_k||_2 = \sigma_{k+1}$. Take any rank-$k$ matrix B. We only need to show that $||A - B||_2 \geq \sigma_{k+1}$, or equivalently, there exists a unit vector z such that $||(A - B)z|| \geq \sigma_{k+1}$. Let $x_1, \ldots, x_{n-k}$ be a basis for $null(B)$. Since $(n - k) + (k + 1) > n$, $span\{x_1, \ldots, x_{n-k}$ and $span\{v_1, \ldots, v_{k+1}\}$ intersect in a non-zero vector (and hence a unit vector z). This will be our z since $z = \sum_{i=1}^{k+1} \beta_i v_i$ for some β_i's and

$$\begin{aligned}
Az &= U\Sigma V^T z = U\Sigma(\beta_1, \ldots, \beta_{k+1}, 0, \ldots, 0)^T \\
&= U(\sigma_1\beta_1, \ldots, \sigma_{k+1}\beta_{k+1}, 0, \ldots, 0)^T \\
&= \sum_{i=1}^{k+1} \sigma_i \beta_i u_i \\
\Rightarrow ||Az||^2 &= \sum_{i=1}^{k+1} \sigma_i^2 \beta_i^2 \geq \sigma_{k+1}^2 \\
\Rightarrow ||Az - Bz|| &\geq \sigma_{k+1}.
\end{aligned}$$

■

There is no reason why the 2-norm should be the norm to use for approximation. While there are many other norms to use, the A_k in the theorem statement minimizes the distance to A for many of these norms. In particular, this is true for the following norm.

Definition 3.2 *The* Frobenius norm *of a matrix is defined by*

$$||A||_F := (\sum_{i,j} a_{ij}^2)^{1/2}.$$

Theorem 3.3 $||A - A_k||_F = \min_{rank(B)=k} ||A - B||_F$.

In the next section, we will see that the Frobenius norm is useful for clustering problems.

4 Clustering

Instead of viewing low-dimensional approximation as dimension-reduction, we can view it in terms of clustering. Given $\{a_i\}$, we may want to find some k "essential features" via clustering, where k here is small compared to n. Here are a few ways to do this.

- k-median : Choose "centers" $c_1, \ldots, c_k \in \mathbb{R}^m$ to minimize $\sum_{i=1}^n d(a_i, \{c_1, \ldots, c_k\})$, where $d(x, S)$ is the Euclidean distance of the point x from the set S.
- k-means : Choose $c_1, \ldots, c_k \in \mathbb{R}^m$ to minimize $[\sum_{i=1}^n d(a_i, \{c_1, \ldots, c_k\})^2]^{1/2}$.
- Subspace clustering : Choose some k-dimensional subspace S to minimize $\sum_{i=1}^n d(a_i, S)^2$ over all k-dimensional subspaces.

It is NP-hard to obtain the k-median and k-means. In contrast, subspace clustering can be solved by SVD, and in fact via A_k.

Let $S = range(A_k)$ and a_i' be the i-th column of A_k. For any k-dimensional subspace S', let b_i be the closest point in S' to a_i. Let B have columns $\{b_i\}$, so that $rank(B) \leq k$. Since

$$\begin{aligned}\sum_{i=1}^n d(a_i, S')^2 &= \sum_{i=1}^n d(a_i, b_i)^2 = ||A - B||_F^2 \\ &\geq ||A - A_k||_F^2 = \sum_{i=1}^n d(a_i, a_i')^2 \geq \sum_{i=1}^n d(a_i, S)^2,\end{aligned}$$

S solves the subspace clustering problem.