Lecture 6 for EE127 (Fall 2018): SVD

Scribes: Khalil Sarwari

9/11/2018

1 Motivation

What better way is there to understand an object than to dissect it? Our object of interest is the matrix $A \in \mathbb{R}^{m \times n}$. We have already been introduced to fundamental procedure of diagonalization, which can be applied to A under certain conditions. Here we set up and introduce a new procedure, singular value decomposition (SVD), which lends insight into the essence of any such matrix A.

2 Dyads

Definition 2.1. A matrix $A \in \mathbb{R}^{m \times n}$ is called a *dyad* if it can be written as

$$A = pq^\top$$

for some vectors $p \in \mathbb{R}^m, q \in \mathbb{R}^n$

Example 2.2. The following matrix A is a dyad.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$$

One interpretation of a dyad A is that the columns of A are scaled versions of p, with the scalar factors coming from q. Similarly, the rows of A are scaled versions of $q^\top$, with the scalar factors coming from p. In short, *the rows are proportional, the columns are proportional*.

Example 2.3. Consider a video that consists of two scenes that have practically no activity (i.e. abandoned mall day/night). The two series of images can then be roughly approximated by two dyads $pq^\top$ and $rs^\top$, one for each scene. Each dyad will have a row vector representing the image of the scene, and the series would be various scalings of this row. The video as a whole can then be approximated by a combination of these two dyads, namely $\begin{pmatrix} p \\ 0 \end{pmatrix} q^\top + \begin{pmatrix} 0 \\ r \end{pmatrix} s^\top$.

We can make the following observations:

- If we wish to represent a matrix as a sum of dyads, increasing the number of dyads generally helps improve the quality of the approximation.
- If we wish to represent a matrix as a sum of dyads, and also reduce redundancy among the dyads, it helps to have the $p \in \mathbb{R}^m$ orthogonal to one another. This insight also applies to the $q \in \mathbb{R}^n$. Reducing redundancy in this way keeps the cost of the approximation low.

3 Singular Value Decomposition (SVD)

The Singular Value Decomposition decomposes any matrix $A \in \mathbb{R}^{m \times n}$ into a product of three matrices.

Theorem 3.1. ***SVD***
Any matrix $A \in \mathbb{R}^{m \times n}$ can be factorized as

$$A = U\tilde{\Sigma}V^\top$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\tilde{\Sigma} \in \mathbb{R}^{m \times n}$ is a matrix having the first $r := rank(A)$ diagonal entries $(\sigma_1, ..., \sigma_r)$ positive and decreasing in magnitude, and all other entries 0.

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}, \ \Sigma = diag(\sigma_1, ..., \sigma_r) \succ 0$$

Compact-form SVD follows directly from this theorem.

Corollary 3.2. ***Compact-form SVD***
Any matrix $A \in \mathbb{R}^{m \times n}$ can expressed as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top = U_r \Sigma V_r^\top$$

where:

- *$U_r = \begin{pmatrix} u_1 \ ... \ u_r \end{pmatrix}$ is such that $U_r^\top U_r = I_r$ and $V_r = \begin{pmatrix} v_1 \ ... \ v_r \end{pmatrix}$ is such that $V_r^\top V_r = I_r$*
- *The positive numbers σ_i are called the singular values of A, and are arranged such that $\sigma_1 \geqslant \sigma_2 \geqslant ... \geqslant \sigma_r > 0$*
- *The vectors u_i are called the left singular vectors of A, such that $Av_i = \sigma_i u_i$ and the vectors v_i are called the right singular vectors of A such that $u_i^\top A = \sigma_i v_i^\top$*
- *$\sigma_i^2 = \lambda_i(AA^\top) = \lambda_i(A^\top A), i = 1, ..., r$ and u_i, v_i are the normalized eigenvectors of $AA^\top$ and of $A^\top A$, respectively.*

4 Interpretations

Dyadic

The SVD theorem allows us to us to write any matrix as a sum of dyads:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

The vectors u_i are orthonormal, and so are the vectors v_i. Thus, we have minimal redundancy, with $\sigma_i > 0$ providing the weight of the dyad i in the overall summation.

The sum of the first k dyads provide the best rank k approximation to the matrix A (discussed further in Section 6).

Geometric

Each part of the SVD has a geometric interpretation. The matrices U and $V^\top$ are both rotation matrices. The matrix $\tilde{\Sigma}$ stretches input vectors (not necessarily all the same amount or same direction). This interpretation means that any linear transformation between two finite dimensional spaces can be broken down into rotation, scaling, and rotation.

5 Matrix Properties

From the singular value decomposition, we can observe essential matrix properties.

Rank, Nullspace and Range

- The rank r of A is the cardinality of the nonzero singular values, that is the number of nonzero entries on the diagonal of $\tilde{\Sigma}$.
- Since $r = rank(A)$, we have dim $\mathcal{N}(A) = n - r$. A basis for $\mathcal{N}(A)$ is the set of the last $n - r$ columns of V. Namely, we have the following

  $$\mathcal{N}(A) = span(\{v_{r+1}, ..., v_n\})$$

  With orthogonality sending the top r entries to zero and the 0s on the diagonal in $\tilde{\Sigma}$ killing off the rest, one can confirm that these vectors constitute $\mathcal{N}(A)$.
- Similarly, an orthonormal basis spanning the range of A is given by the first r columns of U. Namely, we have the following

  $$\mathcal{R}(A) = span(\{u_1, ..., u_r\})$$

Matrix Norms

Recall that the trace of a square matrix is the sum of its eigenvalues. Then the square of the Frobenius matrix norm of a matrix $\in \mathbb{R}^{m\times n}$ can be defined as

$$\|A\|_F^2 = trace(A^\top A) = \sum_{i=1}^{n} \lambda_i(A^\top A) = \sum_{i=1}^{n} \sigma_i^2$$

where σ_i are the singular values of A. Thus, the square of the Frobenius norm is simply the sum of its squared singular values.

Furthermore, recall that the squared spectral matrix norm $\|A\|_2^2$ is equal to the maximum eigenvalue of $A^\top A$. Thus, $\|A\|_2^2 = \sigma_1^2$, the square of the largest singular value. Equivalently, the spectral norm is the largest singular value of A. More formally, we have

Proof.

$$\|A\|_2^2 = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x\neq 0} \frac{\|U\tilde{\Sigma}V^\top x\|}{\|x\|} = \max_{x\neq 0} \frac{\|\tilde{\Sigma}V^\top x\|}{\|x\|}$$

Since orthogonal matrices do not affect the magnitude their inputs, we can drop U. We cannot drop the $V^\top$ since its rotation effect may have consequences when $\tilde{\Sigma}$ is applied. However, we can instead add it into the denominator. Substituting $z = V^\top x$, we get

$$\max_{z\neq 0} \frac{\|\tilde{\Sigma}z\|}{\|z\|} = \max_{z} \|\tilde{\Sigma}z\| : z^\top z = 1 = \max_{z} \sum_{i=1}^{r} \sigma_i^2 z_i^2 : z^\top z = 1$$

rewriting this in terms of $p_i = z_i^2$s, we obtain a probabilistic perspective

$$\max_{\mathbf{1}^\top p=1, p\geqslant 0} \sum_{i=1}^{r} \sigma_i^2 p_i$$

Clearly, the upper bound on the optimal value is $\max_{1\leqslant i\leqslant r} \sigma_i^2$. In fact, this can be easily obtained by setting $p_i = 1$ at the index of the maximum singular value, and 0 everywhere else. ■

Definition 5.1. The *nuclear norm* of a matrix A is denoted as follows:

$$\|A\|_* = \sum_{i=1}^{r} \sigma_i$$

where $r = rank(A)$. This norm appears in low-rank matrix problems.

Condition Number

Definition 5.2. The condition number of an invertible (all $\sigma_i > 0$) matrix is the ratio between the largest and smallest singular value:

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} = \|A\|_2 \cdot \|A^{-1}\|_2$$

This ratio provides a quantitative measure of how close A is to being singular (larger $\kappa(A)$ means closer to being singular).

Matrix Pseudo-Inverses

Definition 5.3. A *pseudoinverse* of a matrix $A \in \mathbb{R}^{m\times n}$ is a matrix $A^\dagger$ that satisfies:

$$AA^\dagger A = A$$
$$A^\dagger AA^\dagger = A^\dagger$$
$$(AA^\dagger)^\top = AA^\dagger$$
$$(A^\dagger A)^\top = A^\dagger A$$

Definition 5.4. The *Moore-Penrose pseudoinverse* of a matrix $A \in \mathbb{R}^{m\times n}$ is:

$$A^\dagger = V\tilde{\Sigma}^\dagger U^\top$$

where

$$\tilde{\Sigma}^\dagger = \begin{pmatrix} \Sigma^{-1} & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},\ \Sigma^{-1} = diag(\sigma_1^{-1}, ..., \sigma_r^{-1}) \succ 0$$

Special Cases

- If A is square and nonsingular, then $A^\dagger = A^{-1}$
- If A is full column rank (so $r = n \leqslant m$), then $A^\dagger A = V_r V_r^\top = VV^\top = I_n$. Thus, $A^\dagger$ is a left inverse of A, and can be written $A^\dagger = (A^\top A)^{-1}A^\top$
- If A is full row rank (so $r = m \leqslant n$), then $AA^\dagger = U_r U_r^\top = UU^\top = I_m$. Thus, $A^\dagger$ is a right inverse of A, and can be written $A^\dagger = A^\top(AA^\top)^{-1}$

Projectors

- $P_{\mathcal{R}(A)} = U_r U_r^\top$ is the matrix for an orthogonal projection onto $\mathcal{R}(A)$.
- $P_{\mathcal{R}(A)^\perp} = I_m - AA^\dagger$ is the matrix for an orthogonal projection onto $\mathcal{R}(A)^\perp$.
- $P_{\mathcal{N}(A)} = I_n - A^\dagger A$ is the matrix for an orthogonal projection onto $\mathcal{N}(A)$.
- $P_{\mathcal{N}(A)^\perp} = A^\dagger A$ is the matrix for an orthogonal projection onto $\mathcal{N}(A)^\perp$.

6 Low Rank Matrix Approximation

The ratio $\eta_k = \frac{\|A_k\|_F^2}{\|A\|_F^2} = \frac{\sigma_1^2+...+\sigma_k^2}{\sigma_1^2+...+\sigma_r^2}$ indicates the proportion of the variance of A explained by the best rank k approximation of A.
Note that

$$\|A - A_k\|_F^2 = \| \sum_{k+1}^{r} \sigma_i u_i v_i^\top \|_F^2 = \sum_{k+1}^{r} \sigma_i^2$$

The approximation error e_k follows directly and is defined as:

$$e_k = \frac{\|A - A_k\|_F^2}{\|A\|_F^2} = \frac{\sigma_{k+1}^2 + ... + \sigma_r^2}{\sigma_1^2 + ... + \sigma_r^2} = 1 - \eta_k$$

Minimum "Distance" to Rank Deficiency

Consider the problem of finding δA, the smallest perturbation of a rank n matrix A, which makes $rank(A + \delta A) = n - 1$. The optimal solution to this problem is $\delta A^* = A_{n-1} - A$, where A_{n-1} is the rank $n-1$ approximation to A.

In other words, $\delta A^* = \sigma_n u_n v_n^\top$ and the "distance" is simply $\|\delta A^*\|_F = \sigma_n$

7 Link with PCA

Let $x_i \in \mathbb{R}^n, i = 1, ..., m$ be data points. Let $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$, such that $\bar{x}$ is the center of the data points. Our data matrix $\tilde{X} \in \mathbb{R}^{n \times m}$ is a matrix containing the centered data points:

$$\tilde{X} = [\tilde{x}_1 \ \dots \ \tilde{x}_m], \text{ where } \tilde{x}_i = x_i - \bar{x}\ , i = 1, ..., m$$

We wish to find the direction $z \in \mathbb{R}^n$ in which the variance of the projection of the data points on the line defined by z is maximized.
For each x_i, the component along z is $\alpha = \tilde{x}_i z^\top$. The mean square variation of the data along z is then

$$\frac{1}{m} \sum_{i=1}^{n} \alpha_i^2 = \frac{1}{m} \sum_{i=1}^{n} z^\top x_i x_i^\top z = \frac{1}{m} z^\top \tilde{X} \tilde{X}^\top z$$

Thus, we have the following optimization problem:

$$\max_{z\,:\,z^\top z = 1} z^\top (\tilde{X}\tilde{X}^\top) z$$

We know the solution is the largest eigenvalue of $\tilde{X}\tilde{X}^\top$. Applying SVD, $\tilde{X} = U_r \Sigma V_r^\top$, we get that the largest eigenvalue of $\tilde{X}\tilde{X}^\top = U_r \Sigma^2 U_r^\top$ is σ_1^2, with maximum direction of variation $z = u_1$, the first column of U_r corresponding to σ_1^2.

Additional principal axes can be found by "removing" the first principal components, and applying the same approach again on the "deflated" data matrix.