CS205 – Class 5

Covered in class: *1, 3, 4, 5.*
Reading: *Heath Chapter 4.*

Singular Value Decomposition (SVD)

1. The Singular Value Decomposition is an eigenvalue-like decomposition for rectangular $m \times n$ matrices. It has the form $A = U \Sigma V^T$ where U is an $m \times m$ orthogonal matrix, *V* is an $n \times n$ orthogonal matrix, and Σ is an $m \times n$ diagonal matrix with positive diagonal entries that are called the *singular values* of A. The columns of *U* and *V* are the *singular vectors*.
   a. Introduced and rediscovered many times: Beltrami in 1873, Jordan in 1875, Sylvester in 1889, Autonne in 1913, Eckart and Young in 1936.
   b. Pearson introduced principle component analysis (PCA) in 1901. It uses SVD.
   c. Numerical work by Chan, Businger, Golub, Kahan, etc.
2. The singular value decomposition of $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$ is given by

$$\begin{bmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .664 & -.509 \\ .750 & -.371 & -.542 & .079 \end{bmatrix} \begin{bmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{bmatrix}.$$

   a. The singular values are 25.5, 1.29, and 0. The singular value of 0 indicates that the matrix is rank deficient. However, even a “small” singular value could indicate a “zero” and a rank deficient matrix.
3. The singular values of A are the non-negative square roots of the eigenvalues of the symmetric positive semi-definite $A^T A$ (and also AA^T), and the columns of U and V are the orthonormal eigenvectors of AA^T and $A^T A$ respectively. (Note the strong connection to the normal equations and least squares problems).
4. The condition number of a matrix A with respect to the Euclidean norm is $\sigma_{\max} / \sigma_{\min}$.
   a. For a square matrix, the condition number measures the closeness to singularity. For a rectangular matrix, the condition number measures the closeness to rank deficiency.
5. The rank of a matrix is equal to the number of nonzero singular values that it has. However, if values are ”close” to “zero” then the condition number $\sigma_{\max} / \sigma_{\min}$ can be very high essentially making these numbers “zero” as far as rank is concerned.
6. The columns of V corresponding to “zero” singular values form an orthonormal basis for the null space of A.
   a. The remaining columns of V form an orthonormal basis for the space perpendicular to the null space of A.
7. The columns of U corresponding to the “nonzero” singular values form an orthonormal basis for the range of A.
   a. The remaining columns of U form an orthonormal basis for the space perpendicular to the range of A.
8. The columns of V corresponding to zero columns of Σ and the columns of U corresponding to zero rows of Σ along with those zero columns and rows can then be omitted without changing their product.

a. Applying this to the SVD of A from part 4 gives us the new reduced SVD,

$$\begin{bmatrix} .141 & .825 \\ .344 & .426 \\ .547 & .028 \\ .750 & -.371 \end{bmatrix} \begin{bmatrix} 25.5 & 0 \\ 0 & 1.29 \end{bmatrix} \begin{bmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \end{bmatrix}$$

9. SVD is a transformation into a diagonal axis aligned space.
    a. Transform b into the space spanned by U^T, $U^T U\Sigma V^T x = \Sigma V^T x = U^T b = \hat{b}$. No information is lost going from b to $\hat{b}$ because U^T is square and orthogonal.
    b. Replace $V^T x$ by $\hat{x}$ to get a diagonal system, $\Sigma V^T x = \Sigma\hat{x} = \hat{b}$.
    c. Now solve the system $\Sigma\hat{x} = \hat{b}$ simply by scaling elements of $\hat{b}$ by the singular values.
    d. The original x is then recovered as $x = V\hat{x}$.
    e. Essentially the SVD solves the matrix by transforming the vectors in a space with eigenvectors along the unit axis.
10. $A = U\sum V^T = \sum_i \sigma_i u_i v_i^T$

    proof: define $l = \min(m,n)$, $\hat{U}$ the first l columns of U, $\hat{\Sigma}$ the square $l \times l$ submatrix from the upper left corner of Σ, $\hat{V}$ the first l columns of V. Then

$$A = U\Sigma V^T = \hat{U}\hat{\Sigma}\hat{V}^T = \begin{pmatrix} u_1 & \cdots & u_l \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_l^T \end{pmatrix} = \begin{pmatrix} u_1 & \cdots & u_l \end{pmatrix} \begin{pmatrix} \sigma_1 v_1^T \\ \vdots \\ \sigma_l v_l^T \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{l} \sigma_i u_i^1 v_i^1 & \cdots & \sum_{i=1}^{l} \sigma_i u_i^1 v_i^n \\ \vdots & & \vdots \\ \sum_{i=1}^{l} \sigma_i u_i^m v_i^1 & \cdots & \sum_{i=1}^{l} \sigma_i u_i^m v_i^n \end{pmatrix} = \sum_{i=1}^{l} \begin{pmatrix} \sigma_i u_i^1 v_i^1 & \cdots & \sigma_i u_i^1 v_i^n \\ \vdots & & \vdots \\ \sigma_i u_i^m v_i^1 & \cdots & \sigma_i u_i^m v_i^n \end{pmatrix} = \sum_{i=1}^{l} \sigma_i u_i v_i^T$$

    f. Note that “zero” or “small” σ_i produce terms that contribute little to the sum, and that large σ_i produce terms that contribute significantly to the sum.
    g. If the “zero” or “small” σ_i are omitted from the summation, one obtains a matrix with lower rank. For example, if only the first k terms are summed, the result has rank k.
        i. Moreover, it can be shown that this new rank k matrix is the closest rank k matrix to A in both the L_2 and the Frobenius norm.
        ii. This is the key idea in PCA, clustering/data mining algorithms, etc.
11. The “pseudo-inverse” of a matrix A is defined by $A^+ = V\sum^+ U^T$ where $\sum^+$ is obtained from $\sum$ by replacing all “nonzero” σ_i with $1/\sigma_i$, and leaving all the zero entries *identically zero*.
    h. If A is square and nonsingular ($\sigma_i \neq 0$), $A^+ = A^{-1}$.
    i. The least squares solution to Ax=b is $x = A^+ b = V\Sigma^+ U^T b = \sum_{\sigma_i \neq 0} \left(u_i^T b / \sigma_i\right) v_i$. (Note Σ^+ contains a transpose)

i. Moreover, small σ_i can be dropped from the summation stabilizing the solution, and effectively improving the condition number. This amounts to “dropping columns” from the original matrix A.