CHAPTER 15

Appendices

1 Collaborative filtering and the SVD

This appendix presents the relationship between collaborative filtering and the singular value decomposition, a fundamental matrix decomposition method in linear algebra.

Recall that the goal of collaborative filtering is to approximate an $n \times m$ matrix Y into a rank k matrix X, where X is decomposed as the product of an $n \times k$ matrix U, and a $k \times m$ matrix V^T, i.e., $X = UV^T$. Let us denote the k *columns* of U as $\{u_i\}$ (for i in $1 \cdots k$), and the k *rows* of V^T as $\{v_i\}$. This decomposition may be visualized as the following matrix product:

$$X = UV^T = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_k \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ & \vdots & \\ — & v_k & — \end{bmatrix} \tag{15.1}$$

Recall that k is the rank of X, and thus each u_i is linearly independent of all other column vectors of U, and similarly for v_i. (This makes the k features independent of each other.) Because of this linear independence, we may choose u_i such that $(u_i)^T u_j = \|u_i\|^2 \delta_{ij}$ is zero for $i \neq j$, meaning that each u_i is *orthogonal* to other column vectors of U. This orthogonalization can be done using a Gram-Schmidt process, for example. The row vectors v_i can similarly be constructed to be orthogonal, and in the following we assume the $\{u_i\}$ and $\{v_i\}$ are each sets of orthogonal vectors.

Consider, now, what happens when the $n \times m$ matrix X is left-multiplied by one of the $1 \times n$ vectors $(u_i)^T$. Evidently

$$(u_i)^T X = \sum_j (u_i)^T u_j v_j = \|u_i\|^2 v_i \,, \tag{15.2}$$

where $\|u_i\|^2$ is the square of the norm of the vector u_i. Similarly, when X is right-multiplied by one of the $m \times 1$ vectors $(v_i)^T$, we get:

$$X(v_i)^T = \sum_j u_j v_j (v_i)^T = \|v_i\|^2 u_i \,. \tag{15.3}$$

Combining these to compute the right-multiplication of $X^T X$ by $(v_i)^T$, we observe something very interesting:

$$X^T X (v_i)^T = \|v_i\|^2 X^T u_i = \|v_i\|^2 ((u_i)^T X)^T = \|v_i\|^2 \|u_i\|^2 (v_i)^T , \tag{15.4}$$

which means that $(v_i)^T$ is a "right" *eigenvector* of $X^T X$, with eigenvalue $\|v_i\|^2\|u_i\|^2$! Similarly, the left-multiplication of XX^T by $(u_i)^T$ gives:

$$(u_i)^T XX^T = \|u_i\|^2 v_i X^T = \|u_i\|^2 (X(v_i)^T)^T = \|u_i\|^2\|v_i\|^2 (u_i)^T , \tag{15.5}$$

which means that $(u_i)^T$ is a "left" eigenvector of XX^T, with eigenvalue $\|u_i\|^2\|v_i\|^2$.

How many eigenvectors do we have? Well, $X^T X$ is an $m \times m$ positive-definite matrix, so it must have m eigenvectors; let us thus extend our definition of $\{v_i\}$ to be all these m eigenvectors. And XX^T is an $n \times n$ positive-definite matrix, which has n eigenvectors, so let us similarly extend our definition of $\{u_i\}$.

Eigenvectors provide orthogonal bases for linear vector spaces, and thus it is convenient to normalize them. Let us define

$$\tilde{u}_i = \frac{u_i}{\|u_i\|} \tag{15.6}$$

$$\tilde{v}_i = \frac{v_i}{\|v_i\|} \tag{15.7}$$

as the columns of matrix $\tilde{U}$ and the rows of matrix $\tilde{V}^T$. Let us also define the diagonal matrix $\Lambda_{ii} = \|u_i\|\|v_i\|$. Using these newly defined matrices, we may now rewrite X as

$$X = UV^T = \tilde{U}\Lambda\tilde{V}^T , \tag{15.8}$$

where, to summarize, the $n \times n$ matrix $\tilde{U}$ has as its columns the normalized left-eigenvectors of XX^T, the $m \times m$ matrix $\tilde{V}$ has as its rows the normalized right-eigenvectors of $X^T X$, and Λ is a diagonal matrix of the products of the square roots of the eigenvalues.

This is known as the *singular value decomposition* of X! The SVD is a standard matrix decomposition, and the diagonal elements of Λ are known as the *singular values* of X. These singular values are non-negative, real values, and thus Λ may be viewed as a scaling matrix. Meanwhile, $\tilde{U}$ and $\tilde{V}^T$ geometrically act as rotation matrices, because they are *unitary* matrices: $\tilde{U}^T\tilde{U} = I$ and similarly for $\tilde{V}^T$.

How does this relate to the collaborative filtering decomposition, where U and V are not square matrices? Well, the largest singular values of Y contribute "most" to Y, and thus an important method of approximating Y to some degree k is to compute $Y' = \tilde{U}\Lambda'\tilde{V}^T$, where Λ' only keeps the k largest singular values, and drops the rest to zero. We may also drop rows of $\tilde{V}^T$ and columns of $\tilde{U}$ corresponding to the dropped singular values, producing rank-k matrices U and V. This is known as taking the rank k *principal component* of Y, providing Y' which has the smallest possible Frobenius norm with Y, i.e., minimizing the square root of the sum of the absolute square of the elements of $Y - Y'$.

This shows how singular value decomposition is mathematically related to collaborative filtering, but how do they compare algorithmically? In practice, Y may have missing (or hidden) entries, and standard techniques for computing the SVD may not be robust against such missing data. Computing full sets of eigenvectors and eigenvalues can also be computationally expensive, especially if you only want those corresponding to the largest k singular values. Thus, there can be advantages to collaborative filtering algorithms, e.g., those based on gradient descent, although modifications can also be made to improve the robustness of the SVD approach.