SVD and LSA

I. SVD Fundamentals

Singular Value Decomposition (SVD) is a matrix decomposition technique. Formally, the singular value decomposition of an m x n matrix A is as follows:

$$A = U\Sigma V^T$$

Where U is an m x m unitary matrix, Σ is the an m x n diagonal matrix of singular values $\sigma_1 ... \sigma_n$, and V is an n x n unitary matrix. To derive the equation, it is useful to think geometrically. Consider the idea that the image of a unit sphere under any m x n matrix is a hyperellipse, which is obtained by stretching the unit sphere in R^n by some factors $\sigma_1 ... \sigma_n$ in orthogonal directions. This motivates the following definition.

Definition 1.1: Let S be the unit sphere in R^n, and A an m x n matrix. Then AS is a hyper ellipse in R^m, and the n singular values $\sigma_1 ... \sigma_n$ of A are the lengths of the n principal semiaxes of AS, with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n > 0$.

Definition 1.2: A non-negative real number σ is a singular value of A if and only if there exist unit-length, orthonormal vectors u_j in R^m and v_j in R^n such that $Av_j = \sigma_j u_j$.

Geometrically speaking, U is the matrix of left singular vectors of A that define the directions of the n principal semiaxes of AS, and V is matrix of right singular vectors of A that are the pre-images of the n principal semiaxes of AS. The equation in Definition 1.2 is the core of SVD. The collection of vectors satisfying $Av_j = \sigma_j u_j$ is $AV = U\Sigma$, and because every v_j is orthonormal, V is unitary, meaning $V^{-1} = V^T$, and we arrive at $A = U\Sigma V^T$.

II. Applying Latent Semantic Analysis

Consider a body of text. We can represent the text with the m x n matrix $X = [x_1, x_2, ..., x_n]$, $X \in R^{mxn}$ where each column vector d_j represents sentence j of the text. Each row of X is a term-frequency vector, represented by t_i^T.

$$t_i^T = [x_{i1} ... x_{i,n}] \text{ and } d_j = [x_{i,j} ... x_{m,j}]$$

The application of SVD on X gives us the following decomposition:

$$\begin{array}{ccccccc} & X & & & U & \Sigma & V^T \\ & (\mathbf{d}_j) & & & & & (\hat{\mathbf{d}}_j) \\ & \downarrow & & & & & \downarrow \\ (\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ & & \\ \vdots & \ddots & \vdots \\ & & \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix} \end{array}$$

The crux of the tool here is approximating X by some other matrix X_k, which has a specific rank $k, k \leq n$. The approximation is based on minimizing the norm of the difference between X and X_k, under the constraint that $rank(X_k) = k$. The solution is given by the SVD of X, namely:

$$X' = U_k \Sigma_k V_k^T$$

Where Σ_k is a truncated version of Σ, in that it contains only the k largest singular values of Σ. The existence of this lower-dimensional approximation is stated by the Eckart-Young Theorem.

Theorem 2.2: (Eckart-Young) Let $A = U\Sigma V^T = Udiag(\sigma_1 ... \sigma_n)V^T$. Let B be some matrix with $rank < v$. For any v with $0 \leq v \leq n, A_v = \Sigma_{i=1}^{v} \sigma_i u_i v_i^T$,

$$||A - A_v||_2 = min||A - B||_2 = \sigma_{v+1}$$

Proof:
First part: Suppose there exists some B with rank(B)≤v such that $||A - B||_2 < ||A - A_v||_2 = \sigma_{v+1}$. Then there exists an (n-v)-dimensional subspace $W \in R^n$ such that $w \in W \Rightarrow Bw = 0$. Then

$$||Aw||_2 = ||(A - B)w||_2 \leq ||A - B||_2 ||w||_2 < \sigma_{v+1}||w||.$$

But there is a (v+1)-dimensional subspace where $||Aw|| \geq \sigma_{v+1}||w||$, namely the space spanned by the first v+1 right singular vector of A. Since the sum of the dimensions of these two spaces exceeds n, there must be a nonzero vector lying in both, and this is a contradiction.

Second part:
We must now show $||A - Av||_2 = \sigma_{v+1}$. Let $\Sigma_v = U(A - A_v)V^T$.

$$\begin{aligned} \Sigma_v &= U(diag(\sigma_1, ..., \sigma_v, \sigma_{v+1}, ..., \sigma_p) \text{ - } diag(\sigma_w, ...\sigma_v, 0, ..., 0))V^T \\ &= Udiag(0, ..., 0, \sigma_{v+1}, ..., \sigma_p)V^T \\ &\Rightarrow ||A - A_v||_2 = ||\Sigma_v||_2 = \sigma_{v+1}. \end{aligned}$$

By mapping the vectors of X into a lower dimensional space, we can now treat the term and document vectors as a "semantic space". The lower dimensional vectors in this semantic space are approximations of the higher dimensional counterparts. The reduction of dimensionality takes out irrelevant noise of the body of text, which inherently has great dimensionality due to the complexity of semantics.

III. Exploring Word Similarity

With these tools, we can compare two sentences at their core level. We can find their correlation by dotting Σ_k and d_j and Σ_k and d_p for two vectors $d_j, d_p \in V$, and finding the angle between them. This is proved below. We multiply the document vectors by the singular values to favor the index values in the matrix V that correspond to the highest singular values, i.e. the most significant topics within the document.

Definition 2.1: The inner product $< \cdot, \cdot >$ is a symmetric, positive-definite bilinear form, so:

$$< x, y >=< y, x >$$
$$< ax, y >= a < x, y > \text{ for some } a \in R$$
$$< x, x >= 0 \text{ if and only if } x = 0$$
$$< x + y, y >=< x, y > + < y, y >.$$

Lemma (a): Let $V = (X_1, X_2, ...X_n)$. If $X_1, ..., X_n$ are probabilistically independent, then the covariance of a matrix X is an inner product on the matrix X.

Proof: Covariance is a symmetric, bilinear form. The positive definite property comes from the fact that $Cov(X, X) = E[(x - \mu)^2] \geq 0$. $Covar(X, X) = 0 \iff X = 0$ follows from the independence of $X_1, ..., X_n$.

Lemma (b): As such, the correlation between to vectors is the angle between them.

Proof:

$$\text{By definition, } Corr(X, Y) = Covar(X, Y)/SD(X)SD(Y).$$
$$||X|| = \sqrt{Covar(X, X)} = \sqrt{Var(X)} = SD(X)$$
$$\Rightarrow Corr(X, Y) = Covar(X, Y)/||X||||Y|| = cos(\theta)$$

Theorem 2.1: $Corr(t_i, t_p) = cos^{-1}(< t_i, t_p > /||t_i||||t_p||)$. Likewise, $Corr(d_i, d_p) = cos^{-1}(< d_i, d_p > /||d_i||||d_p||)$.

Proof: Follows from Lemma a and Lemma b.

Corollary: $sup_{x,y \in X} Corr(x,y) = 0$ and $min_{x,y \in X} Corr(x,y) = \pi/2$.

III. Exploring Text Summarization

Another application of SVD-LSA is text summarization. However we must take a slightly different approach to this. When applied to text summarization, there are two disadvantages to the previously mentioned SVD approach. First, we have to use the same number of dimensions as the number of sentences we want to use for a summary. The higher the number of dimensions of reduced space, the less significant the topic we take into summary. The second disadvantage is that a sentence with large index values but not the largest, will not be in the summary, although its content for the summary is very suitable. To combat these disadvantages, we collect the vectors s_k where

$$s_k = \sqrt{< v_k, \sigma_i >}$$

$s = [s_1, s_2, ...s_n]$ is the vector of the norms of the kth sentence vector in the semantic space. This value is independent of the number of summary sentences. As before, we dot the document vectors by the singular values to favor the index values in the matrix V that correspond to the highest singular values, i.e. the most significant topics within the document. The summary consists of the sentences with the highest values in the vector s.