Lecture 26. Singular Value Decomposition (SVD)

Form

\[ A = U\Sigma V^T \]

\( U \) and \( V \) have orthonormal columns and \( \Sigma \) is diagonal

Transformation

\[ \begin{array}{c}
\text{VT} \\
\Sigma \\
U
\end{array} \]

Figure 1: Demonstrating the steps of SVD components’ transformation

Reasons to be interested in it:

1) \( V^T \) stationary probability of random walk
2) Finding subspace that minimizes sum of square of distance to set of points

Fitting data

The task is to find the subspace that minimizes the sum of squares of distance to a set of points. Our subspace will be the line of the form \( y = f(x) = ax + b \).

This is useful when dealing with word vectors. A word vector is a data structure in which vectors represent documents and elements in the vectors correspond to the frequencies of words in the documents.
When doing linear regression by minimizing mean squares errors, we need to minimize
\[
\sum_{i=1}^{n} [y_i - (ax_i + b)]^2
\]
This can be solved by differentiating \( \frac{\partial}{\partial a} \) and \( \frac{\partial}{\partial b} \).

\[\text{Figure 2: Fitting by minimizing square errors}\]

However, our problem is slightly different than conventional linear regression. We force \( f(x) \) to go through the origin and measure the distance perpendicular to the line rather than vertically.

\[\text{Figure 3: Best fit subspace by minimizing distance}\]
To minimize absolute values or square of norms have different implications. This could be a potential homework problem. The key lies in the treatment of the larger errors relative to the smaller ones.

To find the best fit subspace:

\[
\{(x_i, y_i) | 1 \leq i \leq n\}
\]

\[\]
the same subject will have essentially the same high value. All of the comparisons between documents on different subjects will have a value of about zero.

\[
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

If you consider each row as a point in space, you can create a best fit line through the data. All of the points will cluster around three points, one of which corresponds to each subject. The best fit “line” through the origin for this data will help us pull out the cluster of documents.

Then for example, we would have a vector \( \overrightarrow{v_1} = (1,1,\ldots,0,0,\ldots) \), that runs through the best fit line and \( \overrightarrow{v_2} \perp \overrightarrow{v_1} \) that minimizes the sum of squares of the \( d_i \).

This is exactly what SVD, singular value decomposition does.

**Minimizing distances and SVD**

Let \( A \) be an \( n \times d \) matrix in which the rows are points in \( d \) dimensions. Let \( v \) be the direction of the best fit line through the origin and normalize \( v \) such that \( |v| = 1 \).

The length of the projection of \( i^{th} \) row is \( |a_i \cdot v| \)

\[
\sum_{i=1}^{n} l_i^2 = \sum_{i=1}^{n} |a_i \cdot v|^2 = |Av|^2
\]

Find \( v_1 = \arg \max |Av|^2, \quad |v| = 1 \) (constraint)

(i.e. the unit length vector \( v_1 \) such that \( |Av|^2 \) is maximized)

\( v_1 \) is the first right singular vector.

\( \sigma_1 = |Av_1|^2 \) is the first singular value.
\[ v_2 = \arg \max \|Av_i\|^2, \quad \|v\| = 1, \quad v_2 \perp v_i \text{ (constraints)} \]

\( v_2 \) is the second right singular vector.

\[ u_i = \frac{1}{\sigma_i} Av_i \] is the \( i \)th left singular vector.

This process gives us \( u_i, \sigma_i, v_i \) which form \( U, \Sigma \) and \( V \) matrices in the singular value decomposition.

**Proof of Claim**

To validate this, we need to prove two things:

1) The \( u_i \) vectors are orthogonal

2) \[ A = \sum \sigma_i u_i v_i^T \]

As it is trivial to see \( v_i \) are orthonormal.

Proof: We prove them by induction. For the base case, where it’s one dimension, 1) is trivially true. Only 1 \( u_i \) exists. Then, we need to prove 2). In the next lecture, we will continue the proof, using the following lemma:

**Lemma**: To prove \( A = B \), it suffices to show that for all \( v \), \( Av = Bv \).

(END OF LECTURE)