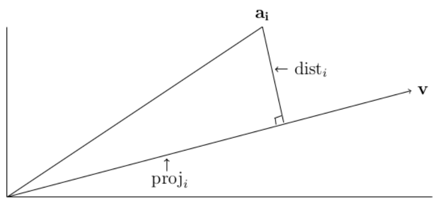
Seminar Theoretische Informatik 14.11.2017

Best-Fit Subspaces and Singular Value Decomposition (SVD)

Michaela Borzechowski *Wolfgang Mulzer*

1 The projection of a point to a line and its properties



The following holds by the Pythagorean Theorem: $(dist_i)^2 = a_{i1}^2 + a_{i2}^2 + ... + a_{id}^2 - (proj_i)^2$
Minimizing $\sum_{i=0}^{n}(dist_i)^2$ is equivalent to maximizing $\sum_{i=0}^{n}(proj_i)^2$.

2 Best-Fit subspace

The best-fit line problem describes the problem of finding the best line for a set of data points, where the quality of the line is measured by the sum of squared (perpendicular) distances of the points to the line.

This can be transferred to higher dimensions: One can find the best-fit d-dimensional subspace, so the subspace which minimizes the sum of the squared distances of the points to the subspace, which is equivalent to maximizing $\sum_{i=0}^{n}(proj_i)^2$.

2.1 Singular Vector

Let A be a $n \times d$ matrix, where the n rows are d-dimensional points. A singular vector v of A is a unit vector along the best-fit line through the origin for the points of A.

Then $proj_i$ is the length of the projection of the i'th row of A (a_i) onto v: $proj_i = |a_i \cdot v|$. And $\sum_{i=0}^{n}(proj_i)^2 = |Av|^2$.

Example 2.1 $A = \begin{pmatrix} 3/5 & 4/5 \\ 6 & 8 \\ 3 & 4 \end{pmatrix}$, first singular vector $v_1 = \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix}$

$proj_1 = |a_1 \cdot v_1| = (3/5, 4/5) \cdot \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix} = 1$

$proj_2 = |a_2 \cdot v_2| = (6, 8) \cdot \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix} = 10$

$proj_3 = |a_3 \cdot v_3| = (3, 4) \cdot \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix} = 5$

$|Av_1|^2 = \left| \begin{pmatrix} 1 \\ 10 \\ 5 \end{pmatrix} \right|^2 = \sqrt{1^2 + 10^2 + 5^2}^2 = 1^2 + 10^2 + 5^2 = \sum_{i=0}^{n} (proj_i)^2$

$|Av_1| = 11,225$

2.2 Find a first singular vector

To find a first singular vector (also called right singular vector) v_1 of A:

$$v_1 = arg \max_{|v|=1} |Av| \Leftrightarrow |Av_1| = \max_{|v|=1} |Av|$$

Note that there may be several maximal v_1's, for example $-v_1$ is as good as v_1.

2.3 First singular value

The first singular value $\sigma_1(A)$ is defined as $\sigma_1(A) = |Av_1|$.

2.4 How to find the best-fit r-dimensional subspace the greedy way (Theorem 3.1)

First, find v_1.

Then find a unit vector v_2 perpendicular to v_1 that maximizes $|Av|^2$:

$$v_2 = arg \max_{v \perp v_1, |v|=1} |Av|$$

Repeat that for v_r perpendicular to $v_{k-1}, v_{k-2}, ...v_1$ until

$$arg \max_{v \perp v_1, v_2, ..., |v|=1} |Av| = 0$$

Example 2.2 Assume all three dimensional points of a matrix A lie in a plane, then the best-fit subspace is a 2-dimensional plane. The first singular vector is the best-fit line and the second singular vector is perpendicular to the first one, spanning a plane. The third singular vector must be perpendicular to the other two, but $proj_i$ of every point i is 0.

This is proved via induction.

2.5 Right- and left singular vectors

$v_1, v_2, ..., v_r$ are called *right singular vectors.* A *left singular vector* u_i is defined as follows, where $|u_i| = 1$

$$u_i = \frac{1}{\sigma_i(A)} Av_i = \frac{1}{|Av_i|} Av_i$$

Example 2.3 $v_1 = \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix}$, $Av_1 = \begin{pmatrix} 1 \\ 10 \\ 5 \end{pmatrix}$, $|Av_1| = 11,225$

$u_1 = \frac{1}{11,225} \cdot \begin{pmatrix} 1 \\ 10 \\ 5 \end{pmatrix} = 0,089 \cdot \begin{pmatrix} 1 \\ 10 \\ 5 \end{pmatrix} = \begin{pmatrix} 0,089 \\ 0,89 \\ 0,445 \end{pmatrix}$, $|u_1| = 1$

3 Singular value decomposition

Idea: To represent a matrix by its singular vectors and values. We will see applications of that next week in the second part of chapter 3.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T = \sum_{i=1}^{r} |Av_i| \cdot \frac{1}{|Av_i|} Av_i \cdot v_i^T = \sum_{i=1}^{r} Av_i v_i^T \qquad \text{(Theorem 3.4)}$$

A different way to write that equation is to define U, V^T and D:

- Let U be a $n \times r$ matrix where u_i is the i'th column.
- Let V^T be a $r \times d$ marix where v_i^T is the i'th row of V^T.
- Let D be a $r \times r$ matrix where σ_i is the i'th entry on its diagonal.

Then $A = UDV^T = AVV^T$.

Example 3.1 $A = \begin{pmatrix} 3/5 & 4/5 \\ 6 & 8 \\ 3 & 4 \end{pmatrix}$, $v_1 = \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix}$, $u_1 = \begin{pmatrix} 0,089 \\ 0,89 \\ 0,445 \end{pmatrix}$

$U = \begin{pmatrix} 0,089 \\ 0,89 \\ 0,445 \end{pmatrix} = u_1$

$V^T = (3/5, 4/5)$

$D = (11,225)$

$A = UDV^T = \begin{pmatrix} 0,089 \\ 0,89 \\ 0,445 \end{pmatrix} \cdot (11,225) \cdot (3/5, 4/5) = \begin{pmatrix} 1 \\ 10 \\ 5 \end{pmatrix} \cdot (3/5, 4/5) = \begin{pmatrix} 3/5 & 4/5 \\ 6 & 8 \\ 3 & 4 \end{pmatrix} = A$

3.1 Why does that work?

$$AVV^T = \begin{pmatrix} a_{11} & ... & a_{1d} \\ ... & & ... \\ a_{n1} & ... & a_{nd} \end{pmatrix} \cdot \begin{pmatrix} v_{11} & ... & v_{r1} \\ ... & & ... \\ v_{1d} & ... & v_{rd} \end{pmatrix} \cdot \begin{pmatrix} v_{11} & ... & v_{1d} \\ ... & & ... \\ v_{r1} & ... & v_{rd} \end{pmatrix} = \begin{pmatrix} a_1 \cdot v_1 & ... & a_1 \cdot v_r \\ ... & & ... \\ a_n \cdot v_1 & ... & a_n \cdot v_r \end{pmatrix} \cdot \begin{pmatrix} v_{11} & ... & v_{1d} \\ ... & & ... \\ v_{r1} & ... & v_{rd} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{r} v_{j1}(a_1 \cdot v_j) & ... & \sum_{j=1}^{r} v_{jd}(a_1 \cdot v_j) \\ ... & & ... \\ \sum_{j=1}^{r} v_{j1}(a_n \cdot v_j) & ... & \sum_{j=1}^{r} v_{jd}(a_n \cdot v_j) \end{pmatrix} = A$$

The following holds for every orthonormal basis: $a_i = \sum_{j=1}^{r} (a_i \cdot v_j) v_j$. Since $v_1, ..., v_r$ are an orthonormal basis, the SVD works.

4 Power Method for SVD

To find the SVD for a matrix A, the singular vectors need to be calculated.

This is a hard thing to calculate for higher dimensions. There is no algebraic solution to polynomia equations with degree ≥ 5 (Abel-Ruffini theorem), so calculating the x_i for which the derivative is 0 may not be possible.

Therefore we need to approximate v_i. The power method approximates the SVD in polynomial time.

Let B be the square and symmetric matrix $B = A^T A$.

$$B^k = \sum_{i=1}^{r} \sigma_i^{2k} v_i v_i^T$$

If $\sigma_1 > \sigma_2$ and for a big enough k, the first column of B^k is a good approximation of v_1.

The problem with this method is that we need to do many matrix multiplications with big matrices.

4.1 Faster method

Instead of computing B^k, compute $B^k x$ for a random vector x represented by the orthonormal basis $v_1, ..., v_r$: $x = \sum_{i=1}^{d} c_i v_i$.

$B^k x \approx (\sigma_1^{2k} v_1 v_1^T)(\sum_{i=1}^{d} c_i v_i) = \sigma_1^{2k} c_1 v_1$

Calculating $B^k x$ is easier than just calculating B^k and results in a vector, which is v_1 when it is normalized.

5 Best Rank k Approximations

Idea: We want to find the best-fit subspace with rank k of matrix A. It might or might not include all points, but it is the one “closest” to A. “closest” can be defined by different norms, in this case by the *Frobenius norm* or the *2-norm*.

Best rank k approximation $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

Frobenius norm: $||A||_F = \sqrt{\sum_{j,k} a_{jk}^2}, \qquad ||A||_F^2 = \sum \sigma_i^2(A)$

2-norm: $||A||_2 = \max_{|x| \leq 1} |Ax|$

Theorem 3.6: For any matrix B of rank at most k holds $||A - A_k||_F \leq ||A - B||_F$

Theorem 3.9: For any matrix B of rank at most k holds $||A - A_k||_2 \leq ||A - B||_2$