Lecture 25 Notes: SVD continued March 27th, 2009

Yuzhe Liu yl435 & Ramu N. rn54

From Previous Lecture:

SVDs can be used to cluster documents based on their similarity. Documents can be summarized by creating a vector with the counts of words in the document. If a word i appears zero times in the document, then entry i in the vector will be zero. The similarity of any two documents can be computed by taking the dot product of their normalized vectors.

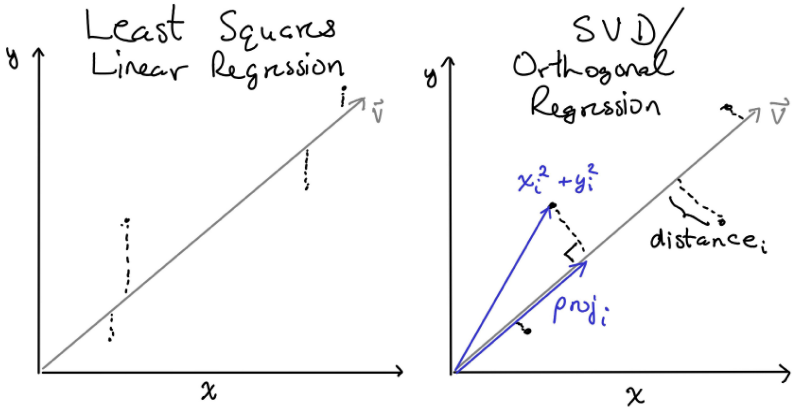
This Lecture:

SVDs or Singular Value Decomposition is one of the best ways to factorize a matrix. It allows us to decompose a matrix A into U & V, two orthonormal matrices and Σ, a diagonal matrix. The rows of matrix A can be considered points in space. Our goal is to find a k dim subspace that is a best fit.

The best fit minimizes sum of square distances of all the points to the subspace.

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

To start with lets begin with the simpler problem where we restrict ourselves to 2 dimensions and hence the subspaces of lines thru the origin.



Our goal is to minimize the distance from the subspace to each of the points, a_i. This can be expressed in the formula below where our goal is to minimize the left hand side of the equation.

$$\sum_i (distance_i)^2 = \sum_i (x_i^2 + y_i^2) - \sum_i (proj_i)^2$$

In the 2D case diagrammed above, this corresponds to minimizing the distance of points from the regression line. This is slightly different problem than regular linear regression where we are only trying to minimize the total vertical distance. From a statistics perspective this corresponds to either assuming that there is error in the measurement of both the x and the y coordinates or just in the y coordinate.

$\sum_i (x_i^2 + y_i^2)$ is a constant since all the points stay the same. Only the vector v changes. Hence, instead of trying to minimize the distances of the points from the line, we can attempt to maximize the projection of the points onto the vector v as below:

$$\sum_i (proj_i)^2 = \sum_i (x_i^2 + y_i^2) - \sum_i (distance_i)^2$$

The projection of a point i onto the vector v, $proj_i$, can be determined by the equation$proj_i = |a_i v|$ where a_i is a row vector and v is column vector. This means that our maximization function can be written as $\sum_i |a_i v|$. This can be written as Av taking advantage of the speed of matrix multiplication.

We can define the (right) singular vector

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

which has a corresponding singular value:

$$\sigma_1 = |Av_1|$$

The remaining singular values and vectors are defined as follows:

$$\begin{aligned}
v_2 &= \arg\max_{v_2 \perp v_1, |v_2|=1} |Av| \\
\sigma_2 &= |Av_2| \\
v_3 &= \arg\max_{v_3 \perp v_2, v_1, |v_3|=1} |Av| \\
\sigma_3 &= |Av_3| \\
&\vdots \\
v_r &= \arg\max_{v_r \perp v_1, v_2 \ldots v_{r-1}, |v_r|=1} |Av| \\
\sigma_r &= |Av_r|
\end{aligned}$$

Note that each singular vector is orthogonal to all the earlier vectors. This means that the maximum number of nonzero singular values that can be found is r, the rank of the matrix where

$r \leq \min(r, c)$ where r and c are the rows and columns in A respectively. Hence, the following equation holds:

$$\max_{v \perp v_1, v_2 \ldots v_r} |Av| = 0$$

The left singular vector, u_i can be computed from the corresponding right singular vector, v_i and the singular value σ_i.

$$u_i = \frac{1}{\sigma_i} A v_i$$

0.1 Theorem

1. $u_1, u_2, \ldots u_r$ are orthogonal.

2. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \Rightarrow A = U \Sigma V^T$

Proof by induction:

Base case:

For the case $r = 1$, 1. is trivially true.
2. $A = \sigma_1 u_1 v_1^T$: Need to show for all v that $Av = \sigma_1 u_1 v_1^T v$.
$v = av_1 + wwherew \perp v_1$
$Av = A(av_1 + w)$ But Aw is 0, since there were no more vectors perpendicular to v_1.
$Av = aAv_1$
$\sigma_1 u_1 v_1^T v = \sigma_1 u_1 v_1^T (av_1 + w) = a\sigma_1 u_1$ ($w \perp v_1$, so $v_1^T = 0$)
$= a\sigma_1 \frac{1}{\sigma_1} Av_1 = aAv_1$
Thus, $A = \sigma_1 u_1 v_1^T$.

Inductive step:

$B = A - \sigma_1 u_1 v_1^T$
$B = v_2 \ldots v_r; \sigma_2 \ldots \sigma_r; u_2 \ldots u_r$
$B = \sum_{i=2}^{r} \sigma_i u_i v_i^T$, when $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

$v_i = argmax_{v \perp v_1 \ldots v_{i-1}; |v|=1} |Av|$ where $v = av_1 + w$
$argmax_{|v|=1}(A - \sigma_1 u_1 v_1^T)v = argmax_{a;w;|v|=1}(A - \sigma_1 u_1 v_1^T)(av_1 + w)$
$= argmax_{w \perp v_1}(aAv_1 + Aw - a\sigma_1 u_1) = v_2$

Following this procedure, we can see that $u_1, u_2, \ldots u_r$ are orthogonal by the induction hypothesis.

Without loss of generality, assume $u_1^T u_i > 0$.
$|A(\frac{v_1 + \epsilon v_i}{\sqrt{1+\epsilon^2}})| = |\frac{1}{\sqrt{1+\epsilon^2}}(Av_1 + \epsilon Av_i)| = |\frac{1}{\sqrt{1+\epsilon^2}}(\sigma_1 u_1 + \epsilon \sigma_i u_i)|$, which is at least as long as component along u_1, $= |\frac{\sigma_1 + \epsilon \sigma_i u_1^T u_i}{\sqrt{1+\epsilon^2}}| = |(\sigma_1 + \epsilon \sigma_i u_1^T u_i)(1 - \frac{\epsilon^2}{2} + O(\epsilon^4))| = |\sigma_1 + \epsilon \sigma_i u_1^T u_i - O(\epsilon^2)| > \sigma_1$

This is a contradiction.