HW 4

Bayard Walsh

May 2023

1

1.1

We apply rank 1 matrix decomposition formula as follows *(note that we only sum to k as rank(X)=k)*

$$X = \Sigma_{i=1}^{k}(u_i \sigma_i v_i^T)$$

1.2

First, as we want the best 1d subspace, we can achieve this through minimizing the sum of squared linear projections. As we know that V^T gives the principal component directions of X, and we know that v_1 is the first principle component, we also know that v_1 is the largest value in V (by first principle component definition) so v_1 will minimize the sum of squared linear projections for a 1 d subspace and therefore offer the best 1 d subspace for the data. Therefore we have v_1 as our basis.

1.3

As we have $X = U\Sigma V^T$ as its full SVD, we will derive the following SVDs.

$$X^T$$
$$(U\Sigma V^T)^T$$
$$V\Sigma^T U^T$$

Therefore the SVD is as follows

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

Next

$$XX^T$$
$$(U\Sigma V^T)(V\Sigma^T U^T)$$
$$U\Sigma V^T V\Sigma^T U^T$$

by orthnormality of V

$$U\Sigma I\Sigma^T U^T$$

$$U\Sigma\Sigma^T U^T$$

Therefore the SVD is as follows

$$XX^T = U\Sigma\Sigma^T U^T$$

Next

$$X^T X$$

$$(V\Sigma^T U^T)(U\Sigma V^T)$$

$$V\Sigma^T U^T U\Sigma V^T$$

$$V\Sigma^T I\Sigma V^T$$

$$V\Sigma^T\Sigma V^T$$

Therefore the SVD is as follows

$$X^T X = V\Sigma^T\Sigma V^T$$

1.4

1.4.1

Note that by SVD nature, if X is a matrix with rank k, the first k rows of V (in the SVD) will span the basis of X, whereas the remaining rows (or rows $k+1\cdots p$ of V) will all cause $X \cdot v_j = 0$ for any j in $k+1\cdots p$ of V. Because every vector in V is orthonormal to each other, any vector w where $Xw = 0$ will be spanned by the basis $v_{k+1}\cdots v_p$ of V, so $v_{k+1}\cdots v_p$ of V is the basis of all weight vectors that could satisfy $Xw = 0$.

1.4.2

We will consider the least squares optimization, which is

$$\min_w ||y - Xw||_2^2$$

I will now show that minimizing

$$||\tilde{y} - \Sigma\tilde{w}||_2^2$$

is equivalent. To minimize this term we must minimize w, so we have

$$\min_w ||\tilde{y} - \Sigma\tilde{w}||_2^2$$

Next

$$\min_w ||U^T y - \Sigma V^T w||_2^2$$

$$\min_w ||UU^T y - U\Sigma V^T w||_2^2$$

by orthnormality

$$\min_w ||y - U\Sigma V^T w||_2^2$$

as we have

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

we have

$$= \min_w ||y - Xw||_2^2$$

This relation proves that the two terms are equivalent.
The minimum norm solution is the

$$\min ||\tilde{w}||_2^2$$

such that $y = X\tilde{w}$. However, as X is not full rank and the pseudo inverse is unique, the solution for $\tilde{w}$ that follows the relation above must be unique (and therefore the minimimum solution), so we solve for

$$\tilde{w}$$

using the terms from above. Therefore we have

$$\tilde{w} = V^T w$$

least squares substitution

$$\tilde{w} = V^T (X^T X)^{-1} X^T y$$
$$\tilde{w} = V^T ((U\Sigma V^T)^T U\Sigma V^T)^{-1} (U\Sigma V^T)^T y$$
$$\tilde{w} = V^T (V\Sigma^T U^T U\Sigma V^T)^{-1} V\Sigma^T U^T y$$
$$\tilde{w} = V^T (V\Sigma^T I\Sigma V^T)^{-1} V\Sigma^T U^T y$$
$$\tilde{w} = V^T V (\Sigma^T \Sigma)^{-1} V^T V\Sigma^T U^T y$$
$$\tilde{w} = (\Sigma^T \Sigma)^{-1} \Sigma^T U^T y$$

1.4.3

As mentioned above the solution is unique, therefore the optimal solution with the smallest norm is the solution for w from above. Therefore, we have the following:

$$\tilde{w} = V^T w$$

from previous derivation

$$(\Sigma^T \Sigma)^{-1} \Sigma^T U^T y = V^T w$$
$$V(\Sigma^T \Sigma)^{-1} \Sigma^T U^T y = VV^T w$$

Therefore the minimum least squares solution is as follows:

$$w = V(\Sigma^T \Sigma)^{-1} \Sigma^T U^T y$$

2

2.1

The predictor is performing poorly. Note that the estimated best joke is joke 35 from our model, with a ranking of 7.396267570675075, whereas the estimated score for 29 (the known highest joke) is 3.515461743382937. Note that 35 was the 27th highest ranked joke (in the true rankings) so our estimate is not very good. Furthermore, in the graph (prediction of unknown ratings trained with 20 users) shows a very non correlated distribution while the true rating has a very predictable increasing behavior

2.2

Now that we have an undetermined data set, we can apply SVD decomposition (as $n > p$ previously SVD prerequisites were not met). By using SVD, we have an almost perfect training set error, with $1.86193636e-15$ error rate, however the test rate is still scattered with 0.40349347 error rate, compared to the initial 0.61913973 error rate from the over determined data set, so we have improved but not by much. Note that the standard SVD formula for least squares is applied, but with X changed to all data. As we derived w earlier from SVD, the formula from 1.4.3 is applied

2.3

First I generate the *diffmatrix* data set, which is the square difference between our model outcome ratings and the actual ratings for a given user. From here, I find the lowest difference and second lowest difference indexes, which will be the two best users for the model, as they resemble the new user the most. After implementing our model where we restrict our test data to the two first users, we achieve a training error of 0.82588456 and a test error of 0.50103103. Note that this is worse than the SVD application offered in 2.2, (especially in its training data), however it has slightly better performance than the over-determined test error rate (which has very poor performance). Therefore this is not an improvement from our previous model.

2.4

Spectrum computed by taking SVD of X and graphing s values. Matrix X has rank 100. Therefore none of its columns are *completely* linearly independent, however based on viewing the values of s from the SVD, there is a sharp drop off after the first 5 entries, with each entry being less than 600 afterwards, which is a steep decline from the first 5 entries. Therefore, based on the composition of the SVD matrix, we can use 5 or so dimensions to represent the majority of the information of matrix X. This suggests that we can decrease the amount of dimensions we are working in and still have a quality estimate of X data through a linear combination of the 5 highest spectrum values

2.5

First we calculate the SVD, then reduce S to its first three columns and V^T to its first three rows. From here, a dot product between these reduced terms will project the first three principal component directions. In the graph, it is noticeable that the amount of variation roughly aligns with $PCA1 > PCA2 > PCA3$, as we are reducing the amount of variation with each PCA.

2.6

In order to show how the power method works, I will show equivalence between the power method and the first eigenvalue of X. As we have $A = XX^T$ by the power method, let us have u to be the first eigenvector of A and λ to be the first eigenvalue of A. Assume that $A_1 = u$, or let the first eigenvector of A be u. Now by eigenvector property and assumption,

$$A \cdot u = \lambda u$$

$$XX^T \cdot u = \lambda u$$

SVD equivilance substitution

$$U\Sigma V^T (U\Sigma V^T)^T \cdot u = \lambda u$$

$$U\Sigma V^T V \Sigma^T U^T \cdot u = \lambda u$$

$$U\Sigma\Sigma^T U^T \cdot u = \lambda u$$

Note that by eigenvector property, we have $U^T u = v_{eigen}$, where v_{eigen} is the first eigenvector

$$U\Sigma\Sigma^T v_{eigen} = \lambda u$$

$$\lambda u u \Sigma\Sigma^T = \lambda u$$

$$\lambda u u \cdot \sigma_1^2 = \lambda u$$

$$\lambda u = \lambda u$$

Note that we also have $\sigma_1^2 = \lambda$ in our calculation. There is a difference in 0.0005179300337658788, so the values are essentially the same (considering the margin cutoff we gave of .0001), given we ignore the sign, though the power method resulted in a different sign than the traditional method. This is because the power method finds the eigenvalue with the largest magnitude, which may be a negative eigenvalue.

2.7

Starting with the zero vector will always cause the power method to fail, because in its process of finding the first left and right singular vectors, the power method multiplies the matrix with a starting vector and normalizing the result to approximate the dominant singular vector, however if it starts with the zero vector, it will never converge to any non-zero vector.

3

3.1

implemented in code

3.2

Generally, there appeared to be no correlation to optimal lambda based on k values, however as sigma increased a higher lambda became more optimal. These correlations are demonstrated in graphs included.