Data Science Assignment 4

Nike Marie Pulow – Henri Paul Heyden

stu239549 – stu240825

1 PCA, SVD

1. We search for matrices U, Σ, V with $X = U\Sigma V^T$ so that the columns of U and V are eigenvectors of XX^T and X^TX respectively.
We'll begin by computing XX^T:

$$\begin{aligned} XX^T &= \begin{bmatrix} 4 & 2 \\ \sqrt{2} & 2\sqrt{2} \\ -\sqrt{2} & -2\sqrt{2} \end{bmatrix} \times \begin{bmatrix} 4 & \sqrt{2} & -\sqrt{2} \\ 2 & 2\sqrt{2} & -2\sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} \end{aligned}$$

Now we have to solve $\det(X - \lambda I_{3\times 3}) = 0$ for the eigenvalues of XX^T.
An expanded form of this equation is:

$$\begin{aligned} 0 &= \det(X - \lambda I_{3\times 3}) \\ &= \det\left(\begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right) \\ &= \det\left(\begin{bmatrix} 20-\lambda & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10-\lambda & -10 \\ -8\sqrt{2} & -10 & 10-\lambda \end{bmatrix} \right) \\ &= (20-\lambda)(10-\lambda)(10-\lambda) + (8\sqrt{2})(-10)(-8\sqrt{2}) + (-8\sqrt{2})(8\sqrt{2})(-10) \\ & \quad -(-8\sqrt{2})(10-\lambda)(-8\sqrt{2}) - (8\sqrt{2})(8\sqrt{2})(10-\lambda) - (20-\lambda)(-10)(-10) \end{aligned}$$

This is a third degree polynomial equation with expanded form $-x^3 + 40x^2 - 144x = 0$ [1]
We get the solutions:

$$\lambda_1 = 36 \wedge \lambda_2 = 4 \wedge \lambda_3 = 0$$

With this we have

$$\Sigma = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

With the eigenvalues of XX^T being calculated, we can now solve for the eigenvectors of XX^T:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} &= \left(\begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} - \begin{bmatrix} 36 & 0 & 0 \\ 0 & 36 & 0 \\ 0 & 0 & 36 \end{bmatrix} \right) \times \begin{bmatrix} v_{1_1} \\ v_{1_2} \\ v_{1_3} \end{bmatrix} \\ &= \begin{bmatrix} -16 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & -26 & -10 \\ -8\sqrt{2} & -10 & -26 \end{bmatrix} \times \begin{bmatrix} v_{1_1} \\ v_{1_2} \\ v_{1_3} \end{bmatrix} \\ &= \begin{bmatrix} -16v_{1_1} + 8\sqrt{2}v_{1_2} - 8\sqrt{2}v_{1_3} \\ 8\sqrt{2}v_{1_1} - 26v_{1_2} - 10v_{1_3} \\ -8\sqrt{2}v_{1_1} - 10v_{1_2} - 26v_{1_3} \end{bmatrix} \end{aligned}$$

[1] Actually writing each step of this down would take too much space.

When numbering the equations in collumns **I**, **II** and **III** we can see that **I** $\Rightarrow v_{1_1} = \frac{\sqrt{2}v_{1_2}}{2} - \frac{\sqrt{2}v_{1_3}}{2}$ and **II** + **III** $\Rightarrow v_{1_2} = -v_{1_3}$ which results in:

$$v_{1_1} = \sqrt{2}v_{1_2}$$

and respectively:

$$v_{1_2} = \frac{v_{1_1}}{\sqrt{2}} \land v_{1_3} = -\frac{v_{1_1}}{\sqrt{2}}$$

There are infinite solutions for this, with one being:

$$v_{1_1} = 1 \land v_{1_2} = \sqrt{2}^{-1} \land v_{1_3} = -\sqrt{2}^{-1}$$

So we get one eigenvector:

$$v_1 = \begin{bmatrix} 1 \\ \sqrt{2}^{-1} \\ -\sqrt{2}^{-1} \end{bmatrix}$$

Because writing the processes down takes a lot of space, the following linear equation systems will not be solved, but solutions provided. We have already demonstrated that we can solve one and the probability of minor mistakes when writing it down ruining the whole task is too great.
The second eigenvector we get by this time subtracting the diagonal matrix times 4 instead of 36 and we get a solution:

$$v_2 = \begin{bmatrix} 1 \\ -\sqrt{2}^{-1} \\ \sqrt{2}^{-1} \end{bmatrix}$$

And the third eigenvector is quite trivial:

$$v_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Something that wasn't specified in the slides was that the eigenvectors must be normalized so that U and V are orthonormal.
For that we have to divide all these vectors by $\sqrt{2}$.
We get:

$$U = \begin{bmatrix} \sqrt{2}^{-1} & \sqrt{2}^{-1} & 0 \\ 1/2 & -1/2 & \sqrt{2}^{-1} \\ -1/2 & 1/2 & \sqrt{2}^{-1} \end{bmatrix}$$

For a quick verification, we can compute UU^T and yes we get $I_{3\times 3}$, so U is indeed orthonormal.
Now to V
As the lecture specifies, the eigenvalues of XX^T are the same like those of X^TX.
This means we „just" have to solve the linear equations but for X^TX and not XX^T.
We'll begin by computing X^TX:

$$\begin{aligned} X^TX &= \begin{bmatrix} 4 & \sqrt{2} & -\sqrt{2} \\ 2 & 2\sqrt{2} & -2\sqrt{2} \end{bmatrix} \times \begin{bmatrix} 4 & 2 \\ \sqrt{2} & 2\sqrt{2} \\ -\sqrt{2} & -2\sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 20 & 16 \\ 16 & 20 \end{bmatrix} \end{aligned}$$

Thank god that this time it's small and regular.
We get the following systems of linear equations: With eigenvalue 36 we get:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} -16 & 16 \\ 16 & -16 \end{bmatrix} \times \begin{bmatrix} v_{1_1} \\ v_{1_2} \end{bmatrix} \\ &= \begin{bmatrix} 16v_{1_1} - 16v_{1_2} \\ 16v_{1_1} - 16v_{1_2} \end{bmatrix} \end{aligned}$$

With one solution being:

$$v_1 = \begin{bmatrix} \sqrt{2}^{-1} \\ \sqrt{2}^{-1} \end{bmatrix}$$

With eigenvalue 4 we get:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} 16 & 16 \\ 16 & 16 \end{bmatrix} \times \begin{bmatrix} v_{2_1} \\ v_{2_2} \end{bmatrix} \\ &= \begin{bmatrix} 16v_{2_1} + 16v_{2_2} \\ 16v_{2_1} + 16v_{2_2} \end{bmatrix} \end{aligned}$$

With one solution being

$$v_2 = \begin{bmatrix} \sqrt{2}^{-1} \\ -\sqrt{2}^{-1} \end{bmatrix}$$

For the last eigenvalue 0 we get $v_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, which formally isn't an eigenvector, since it would be an eigenvector of every matrix, but it is the only solution to the corresponding eigenvalue.
With that we get:

$$V^T = \begin{bmatrix} \sqrt{2}^{-1} & \sqrt{2}^{-1} \\ \sqrt{2}^{-1} & -\sqrt{2}^{-1} \\ 0 & 0 \end{bmatrix}$$

And now we can calculate the product $U\Sigma V^T$ to verify that it is actually right, which I did, but to be honest, writing all that down with each addition and so on would take a lot of time.

2. TODO

2 TF-IDF

1. Vocabulary: [”fast”, ”car”, ”highway”, ”road”, ”bike”, ”wheel”]

2. term frequencies:

t	$d = 1$	$d = 2$	$d = 3$	$d = 4$
fast	0.2	0.4	0.2	0
car	0.4	0.4	0	0.25
highway	0.2	0	0.2	0
road	0.2	0	0.4	0
bike	0	0.2	0	0.25
wheel	0	0	0.2	0.5

3. document frequency:

t	$\#d$ containing t	IDF
fast	3	$\log\frac{4}{3}$
car	3	$\log\frac{4}{3}$
highway	2	$\log\frac{4}{2} = \log 2$
road	2	$\log(2)$
bike	2	$\log(2)$
wheel	2	$\log(2)$

4. calculating TF-IDF vectors:

t	$d=1$	$d=2$	$d=3$	$d=4$
fast	0.025	0.05	0.025	0
car	0.05	0.05	0	0.031
highway	0.06	0	0.06	0
road	0.06	0	0.120	0
bike	0	0.06	0	0.075
wheel	0	0	0.06	0.151

Which gives us the following vectors for each d:

$$r(d=1) = (0.025, 0.05, 0.06, 0.06, 0, 0)$$
$$r(d=2) = (0.05, 0.05, 0, 0, 0.06, 0)$$
$$r(d=3) = (0.025, 0, 0.06, 0.12, 0, 0.06)$$
$$r(d=4) = (0, 0.031, 00, 0.075, 0.151)$$