6 Questions 2h

[1] $A = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}$

(a) find λ_1, λ_2 of A^TA and their corresponding eigenvectors

(b) find S, U, V^T of SVD of $A = USV^T$. consider last column of U: $\frac{1}{\sqrt{6}}\begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$

(c) find pseudo code of A sing SVD of A.

[2] $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ $\quad \varepsilon_i \sim N(0,1)$

x_i							
y_i							

(values of both)

marks:
(3) (a) determine $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]^T$
(3) (b) fit $\hat{\beta}_{ridge}$ where $\lambda = 1$
(1) (c) compare $\hat{\beta}, \hat{\beta}_{ridge}$ considering $\beta = \begin{bmatrix} 5 \\ 15 \end{bmatrix}$, Analyse + Comment

[3] multiarmed bandits: $k = 4$
reward 1 in each epoch for each arm
$\mu_1 = 0.3, \mu_2 = 0.15, \mu_3 = 0.55, \mu_4 = 0.7$

epoch	chosen arm in i	reward arm 1	reward arm 2	reward arm 3	reward arm 4
1					
10					

(values for each epoch.)

(a) find associated regret?
(b) is it possible to draw a sample $x = 0.9$ from beta dist associated with arm 3 within Thompson Sampling? (beta formula given)
(c) is arm 3 a good choice to pick in next epoch using Thompson Sampling? Comment + analyze

(4) [4] $E(w,b) = \frac{1}{2}\sum_{k=\mu_0}^{n}(o_k - t_k)^2$

find $\frac{\partial E}{\partial w_{ji}}$?

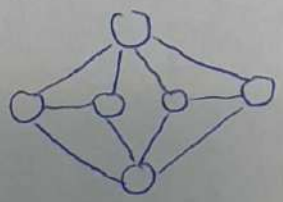
use hyperbolic tangent function as activation: $\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$\sinh(x) = \frac{e^x - e^{-x}}{2}$, $\cosh(x) = \frac{e^x + e^{-x}}{2}$

$\frac{d(\tanh(x))}{dx} = 1 - \tanh^2(x) = \frac{1}{\cosh^2(x)}$

marks
[5] (3) (a) find likelihood of λ in poisson dist?
(4) (b) find MLE?
(2) (c) MLE calculation on sample set. (numbers given)
(1) (d) what is p(5 born in hour)? (calculation)
(3) (e) prof MLE unbiased?

[6] graph:
marks
(3) (a) Calculate $L(G)$?
(6) (b) Calculate nCut, Ratiocut?
(2) (c) Comment on the results



(weights are different on each)

1

(1a) Find eigendecomposition of A^TA.

Solution: we only need the following formulas:

$A^TA\bar{v} = \lambda\bar{v}$ ← its eigenvalue, ← one vector

$(A^TA - \lambda I)\bar{v} = 0$

$\det(A^TA - \lambda I) = 0$

$$A^TA_{2\times 2} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

$$\det\left(\begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0$$

$(5-\lambda)(2-\lambda) - 4 = 0$

$\lambda^2 - 7\lambda + 6 = 0 \Rightarrow \lambda_1 = 6,\ \lambda_2 = 1$

Let's derive 1st eigenvector:

$$\left(\begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix} - \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}\right)\bar{v}_1 = 0$$

$$\begin{bmatrix} -1 & 2 \\ 2 & -4 \end{bmatrix}\bar{v}_1 = 0 \Rightarrow -v_{11} + 2v_{12} = 0 \Rightarrow v_{12} = \frac{v_{11}}{2}$$

$$\bar{v}_1 = k_1\begin{bmatrix} 2 \\ 1 \end{bmatrix} = k_1\begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

any ~~scalar~~ non-zero real value

we only need the ratio between all the components of vector, not precise numbers

Let's derive 2nd vector:

$$\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}\bar{v}_2 = 0 \Rightarrow 4\bar{v}_{21} + 2\bar{v}_{22} = 0$$

$$\Rightarrow \bar{v}_{22} = -2\bar{v}_{21}$$

$$\bar{v}_2 = k_2\begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

can also make them unit length

is that derived for eigendecomposition?

Done.

see slides of de Weijer + Wiki

(1c) Find the pseudoinverse of A using SVD of A.

We first define

$$\Sigma^+ = \begin{bmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$\sigma_1 \geq \sigma_2$

$A^+ = V\Sigma^+ U^T$ where $V = (V^T)^T$, $U^T = (U)^T$, U and V^T are the components of SVD of A, Σ^+ is the transpose of Σ with its non-zero values inverted, that is:

$$\Sigma^+ = \left(\begin{bmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \\ 0 & 0 \end{bmatrix}\right)^T = \begin{bmatrix} 1/\sqrt{5} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

← de Wiljes said in her slides that $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \dots \geq \sigma_n$, i.e., they must be sorted

$$V^T = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix} = V$$

$$U^T = \begin{bmatrix} 2/\sqrt{30} & 5/\sqrt{30} & 1/\sqrt{30} \\ 1/\sqrt{5} & 0 & -2/\sqrt{5} \\ -2/\sqrt{6} & 1/\sqrt{6} & -1/\sqrt{6} \end{bmatrix}$$

$$V\Sigma^+ = \frac{1}{\sqrt{5}} \begin{bmatrix} 2/\sqrt{5} & 1 & 0 \\ 1/\sqrt{5} & -2 & 0 \end{bmatrix}$$

error-prone and troublesome to compute by hand

$$A^+ = V\Sigma^+ U^T = \frac{1}{\sqrt{5}} \begin{bmatrix} 4/\sqrt{180} + 1/\sqrt{5} & 10/\sqrt{180} & 2/\sqrt{180} - 2/\sqrt{5} \\ 2/\sqrt{180} - 2/\sqrt{5} & 5/\sqrt{180} & 1/\sqrt{180} + 4/\sqrt{5} \end{bmatrix}$$

$\rightarrow \sqrt{180} = \sqrt{5}\sqrt{36} = 6\sqrt{5}$

taking out $\frac{1}{\sqrt{5}}$

$$= \frac{1}{5} \begin{bmatrix} 1.667 & 1.667 & -1.667 \\ -1.667 & 0.83 & 4.167 \end{bmatrix}.$$

(16)

(1) Find SVD of A. Last column of U is $\frac{1}{\sqrt{6}}\begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$

$m \times n$

Plan:

1. Determine dimensions of U, Σ, V^T. Write them down.

$m \times m$ $m \times n$ $n \times n$

2. You already have V and almost have Σ. Write them down, align Σ to its dimensions.

3. To compute U, either:
   - ~~use for~~ compute eigenvectors of AA^T, they'll be columns of U
   - use formula $AV = U\Sigma$ (Ben Hawis) ← easier

4. Attach col of U from last decomposition

NB: formulas of G. Strang are no use

Solution: 1) $U \in \mathbb{R}^{3\times 3}$, $\Sigma \in \mathbb{R}^{3\times 2}$, $V^T \in \mathbb{R}^{2\times 2}$

2) Rows of V^T are the eigenvectors of A^TA, the columns of U are the eigenvectors of AA^T.

Let's write out V using the eigenvectors of A^TA found in 2(10) and also make them unit-length vectors:

$\overline{v_1} = k_1\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\overline{v_2} = k_2\begin{bmatrix} 1 \\ -2 \end{bmatrix}$. To make them unit-length, we set $k_1 = k_2 = 1$ and adjust the values in the vectors so that

sqrt of the
the sum of squares of the vector components is 1.

Thus: $\overline{V_1} = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$, $\overline{V_2} = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$.

– (minus) can be here, it won't change anything

$$V^T = \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix} = \begin{bmatrix} \overline{V_1}^T \\ \overline{V_2}^T \end{bmatrix}$$

↳ $1/\sqrt{5}\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}$

ns this is correct notation

~~$\begin{bmatrix} -2 \end{bmatrix}$~~

OR

$V^T = \begin{bmatrix} -2/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix}$, this is what ~~numpy~~ np returns

$\lambda = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$, entries in Σ are sqrt of that; + we need to align it to match the dimensionality 3x2 by adding a zero-valued row. That is:

$$\Sigma = \begin{bmatrix} \sqrt{6} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

3. To derive U, we'll use the formula: $AV = U\Sigma$ as follows:

$AV = U\Sigma$ → just write: $\frac{1}{\sqrt{5}}\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}$

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix} = U\Sigma$$

$$\begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 5/\sqrt{5} & 0 \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix} = U\Sigma$$

$$\rightarrow = \frac{\begin{bmatrix} 2 & 1 \\ 5 & 0 \\ 1 & -2 \end{bmatrix}}{\sqrt{5}} \quad \text{OR} \quad \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & 1 \\ 5 & 0 \\ 1 & -2 \end{bmatrix}$$

To get U, we can use this trick: we divide the columns of AV by the nonzero values in Σ in the respective col's:

$$U = \frac{1}{\sqrt{5}} \begin{bmatrix} 2/\sqrt{6} & 1 \\ 5/\sqrt{6} & 0 \\ 1/\sqrt{6} & -2 \end{bmatrix}.$$

TODO: these should be minuses. numpy yield diff + result for A if you use this matrix. Minuses should be in 1st column.

this didn't work out because I used the V^T returned by np which was different from my V^T. Minuses are crucial, but you only need to not mix up your results with numpy's. I.e. don't play w/ them, don't negate anything for no good reason. Recall that there can be multiple SVDs to A.

4) Now, let's attach the column $\frac{1}{\sqrt{6}}\begin{bmatrix}-2\\1\\-1\end{bmatrix}$ to U, to align it w/ the dim-ty 3×3. (Technically, we don't really care about that column b/c it's gonna be nullified by last row of Σ anyway. i.e., its values This is indeed true, I verified this in numpy. You can also throw that col out and last row of Σ, that would work just as well: 3×2 @ 2×2 @ 2×2 = 3×2):

$$U = \begin{bmatrix} 2/\sqrt{30} & 1/\sqrt{5} & -2/\sqrt{6} \\ 5/\sqrt{30} & 0 & 1/\sqrt{6} \\ -1/\sqrt{30} & -2/\sqrt{5} & -1/\sqrt{6} \end{bmatrix}.$$

(2) (a) $\hat{\beta}_{OLS} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X^T X)^{-1} X^T y$

x_i	1	2	3	4	5
y_i	6.6	7.9	9.4	11.1	12.4

~~$X^T X$~~ $=$

2×2

$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$ $\quad X^T X = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$ (2×2)

Finding inverse $(X^T X)^{-1}$ via Gaussian elim-

0) $\left[\begin{array}{cc|cc} 5 & 15 & 1 & 0 \\ 15 & 55 & 0 & 1 \end{array}\right]$ initial state

1) $R_2 = R_2 - 3R_1$ $\left[\begin{array}{cc|cc} 5 & 15 & 1 & 0 \\ 0 & 10 & -3 & 1 \end{array}\right]$

2) $R_1 = R_1 - 1.5R_2$ $\left[\begin{array}{cc|cc} 5 & 0 & 5.5 & -1.5 \\ 0 & 10 & -3 & 1 \end{array}\right]$

3) $R_1 = \frac{1}{5} R_1$ $\left[\begin{array}{cc|cc} 1 & 0 & 1.1 & -0.3 \\ 0 & 10 & -3 & 1 \end{array}\right]$

4) $R_2 = \frac{1}{10} R_2$ $\left[\begin{array}{cc|cc} 1 & 0 & 1.1 & -0.3 \\ 0 & 1 & -0.3 & 0.1 \end{array}\right]$

inverse

$(X^T X)^{-1} X^T = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$

(2×5)

$= \begin{bmatrix} 0.8 & 0.5 & 0.2 & -0.1 & -0.4 \\ -0.2 & -0.1 & 0 & 0.1 & 0.2 \end{bmatrix}$

$(X^T X)^{-1} X^T y = \begin{bmatrix} 0.8 & 0.5 & 0.2 & -0.1 & -0.4 \\ -0.2 & -0.1 & 0 & 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 6.6 \\ 7.9 \\ 9.4 \\ 11.1 \\ 12.4 \end{bmatrix} = \begin{bmatrix} 4.96 \\ 1.52 \end{bmatrix}$

(2b) $\hat{\beta}_{ridge} = (X^TX - \lambda I)^{-1} X^T y = \begin{bmatrix} 2.66 \\ 2.10 \end{bmatrix}$

$\lambda = 1$

(2c) $\hat{\beta}_{OLS}$ is far closer to $\beta = \begin{bmatrix} 5 \\ 1.5 \end{bmatrix}$, apparently, $\lambda = 1$ imposes too strict regularization, reducing it should bring $\hat{\beta}_{ridge}$ closer to β.

(3) See this video on it, and Annika Bätz's notebook (Problem Sheet 2, Ex 5)

(a) May find expected and/or actual regret.

Expected regret is concerned w/ expected rewards, i.e., probabilities of getting rewards, whereas actual regret is based on actual rewards (that is, 1's), i.e., on observed rewards.

$$\text{Exp. regret} = \sum_{t=1}^{T} E(r(Y_t(a^*)) - E(r(Y_t(a_t)))$$

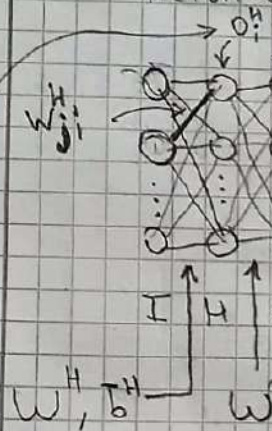
$$= T \cdot E(r(Y_t(a^*)) - \sum p_t$$

$$= T \cdot p^* - \sum p_t$$

Act. regret: ≠ ← for this, you need to have data where all actions were taken and rewards for them were observed, like in Ads csv in Problem Sheet 2, Ex 5. See how it's computed in Annika's notebook.

(4) See slides on NN by de Wiljes, a solved example for sigmoid activation is in PS7. In a nutshell, sol. for w^H_{ji} is same as for b^H_k in slides but multiplied by O^I_j at the end + with $sig(x)(1-sig(x))$ replaced w/ $1/\cosh^2(x)$.

Detailed solution:



$$\frac{\partial E}{\partial w^H_{ji}} = \sum_{k=1}^{N_O} (O^O_k - t_k) \frac{\partial}{\partial w^H_{ji}} (O^O_k - t_k)$$

(not a term of w^H_{ji})

$$\frac{\partial O^O_k}{\partial w^H_{ji}} = \frac{\partial \tanh(X^O_k)}{\partial w^H_{ji}} = \frac{1}{\cosh^2(X^O_k)} \cdot$$

$$\cdot \frac{\partial X^O_k}{\partial w^H_{ji}}$$

$$\frac{\partial X^O_k}{\partial w^H_{ji}} = \frac{\partial}{\partial w^H_{ji}} \sum_{m=1}^{N_H} \underbrace{O^H_m}_{\text{func of } w^H_{ji} \text{ when } m=i} w^O_{mk} + b^O_k$$

just look at any neuron in O layer

$$= w^O_{ik} \frac{\partial O^H_i}{\partial w^H_{ji}} = w^O_{ik} \frac{\partial \tanh(X^H_i)}{\partial w^H_{ji}} = w^O_{ik} \frac{1}{\cosh^2(X^H_i)} \frac{\partial X^H_i}{\partial w^H_{ji}}$$

$$\frac{\partial X^H_i}{\partial w^H_{ji}} = \frac{\partial}{\partial w^H_{ji}} \sum_{v=1}^{N_I} \underbrace{O^I_v w^H_{vi}}_{\text{func of } w^H_{ji} \text{ when } v=j} + b^H_i = O^I_j$$

Putting everything together:

$$\frac{\partial E}{\partial w^H_{ji}} = \sum_{k=1}^{N_O} (O^O_k - t_k) \frac{1}{\cosh^2(X^O_k)} \cdot$$

$$\cdot w^O_{ik} \frac{1}{\cosh^2(X^H_i)} O^I_j .$$

for sigmoid, just replace these w/ $sig(x)\cdot(1-sig(x))$

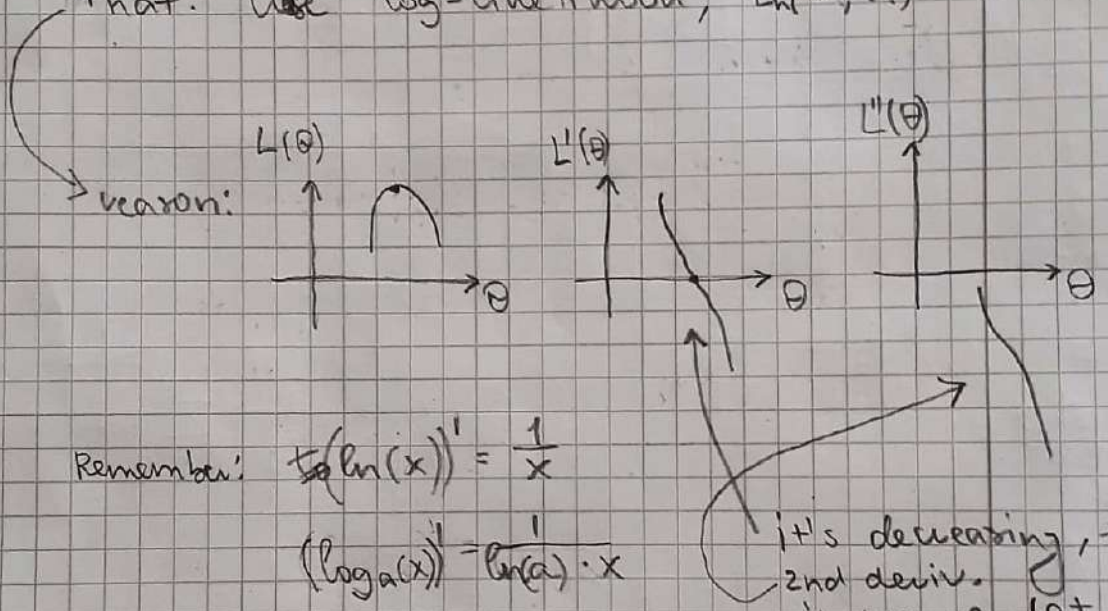
To do (SGD step): $w_{ij} = w_{ij} - \eta \frac{\partial L}{\partial w_{ij}}$ ← plug in here actual values

(5) (a) ~~Depending on how many trial outcomes you have (1 or more)~~

PMF of Poisson distr.

$$L_n(k;\lambda) = \prod_{i=1}^{n} P(X=k)$$

(b) See PDF w/ notes. Note that you must prove the found stationary point is a maximum via 2nd derivative, its sign must be negative for that. Use log-likelihood, $\ell_n(k;\lambda)$.



Remember: $(\ln(x))' = \frac{1}{x}$

$(\log_a(x))' = \frac{1}{\ln(a)\cdot x}$

it's decreasing, thus 2nd deriv. is neg-ve (at least at that point)

For single trial (i.e., not a set of iid $Y_1, Y_2, \ldots$), $\hat{\theta}_{MLE}(\lambda) = k$. This is really logical, knowing the intuition behind the Poisson distribution. That is, if a certain event occurred k times, then most likely that's the value of λ. Try to find such logic elsewhere.

Steps for finding MLE:

1) Write down its formula
2) Find der-ve w.r.t. target param

3) ~~Solve~~ Set der to 0 and solve for param

4) Verify that's a maximus using 2nd der

(5c) Just use the MLE formula:

$$\hat{\theta}_{MLE}(\lambda) = \frac{1}{n}\sum k_i$$

(d) $P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$ ← plug in $k=5$ and $\lambda = \hat{\theta}_{MLE}(\lambda)$

(e) ~~k~~ k_i are actually iid r.v.s, thus:

$$E\left[\hat{\theta}_{MLE}(\lambda)\right] = E\left[\frac{1}{n}\sum k_i\right] = \frac{1}{n}\sum E[k_i]$$

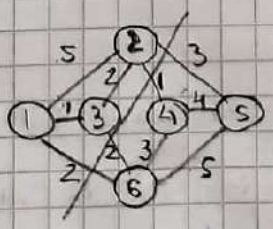
($k_i \sim$ Poisson; $= \lambda$; by linearity of expectation)

$$= \lambda = E[Y], \quad Y \sim \text{Poi}\,\lambda \quad \blacksquare$$

at least, you can treat them like this

(6) see material in Woid file graph and

Assume next ~~g~~ cut is given:



a) $D(G) = \begin{bmatrix} d(1) & & & & & 0 \\ & d(2) & & & & \\ & & d(3) & & & \\ & & & d(4) & & \\ & & & & d(5) & \\ 0 & & & & & d(6) \end{bmatrix} =$

$\forall d_i, \sum w_{ij}$ →

$$= \begin{bmatrix} 8 & & & & & 0 \\ & 11 & & & & \\ & & 5 & & & \\ & & & 8 & & \\ & & & & 12 & \\ 0 & & & & & 12 \end{bmatrix}$$

$$W(G) = \begin{bmatrix} 0 & 5 & 1 & 0 & 3 & 2 \\ 5 & 0 & 2 & 1 & 3 & 0 \\ 1 & 2 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 4 & 3 \\ 0 & 3 & 0 & 4 & 0 & 5 \\ 2 & 0 & 2 & 3 & 5 & 0 \end{bmatrix}$$

$$L(G) = D(G) - W(G) = \begin{bmatrix} 8 & -5 & -1 & 0 & 0 & -2 \\ -5 & 11 & -2 & -1 & -3 & 0 \\ -1 & -2 & 5 & 0 & 0 & -2 \\ 0 & -1 & 0 & 8 & -4 & -3 \\ 0 & -3 & 0 & -4 & 12 & -5 \\ -2 & 0 & -2 & -3 & -5 & 12 \end{bmatrix}$$

6 (b) $A_1 = \{1, 2, 3\}$, $A_2 = \{4, 5, 6\}$

$$\text{RatioCut}(A_1, A_2) = \sum_{i=1}^{k=2} \frac{\text{cut}(A_i, \overline{A_i})}{|A_i|}$$

$$= \frac{8}{3} + \frac{8}{3} = \frac{16}{3} = 5\frac{1}{3}$$

$$\text{NCut}(A_1, A_2) = \sum_{i=1}^{k=2} \frac{\text{cut}(A_i, \overline{A_i})}{\text{vol}(A_i)} = \frac{8}{8+11+5} +$$

$$+ \frac{8}{8+12+12} = \frac{8}{32} + \frac{8}{32} = \frac{7}{12}$$

(6c) Comment on the results – no idea what to say
Maybe say that denominator of RadioCut is optimized (minimized), but that of NCut isn't (in my example) – well, it still may be best for this graph.

$\Sigma \frac{1}{|A_i|}$ $\Sigma \frac{1}{\text{vol}(A_i)}$

Say smth about $L(G)$?