KEN3450: Data and Image Analysis
Final Exam (4/7/2017) *

Your name: KEY

Your student number: _____

Expected grade (optional): ____ / 80

Part	Points	/max
1		/13
2		/13
3		/17
4		/6
5		/15
6		/10
7		/6
B		/1
Total		**/81**

General remarks

- This exam consists of 14 pages and 7 parts (and 1 bonus question). Each part indicates a number of points. These points are indicative of how important the questions are and will be used proportionally for grading.
- If you multiply points by 2 you get how much time (in min. approx.) you have to spend on each question (e.g. 5 points, means 10 minutes).
- The exam contributes 80% to your final grade (data clinics is 10% and data madness is 10%), unless otherwise agreed.
- The answer boxes are small on purpose: You should try and give short answers that are to the point.
- If you need "experimental space" just use the back (white) pages.
- While you shouldn't spend too much time on calligraphy, please make sure I will have at least a chance deciphering your exam

*: Previous exam was on 7/4/2017, coincidence?

1 EDA warm up! (13 points)

(6) a. To honor 4th of July (i.e. today!), a popular US news magazine wants to write an article on how much Americans know about capitals of their states. They devise a test that lists 100 cities and each respondent must guess the state in which the city can be found. Each correct answer earns one point, for a maximum of 100. The random sample of 5000 people had a distribution of scores that was *normally* distributed with mean 62 and standard deviation 12.

(2)i. How many countries can identify correctly the central 95% of the people in this sample?

Normal model : 95% of the answers are in the interval: $[\mu-2\sigma, \mu+2\sigma] \rightarrow$
$[62-2\times12, 62+2\times12] \rightarrow$
$[39, 86]$
95% of people can identify between 38 and 86 countries

(2)ii. A journalist claims that a score of 45 (and below) should be considered poor. Do you agree or not? Justify your answer scientifically.

$$z_{score} = \frac{x-\mu}{\sigma} = \frac{45-62}{12} \approx -1.42$$

↓

not between $[-1, 1]$
So unusual but not rare

(2)iii. Can you decide from the given information whether there are any outliers or not (either too smart or not that smart people)? What extra diagram/sketch (except the actual scores) would you need and how you were going to use it for answering?

No. I need a boxplot

(7) b. The boxplots show the age of people involved in accidents according to their role in the accident.

(1) i. Which role involved the youngest person and what is the age?

passenger <1

(1) ii. Which role had the lowest median age and what is the age?

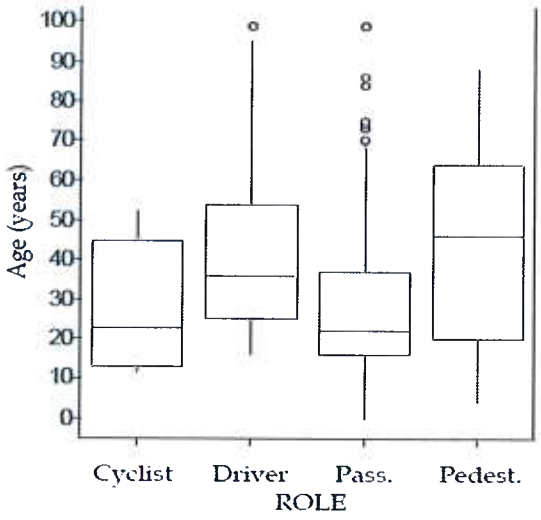
passenger 21

(1) iii. Which role had smallest range of ages and what is it?

cyclist 40

(1) iv. Which role had the largest IQR of ages and what is it?

pedestrian 44



(3) v. A journalist asks you the tricky question: "Which role generally involves the oldest people?" Justify your answer scientifically in order to impress the journalist.

pedestrian

- median age for pedestrians is ~45, while for the other groups is between 22-35
- the oldest 50% of pedestrian group (45-87) is older than the younger 75% of cyclist & passenger

....

2 Ranking documents (13 points)

You are looking for information on the Economic Growth in Scotland in a large document collection. You decide to search using the terms: economy, growth, Scotland, banks and business using an information retrieval system and this recommends three possible documents. You are given the frequency of each of the terms in each document (as well as your query), shown in the table below:

	t1	t2	t3	t4	t5
Terms	economy	Scotland	growth	banks	business
Doc #1 d1	10	8	0	2	1
Doc #2 d2	0	0	9	9	8
Doc #3 d3	2	2	4	4	6
Query q	1	1	1	1	1

You have no additional information about the documents.

(6) a. One possible measure for determining which of the 3 documents is the best match is cosine similarity, which measures the cosine of the angle between the query vector and that of each document. Compute this measure for each of the three documents and based on these results indicate which document is the best match for this query.

$$\cos(q,d_1) = \frac{10\times1+8\times1+0+2\times1+1\times1}{\sqrt{1^2+1^2+1^2+1^2+1^2}\times\sqrt{10^2+8^2+0^2+2^2+1^2}} \simeq 0.722$$

$$\cos(q,d_2) = \frac{1\times0+1\times0+1\times9+1\times9+1\times8}{\sqrt{5}\times\sqrt{9^2+9^2+8^2}} \simeq 0.773$$

$$\cos(q,d_3) = \frac{1\times2+1\times2+1\times4+1\times4+1\times6}{\sqrt{5}\times\sqrt{2^2+2^2+4^2+4^2+6^2}} \simeq 0.923$$

Best match is document 3 (cos ↑, angle ↓)

(4) b. Do you agree with the results of this analysis (take into account the ranking of the documents to the query)? What are (some of) the strengths and weaknesses of cosine measure? Use the above example to justify your answer. How would you attempt to correct the results?

Strength: takes into account the angle (normalization does not matter)

Weakness: sparsity (zeros)

perform normalization

(3) c. One person complained that searched for economy, growth, scotland, bank, business and got different results than the above. Explain how is that possible and how would you solve this problem in a future system.

Scotland ≠ Scotland

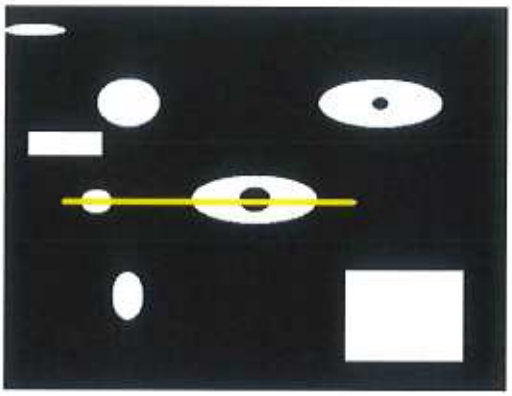
banks ≠ bank

Solution: capital case → lower case

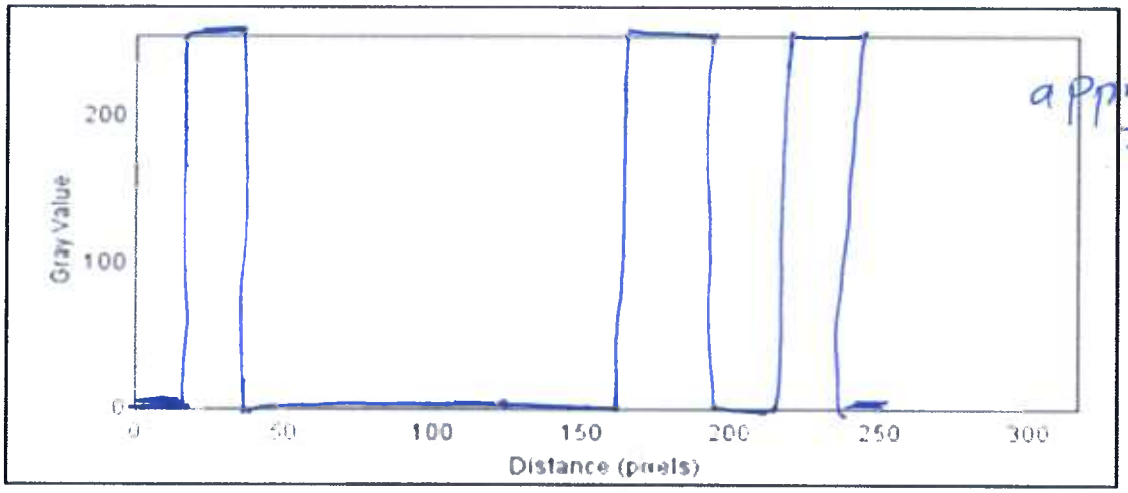
lemmatization / stemming

3 Understanding images (17 points)

(10) a. On the image shown on the right, we are interested in the line profile, as indicated by the line (crossing two patterns in the middle).



(2)i. Can you draw the approximate histogram of this line profile below? (suppose that length of the line is 300 pixels)

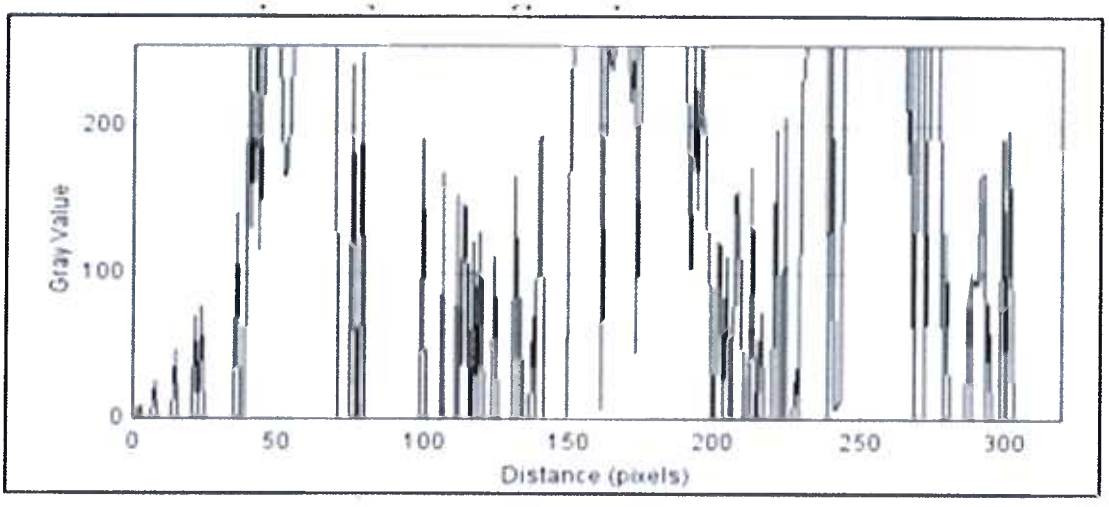


approximation

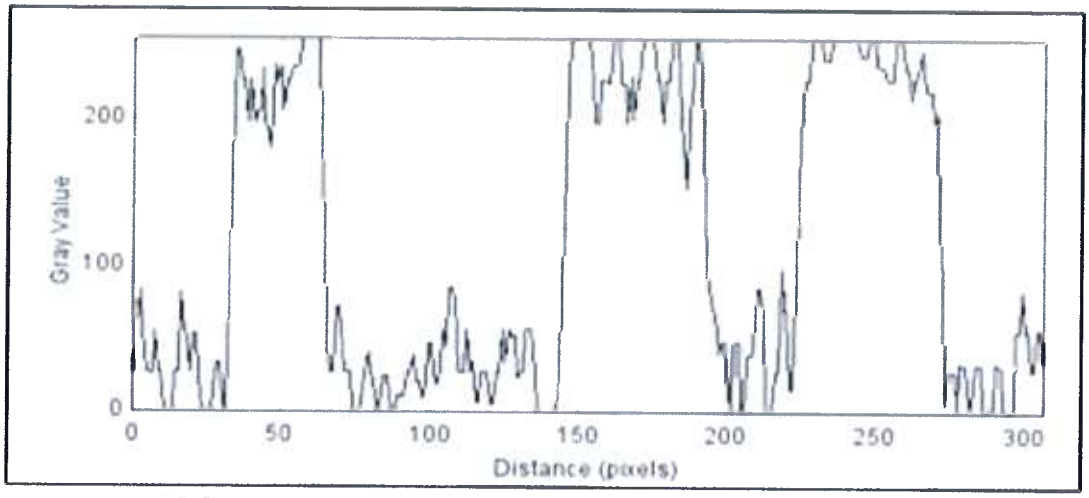
(6)ii. We perform some processing on the image and we obtain the three line profiles below. Indicate (in the given choices) which processing were done to obtain all three A, B and C. Explain your reasoning only if you are not certain.

Possibilities of algorithms (in alphabetical order):

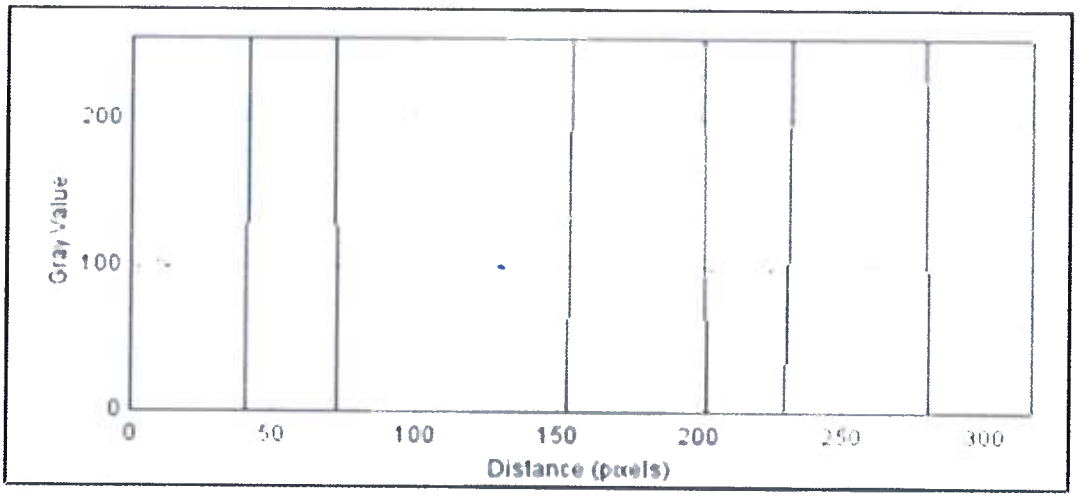
(I) Mean filter
(II) Median filter
(III) Salt and pepper noise



A: Line Profile after processing on original image



B: Line Profile after processing on Image A **(i.e. the previous one)**



C: Line Profile after processing on Image A

Line profile ~~**Filter**~~		**Short justification (only if you are not sure)**
A	III	
B	I	
C	II	

(2)iii. Say that someone wants to get rid of the black holes in the (two) rounded shapes of the original image. Which operator/filter you would apply and why?

dilation (on the reverse image)

(7) b. Given the 4x4 (grayscale) image and the 3x3 convolution mask following:

$$\text{Image: } \begin{bmatrix} 0 & 0 & 0 & 100 \\ 0 & 0 & 100 & 100 \\ 0 & 100 & 100 & 0 \\ 100 & 100 & 0 & 0 \end{bmatrix} \qquad \text{Filter: } \begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

(5)i. Compute the convolution of the image by the mask. (*NB*: Treat the borders any way you like, as long as you mention it in the answer box below. There is no need to show your calculations, just show the final result).

0	200	400	100
200	400	200	-200
200	200	-200	-400
100	-200	-400	-200

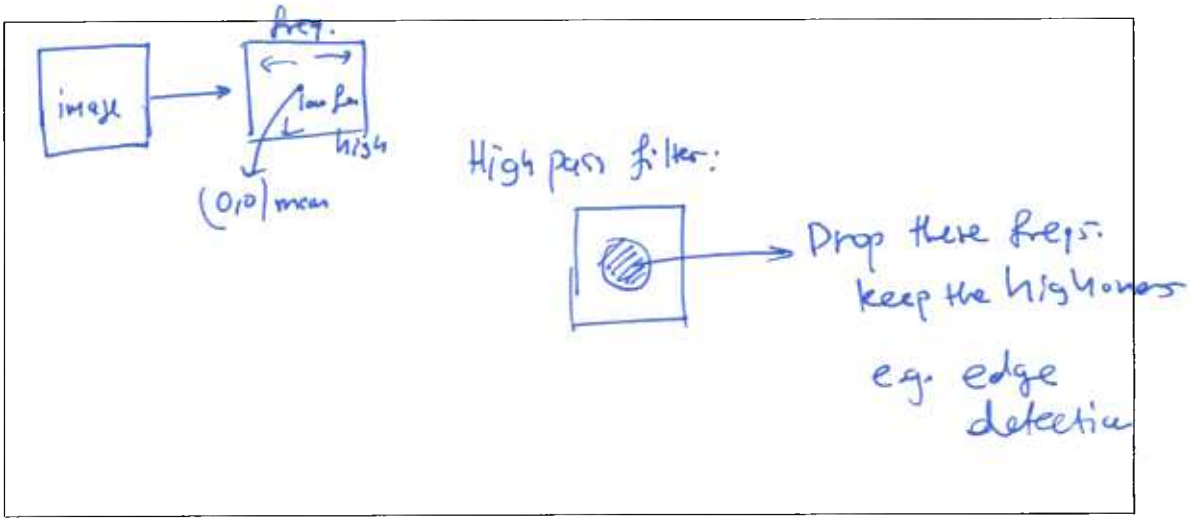
Zero padding (other possibilities welcome)
→ note that the result needs normalization

(2)ii. Given the description of this convolution mask/filter, can you guess for what reason is used for? (*Hint:* Does it remind you of a known filter?)

Sobel mask (gradient) ~~for~~ diagonal edges (45°)

4 Frequency enhancement (6 points)

(3)a. Explain how the Fourier Transform is used for image enhancement by applying frequency domain filters, such as high-pass filtering. What is the main task for a high-pass filter as far as image processing is concerned? Use appropriate sketches to justify your answer.



(3)b. Explain why sampling frequency plays an important role in timeseries modeling and how it can affect Fourier transformation results.

Mention Nyquist frequency : $f_N = \frac{f_s}{2}$

aliasing in spectrum : nothing $> f_N$ is visible

5 Modeling homelessness (15 points)

Homelessness is a problem in many large U.S. cities. To better understand the problem, a multiple regression was used to model the rate of homelessness based on several explanatory variables. The following data were collected for 50 large U.S. cities. The regression results appear below.

Homeless	number of homeless people *per 10,000 in a city*
Poverty	percent of residents with income under the poverty line
Unemployment	percent of residents unemployed
Temperature	average yearly temperature (in degrees F.)
Vacancy	percent of housing that is unoccupied
Rent Control	indicator variable, 1 = city has rent control, 0 = it has not

Dependent variable is **Homeless.** R squared (R^2)= 38.4%, Adjusted R^2 = 31.5%

Variable	Coefficient	SE (Coeff.)	t-ratio	p-value
Intercept	-4.275	3.465	-1.23	0.2239
Poverty	0.0823	0.0823	1.00	0.3228
Unemployment	0.159	0.218	0.73	0.4699
Temperature	0.135	0.0587	2.30	0.0262
Vacancy	-0.247	0.138	-1.79	0.0809
Rent Control:1	2.944	1.37	2.15	0.0373

(2) a. Which variables are associated with the number of homeless people in a city (using a 95% significance level)?

temperature & rent-control (p-value < 005)

NB: vacancy is >005

(4) b. Explain the meaning of the coefficients: **Temperature** and **Rent Control** in the *context of this problem.*

temperature

For cities with fixed poverty, unemployment, etc (all variables), an increase in 1°F leads to a decrease of 0.14 homeless per 10,000

Rent control: for cities with fixed other variables, cities with rent control have 2.94 homeless people more than cities without

(2) c. Do the results suggest that having rent control laws in a city causes higher levels of homelessness? Explain.

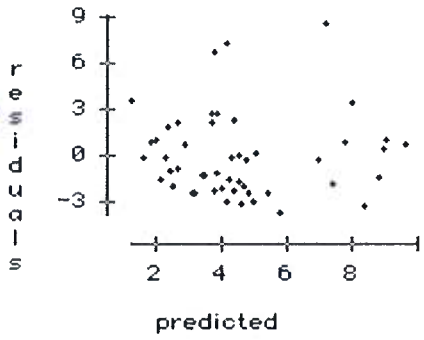
No.
this is an observational study, so I cannot establish causal relations

(4) d. If we created a new model by adding several more explanatory variables, which statistic should be used to compare them—the R^2 or the adjusted R^2? Explain.

Adjusted R^2 because it controls for models with different # of predictors

(3) e. Using the plot of residuals vs predicted values below, can you conclude that the above regression is a good fit (in terms of distribution of residuals)? Justify your choice.

* seems random enough (no obvious relation), so it is a good fit



6 PCA-phobia (10 points)

Ten participants are given a battery of personality tests, comprising the following items: Anxiety; Agoraphobia; Arachnophobia; Adventure; Extraversion; and Sociability (with a scoring range of 0 to 100). The purpose of this project is to ascertain whether the correlations among the six variables can be accounted for in terms of comparatively few latent variables or factors.

Part	Anx	Agora	Arach	Adven	Extra	Socia
1	71	68	80	44	54	52
2	39	30	41	77	90	80
3	46	55	45	50	46	48
4	33	33	39	57	64	62
5	74	75	90	45	55	48
6	39	47	48	91	87	91
7	66	70	69	54	44	48
8	33	40	36	31	37	36
9	85	75	93	45	50	42
10	45	35	44	70	66	78

(3) i. Provide some descriptive statistics on the dataset (mean, st. deviations for each item) and justify whether normalization is needed for this problem.

anx: 53.1 sd: 19.04
agora: 52.8 sd: 18.08
arach: 58.5 sd: 22.24
adven: 56.4 sd: 17.99
Extra: 59.3 sd: 17.70
Socia: 58.5 sd: 18.44

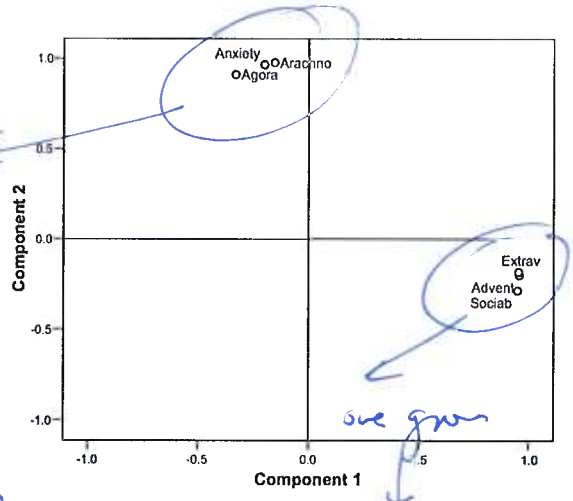
all close so might not be necessary to normalize

(2) ii. Running PCA on this (small) dataset gives the following results in terms of the principal components (NB: here the full results with 6 components are presented), along with the eigenvalues and the % of variance explained by each eigenvalue. As an experienced data scientist how many eigenvalues would you pick? Justify your choice.

Component	Total	% of Variance
1	4.164	69.397
2	1.612	26.862
3	.144	2.396
4	.052	.867
5	.023	.383
6	.006	.095

First two PCAs explain ≅ 96.3% of the variance in data. They also are > 1 (eigenvalues)
So we pick (2)

(4) iii. The plot for two PCs is given below. How would you communicate to a non-data-scientist your final conclusion from this dataset by using this figure and the results from ii.)? Can you predict (without computing) what would be the correlations between the initial 6 variables?



one group
↓
expect correlation

I have two groups of phobias one general and one related to social

one group
↓
high correlation

7 (fast) regularization check questions (6 points)

Explain in one or two sentences why the statements are true (or false).

(2)a. L2 (Ridge) regularization is more robust to outliers than L1 (Lasso).

there is no difference.

(2)b. In terms of feature selection, L2 (Ridge) regularization is preferred (to L1-Lasso) since it comes up with sparse solutions.

No, lasso ⇒ feature selection (sparse) because many coefficients go to 0

(2)c. Nearly all the algorithms we have been through this course, have a tuning parameter which is performing regularization (even if we didn't call it that). In the concept of using PCA as a dimension reduction technique (i.e. reduce training data down to k dimensions with PCA and use these as features), one claims that higher k means less regularization. Do you agree or not?

Yes, ↑k more parameters ⇒ less generalization
less regularization

Bonus question (1 point)

What is Jerry's favorite color?

Answer: _____