Maastricht University

Department of Data Science and Knowledge Engineering

Data Analysis 2017/2018

Final Exam

– Do not turn this page before the official start of the exam! –

First name, Surname: KEY

Student ID: _____

Expected Exam Grade (out of 50): ____ (optional)

Program: Bachelor Data Science and Knowledge Engineering

Course code: KEN3450

Examiner: Dr. Gerasimos (Jerry) Spanakis

Exam version: Z

Date/time: Tuesday, 3rd July 2018, 14.00-17.00h

Format: Open book exam

Allowed aides: Pens, simple (non-programmable) calculator from the DKE-list of allowed calculators.

Instructions to students:

- The exam consists of 7 questions (plus one bonus question) on 15 pages.
- Fill in your name and student ID number on each page, including the cover page.
- Answer every question at the reserved space below the questions (keep your answers short and to the point). If you run out of space, continue on the back side, and if needed, use the extra blank page.
- Ensure that you properly motivate your answers.
- This exam sums to 50 points (plus 1 bonus point) and counts for 50% of your final grade. The rest 50% comes from data clinics (30%), data madness (10%) and Kaggle competition (10%)
- While you shouldn't spend too much time on calligraphy, please make sure that I have at least a chance at deciphering your exam.
- You are not allowed to have a communication device within your reach, nor to wear or use a watch.
- You have to return all pages of the exam. You are not allowed to take any sheets, even blank, home.
- If you think a question is ambiguous, or even erroneous, and you cannot ask during the exam to clarify this, explain this in detail in the space reserved for the answer to the question.
- If you have not registered for the exam, your answers will not be graded, and thus handled as invalid.
- **Success!**

The following table will be filled by the examiner:

Question:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**B**	**Total**
Maximum points:	7	9	7	10	6	5	6	1	**51**
Achieved points:									

Question 1: EDA warmup (7 Points)

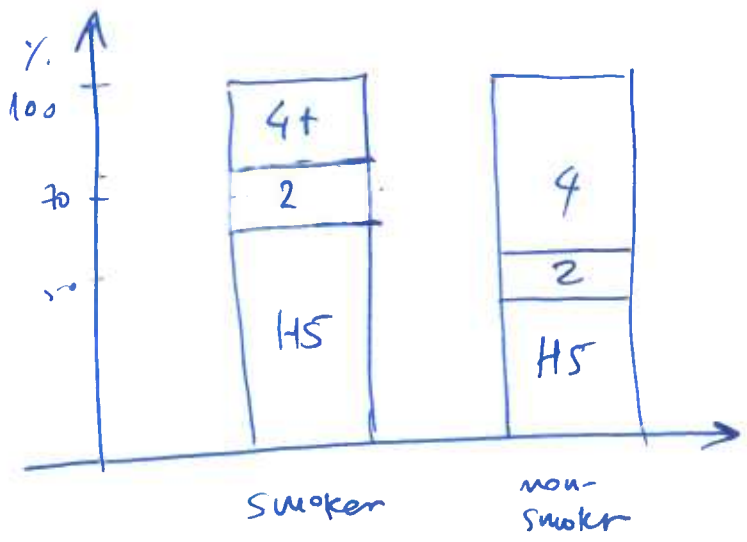
A.(5) 200 adults shopping at a supermarket were asked about the highest level of education they had completed and whether or not they smoke cigarettes. Results are summarized in the table.

	Smoker	Non-smoker	Total
High school	32	61	93
2 yr college	5	17	22
4+ yr college	13	72	85
Total	50	150	200

i.(1) What percent of the shoppers were smokers with only high school educations?
What percent of the shoppers with only high school educations were smokers?
What percent of the smokers had only high school educations?

(a) $32/200 = 16\%$
(b) $32/93 \cong 34.4\%$
(c) $22/50 = 64\%$

ii.(2) Sketch an appropriate graph/plot to for comparing education level among smokers and non-smokers. Label your graph clearly.



segmented bar chart is more appropriate here

iii.(2) Do these data suggest there is an association between smoking and education level? Give *statistical evidence* to support your conclusion. Does this indicate that students who start smoking while in high school tend to give up the habit if they complete college? Explain *shortly*.

(a) There is possible association between smoking & education level
64% of smokers had only a HS diploma
40.7% of non-smokers had only a HS-diploma
Similarly, 26% of smokers had 4+ year college
compared to 48% of non-smokers

(b) No, these data were collected at one time, about 2 different groups (smokers/non-smokers). We have no idea if smoking behavior changes over time

B.(2) An automobile service shop reported the summary statistics shown for repair bills (in €) for their customers last month.

Min	27
Q1	88
Median	132
Q3	308
Max	1442
Mean	284
SD	140

i.(1) Were any of the bills outliers? Show (using statistical evidence) how you made your decision.

Yes. IQR = Q3 − Q1 = 308 − 88 = 220
Upper Fence = Q3 + 1.5 IQR = 638 << 1442
which means that there are (at least one) outliers

ii.(1) After checking out a problem with your car the service manager gives you an estimate of "*only* €90." Is the manager right to imply that your bill will be unusually low? Explain briefly.

No. €90 is higher than Q1 (i.e. 25% of all bills) so it's not unusually low

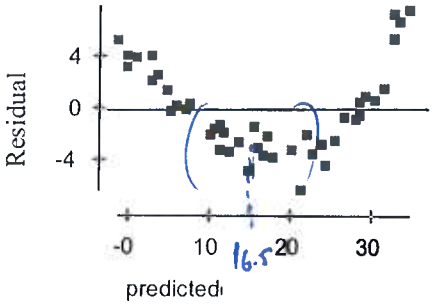
Question 2: Let's (re)discover penicillin (9 Points)

Doctors studying how the human body assimilates medication inject some patients with penicillin, and then monitor the concentration of the drug (in units/cc) in the patients' blood for seven hours. First, they tried to fit a linear model. The regression analysis and residuals plot are shown below.

Dependent variable is: Concentration
No Selector
R squared = 90.8% R squared (adjusted) = 90.6%
s = 3.472 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4900.55	1	4900.55	407
Residual	494.199	41	12.0536	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	40.3266	1.295	31.1	≤ 0.0001
Time	-5.95956	0.2956	-20.2	≤ 0.0001



A.(2) Explain what the value of R squared (90.8%) means in this context and what is the difference from R squared adjusted.

90.8% in the variance of penicillin concentration can be explained by "time" variable

B.(1) Using this model, estimate what the concentration of penicillin will be after *4 hours*.

$\hat{y} = 40.3266 - 5.95956 \times 4 \cong 16.5$ units/cc

C.(2) Is that estimate likely to be accurate, too low or too high? Use the residual plot to explain.

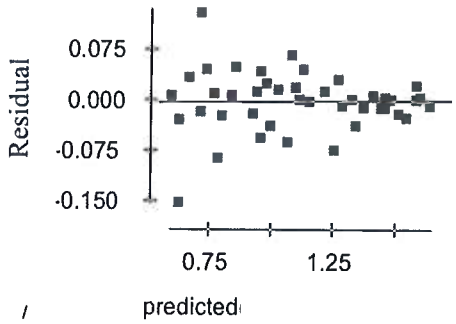
See figure: Predicted value (4 hours) is in an area of negative residuals meaning that the estimate will be too high

Now the researchers try a new model, using the re-expression log(*Concentration*). Examine the regression analysis and the residuals plot below.

Dependent variable is: LogCnn
No Selector
R squared = 98.0% R squared (adjusted) = 98.0%
s = 0.0451 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4.11395	1	4.11395	2022
Residual	0.083412	41	0.002034	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1.80184	0.0168	107	≤ 0.0001
Time	-0.172672	0.0038	-45.0	≤ 0.0001



D.(1) Is this model better than the previous one? Explain why you think yes or no.

Yes
(a) residuals show no pattern (random)
(b) better adj. R^2

E.(1) Using this new model, estimate the concentration of penicillin after 4 hours.

$\log_{10} C = 1.80184 - 0.172672 \times 4 = 1.11152$

$C = 10^{1.11152} \simeq 12.9$ units/cc

F.(2) Explain (in the context of this specific problem) what the coefficient of *"time"* means.

Every 1 hour passes ⇒ the logarithm of concentration is reduced by −0.172672

Question 3: Find the lemur (7 points)

The goal of a retrieval model is to score and rank documents for a query. Different retrieval models make different assumptions about what makes a document more (or less) relevant than another. Suppose you issue the query *"lemur"* to a search engine. And, suppose that two documents both contain the term *"lemur"* twice. Answer the following questions.

A.(1) Would the ranked Boolean (i.e. we give 1 if word is present, 0 if it's not) retrieval model *necessarily* give both documents the same score? If not, what information would determine which document is scored higher?

The boolean model ranks documents based on the number of times each term occurs in each doc

In this case, each document would obtain a score of 2.

So yes both documents would have the same score

B.(2) Would the cosine similarity (still using the boolean/binary representation) *necessarily* give both documents the same score? If not, what would determine which document is scored higher?

cosine similarity = inner product divided by the length of the query and the document

So, the scores would be different, if the # of unique terms in each document were different.

The document with fewer unique terms would get a higher score

(note: $\cos(q,d) = \frac{\langle q,d \rangle}{\|q\| \|d\|}$)

Suppose you have a collection of 4 documents (see below). A person searches for documents that contain information about "yellow lemur" and uses the query "yellow lemur" for this purpose. The system used for representing the documents in the Boolean/binary one.

C.(2) Given that what would be the ranking of the documents in terms of relevance? Comment on the results.

Doc1: yellow lemur sat on a wall
Doc2: lemur ate something yellow
Doc3: all the king's horses and all the king's men
Doc4: couldn't photo lemur.

Boolean model. one word = one term (I don't care about order)

rank

doc1, doc2 (2 terms)
doc4 (1 term)
doc3 (0 term)

since order is ignored doc1 and doc2 are ranked in the same position

D.(2) One student claims that implementing the TF-IDF weighting measure will improve the ranking results (with no other change or pre-processing). Do you agree? Explain shortly *(no need to compute the TF-IDF)*.

Even with TF-IDF, the results would not change since the words we are interested (yellow, lemur) would not get different weights (relatively). So no

Question 4: Images, filters and other fairytales (10 points)

A.(3) Suppose we have a very small binary image (I) and a structuring element (K).

$$I = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad K=[1\ 1]$$

Show how erosion and dilation are applied in this context (show the resulting images) and give justification for the reason that they might have been used in this specific problem.

Erosion:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- When overlaying the K with a pixel everything needs to ~~be~~ foreground to remain foreground

Example: removal of noise/lines

Dilation:

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

- When overlaying K with a pixel ~~[illegible]~~ at least one pixel needs to be foreground, to turn to foreground

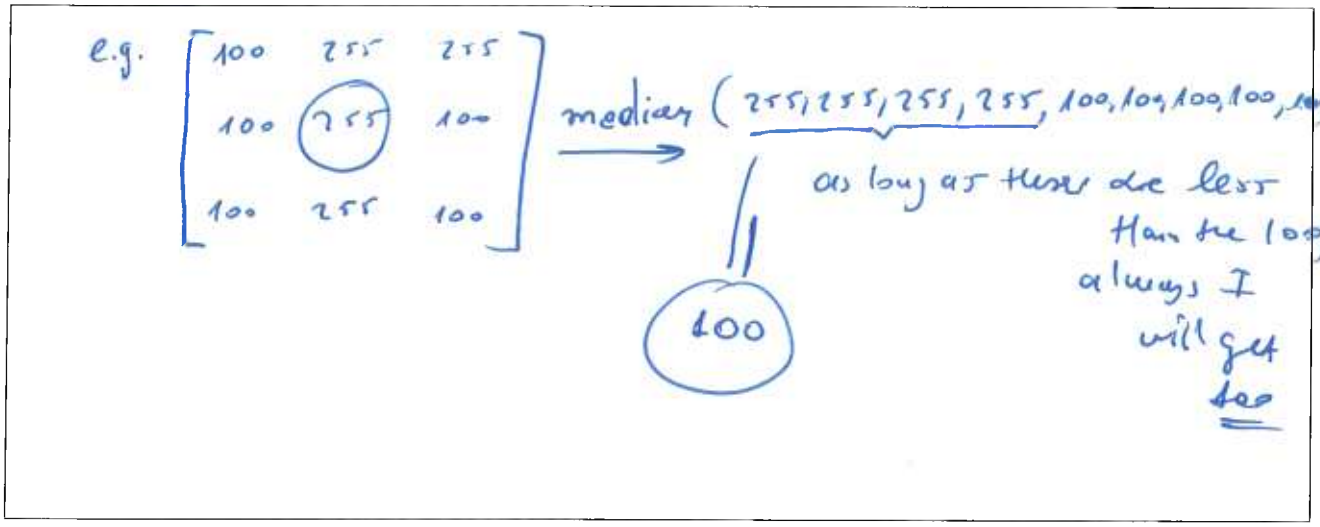
Example: Highlighting details

B.(2) Explain shortly why the filter below can detect edges in images. More specifically, explain what is the expected value of the filter at the edges and what is the expected value elsewhere.

$$L = \begin{bmatrix} 2 & -1 & 2 \\ -1 & -4 & -1 \\ 2 & -1 & 2 \end{bmatrix}$$

- Computes differences between pixels on x and y directions
- Expect 0 at the edges, non-zero elsewhere

C.(2) Suppose that *salt and pepper* noise is added to an otherwise *uniform* region of the image. Show *(with an example)* that provided <u>less than half</u> of the elements in the neighbourhood are noise values, then a 3 × 3 median filter will give the correct (i.e. uniform again) result.



D.(3) Consider the square shown below to be represented by a function f (x_1, x_2) of x_1 and x_2. The gray level shows the value at a given point, with black being 1 and white being 0. Next to f (x_1, x_2) is a plot of the magnitude of its Fourier transform |F f (ξ_1, ξ_2)|. Let a > 1 be a fixed constant, and let A be the matrix

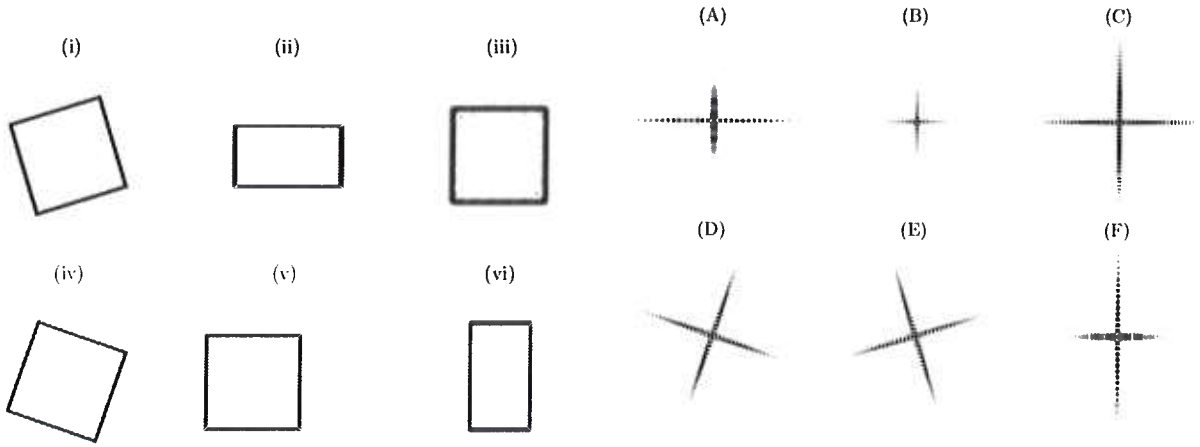
$$A = \begin{pmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{pmatrix}$$

Consider the following modifications of f (x_1, x_2):

1. $f(ax_1, x_2)$ 2. $f(x_1, ax_2)$ 3. $f(x_1 + a, x_2)$

4. $f\left(A\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)$ 5. $f\left(A^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)$ 6. $f(x_1, x_2) * \mathrm{sinc}(ax_1)\,\mathrm{sinc}(ax_2)$

Match the modification 1 – 6 of f(x_1, x_2) with the corresponding plot (i) – (vi) and with the plot of the corresponding Fourier transform (A) – (F). Give brief explanations.



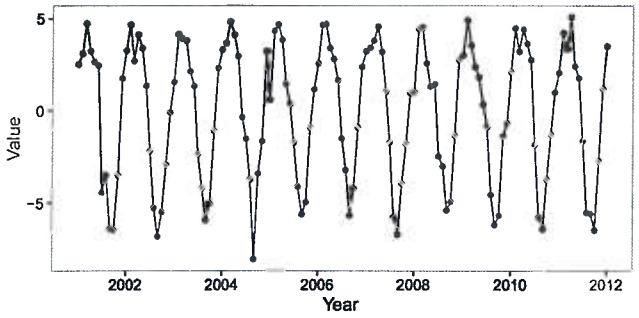
1. vi - A : $(x_1, x_2) \leftrightarrow (ax_1, x_2)$: shrinks in the x_1 direction
   FT: $\mathcal{F}(f(ax_1, x_2)) = \frac{1}{a} f(f(\xi_1/a, \xi_2)$: stretched in ξ_1 direction and the whole figure streched
   ii - F
2. $f(x_1, ax_2)$ similar with 1.
3. V - C : shift x_1 by a to the left
   FT: only a phase change $\Leftrightarrow$ same
4. iv - D : A: rotation by $\pi/6$ $\Rightarrow$ similar with FT
5. i - E : similar with 4.
6. iii - B : well, only one choice left ☺
   $\mathcal{F}(f(x_1, x_2) \cdot \text{sinc}(ax_1) \cdot \text{sinc}(ax_2)) = \mathcal{F}(f(f_1, f_2) \cdot \Pi_a(f_1) \Pi_a(f_2)$
   FT: cuts-off the FT by a function of a
   spatial domain: low-pass

Question 5: Time is of the essence (6 points)

A. (2) Why is stationarity a desirable property for a time series process?

With a stationary process, a longer time series (N??) gives more info about the statistical properties (mean, autocorrelation etc) and they don't change

B. (2) Below is a plot of a monthly time series. The investigator who collected it expects there to be an annual seasonal pattern, but she is more interested in the long-term trend.



Suggest a method the investigator could use to model the seasonality, along with shortly motivating why.

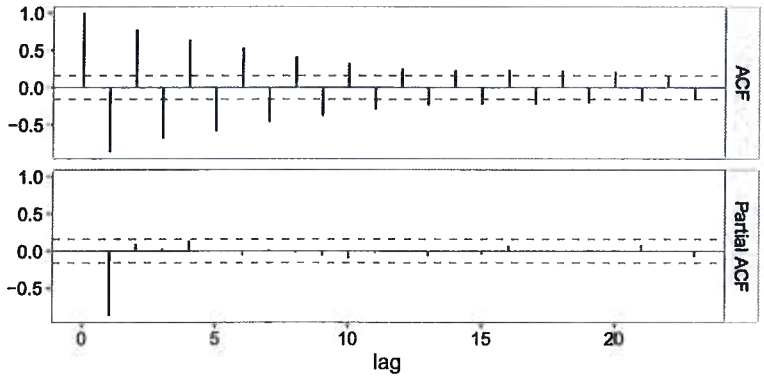
More than one correct answer.

Examples:

(a) calculate monthly means and model deviations from the mean

(b) model the residuals of regression against the month

C. (2) Given the autocorrelation plot and partial autocorrelation plot below, what is your estimate for which kind of ARIMA model the researcher should attempt?



ARIMA(1, 0, 0) :

Question 6: Small Dimensionality Reduction (5 points)

We have a small database of ratings of 5 movies by 6 users (negatives are allowed, in the case that a user dislikes a movie much). They are represented by a 6 × 5 matrix X, where each row corresponds to a user and each column corresponds to a movie.

movies

users

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

A.(1) What is the sample user mean, i.e. how is the average user rating movies?

obviously [0, 0, 0, 0, 0]

B.(2) If SVD (Singular Value Decomposition) is applied, how many singular values you expect to have? Explain why. What would change if two of the zeros were replaced by ones?

2 singular values (due to the structure of X)

If zeros are removed, that accounts for noise meaning that I expect more than 2 (3 or 4)

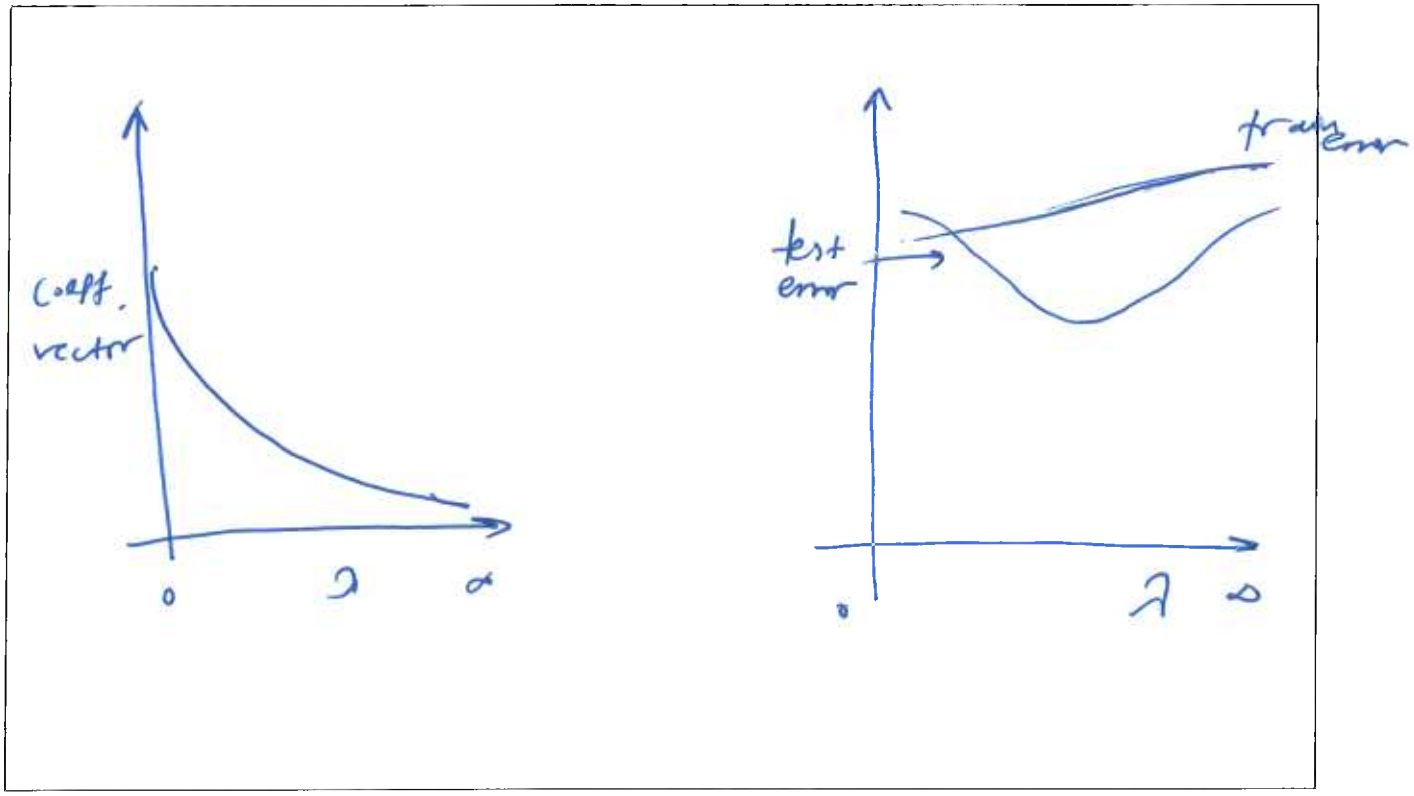
C.(2) A student very eager to apply a recommendation system on the above data, decided to do a per-user normalization (i.e. find the average rating of each user and subtract it from the values). Explain why the student decided to do this pre-processing step and what might be the problem in the dataset above.

- Remove rating bias for each user

- Problem because everything will become 0

Question 7: Regularization magic (6 points)

A.(3) Suppose you are doing linear regression using an L2 regularization penalty. Sketch the average cross-entropy training and testing error you would expect to see when you run the algorithm on a dataset, as well as the norm of the coefficient vector, as a function of the regularization parameter λ (make one graph for the error curves and another for the coefficient vector norm). No explanation is necessary.



B.(3) Suppose you used cross-validation to find a good value for λ, on a dataset of 500 examples with 100 features. But, you have now acquired more data, so instead of the original 500 examples, now you have 50000. Would the same value of λ work best, or would you expect a higher or lower value of λ to work better? Justify your answer.

I expect a lower value of λ

λ: gives a bias-variance trade-off

data volume ↑, bias ↓, no need for high λ

Bonus (Meta)-Question (1 point)

Find a mistake (typo, syntax or other) in this exam or the course slides.

Extra answer sheet.