


BA&ML Past Exam

- The following pages contain one specific configuration of the Business Analytics Retake Exam from the academic year 2020/2021. (Note: The course has since been renamed to “Business Analytics and Machine Learning”).
- **Please note that exam questions are copyrighted works. Do not redistribute this document.**
- This old exam is meant to give you a general idea of the types and scope of the questions that may be on future exams. Of course, this could be subject to change in future installments of the course.
- There will be a **separate, compulsory** preparatory exercise to familiarize yourself with the technical exam submission system and the Code of Conduct. Practicing with this old exam does NOT replace the mandatory preparatory exercise. See Moodle for details.
- The exam below was held **remotely** via TUMexam and was **open-book**. All notes and software tools were allowed. Any communication or cooperation with other students or third parties was forbidden. The limit was **90 minutes**.
- The exam contains questions worth 90 credits, thus on average, students will have 1 minute per credit. Use this information to decide how much time you spend on each problem.
We are aware that this constitutes significant time pressure. Unfortunately, this is a necessity for open-book remote exams to prevent cheating. You can expect similar time pressure in future exams.
- To reach the top grade of 1.0 in this exam in 2021, students needed to reach circa 75 credits.
- We will **not** publish solutions or a detailed grading rubric to this old exam.
- If you have questions about the exam’s format, please ask questions in the moodle forum. However, we will **not** answer questions that are specific to individual problems on this old exam.
- Please note that each student’s exam configuration will be unique, so different students will receive different problems.



Personal sticker



S5001

Compliance to the code of conduct

I hereby assure that I solve and submit this exam myself under my own name by only using the allowed tools listed below.

Signature or full name if no pen input available

Business Analytics: Retake Exam

Exam: IN2028 / Retake **Date:** Tuesday 30th March, 2021
Examiner: Prof. Dr. Martin Bichler **Time:** 14:15 – 15:45

	P 1	P 2	P 3	P 4	P 5
I					
II					

Working instructions

- This exam consists of **16 pages** with a total of **5 problems**.
Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 90 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources: open book.
- Subproblems marked by * can be solved without results of previous subproblems.
- **Answers are only accepted if the solution approach is documented.** Give a reason for each answer unless explicitly stated otherwise in the respective subproblem.
- Do not write in red or green colors.
- Any intermediate or final numbers in your calculations may be **rounded to three (3) digits**.

Left room from _____ to _____ / Early submission at _____



☐ Exam empty





Problem 1 Regression Analysis (25 credits)

Your pharmaceutical company employs sales agents to market their products to doctors. Some agents have recently been given tablet computers in order to improve their product pitches and to boost sales. You recorded the following data on sales presentations of your bestseller product to 1,000 different doctors from the last month.

Variable	Range	Explanation
sales	$\mathbb{R}_{\geq 0}$	sales due to the sales presentation [kEUR]
experience	$\mathbb{R}_{\geq 0}$	years of experience of the sales agent
prev	$\{0, 1\}$	previous sale to doctor occurred (0: no, 1: yes)
time	$\mathbb{R}_{\geq 0}$	duration of the sales presentation [min]
tablet	$\{0, 1\}$	tablet used for sales presentation (0: no, 1: yes)



a)* Does the data reflect cross-sectional, time-series or panel data? Explain your reason.



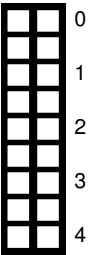


You estimate the following linear regression model:

Residuals:					
Min	1Q	Median	3Q	Max	
-8.9703	-2.7733	-0.0218	2.6774	9.8584	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.508246	0.323180	7.761	2.08e-14	***
experience	0.006429	0.009159	0.702	0.4829	
prev	0.488368	0.233287	2.093	0.0366	*
tablet	1.916271	0.238477	8.035	2.63e-15	***
time	0.104782	0.002762	37.934	< 2e-16	***
—					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Multiple R-squared: 0.6377, Adjusted R-squared: 0.6363					
F-statistic: 437.9 on 4 and 995 DF, p-value: < 2.2e-16					

b)* Interpret the estimated model:

1. Which variables are statistically significant at 5% and why?
2. Interpret the p-value of the F-statistic. What is the idea and null hypothesis of the F-test?





A colleague suggests to include an interaction effect for *tablet* and *time*. You receive the following output:

```
Residuals:
Min       1Q       Median       3Q      Max
-9.1694   -2.7383    0.1029    2.5894    9.6757

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.957551   0.366457   5.342  1.14e-07 ***
experience    0.004978   0.009130   0.545  0.58573
prev         0.466000   0.232365   2.005  0.04518 *
tablet       3.152159   0.459706   6.857  1.23e-11 ***
time         0.114169   0.004062  28.105 < 2e-16 ***
time:tablet  -0.017269   0.005500  -3.140  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Multiple R-squared:  0.6413, Adjusted R-squared:  0.6395
F-statistic: 355.4 on 5 and 994 DF, p-value: < 2.2e-16
```

0			
1			
2			
3			
4			
5			
6			
7			
8			

c)* Interpret the estimated model:

1. According to the model, what is the effect on sales if a presentation is extended by one minute (while keeping the other variables constant)? Make a case differentiation and demonstrate how you arrive at your conclusions.
2. Interpret the coefficient of *time:tablet* in one sentence.
3. Another colleague argues: "The regression is no longer linear w.r.t. *time* and *tablet*. Thus, the linearity assumption of the Gauss-Markov Theorem is violated". Explain why you agree or disagree.





You learn that ten additional features have been recorded that can be added to the set of independent variables, including wage level of the sales agent, age of the sales agent, distance travelled by the sales agent, etc.

d)* Briefly explain how a principal component regression and a ridge regression work. Discuss bias and variance of the estimators. Which problem do these techniques address and why are they appropriate here?

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2
<input type="checkbox"/>	3
<input type="checkbox"/>	4
<input type="checkbox"/>	5
<input type="checkbox"/>	6
<input type="checkbox"/>	7

You consider to extend the study to a longer period of time, recording a multitude of sales presentations to the same set of doctors over time. In addition to the recorded features, you assume that characteristics of the doctors themselves influence the sales, and that those characteristics are uncorrelated with your independent variables. Unfortunately, you have no possibility to record such individual characteristics of the doctors.

e)* Does this constitute a problem for your study? Give reasons for your answer. Describe one appropriate model and one associated statistical test.

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2
<input type="checkbox"/>	3
<input type="checkbox"/>	4
<input type="checkbox"/>	5





Problem 2 Naive Bayes (17 credits)

0

1

2

a)* What are the main assumptions for using a *Naive Bayes classifier*?

0

1

2

3

b)* What is the *zero-frequency problem* and how can it be addressed?

Table 2.1 contains product data of your online store, where “Sold” is the binary class attribute that indicates whether or not the corresponding item was sold within a week.

Item-ID	Category	Size	Price	Sold
31245	Tools	Large	High	No
36545	Accessories	Small	High	Yes
36745	Accessories	Small	Low	Yes
12444	Electronics	Small	High	Yes
29453	Accessories	Small	Low	Yes
21205	Electronics	Small	High	No

Table 2.1: Product data.

0

1

2

3

4

5

6

7

8

9

10

c) Train a *Naive Bayes classifier* on Table 2.1 using the zero-frequency correction. Then, classify the following instance:

Item-ID	Category	Size	Price	Sold
24531	Electronics	Large	Low	?

Notes:

- Explicitly state all probabilities you have used in the calculation.
- State and explain how you handle the ID column.





d) What problem do you see with this prediction in light of this business application and the available features and data?

	0
	1
	2





Problem 3 Clustering (16 credits)

The table below contains five records with the numeric attribute “Age”.

ID	Age
1	12
2	25
3	30
4	16
5	19

Note: ID is not an attribute but may help you formulate your answers. As distance measure, use Euclidean distance.

a) * Apply 2-means clustering to the five instances from the table above, using point A (30) and point B (25) as initial cluster centers. *Write down all calculation steps of the algorithm until the algorithm stops!*

0

1

2

3

4

5

6

7

8

9

10





b) * Imagine additional to the age you now also have the yearly income of a person in dollars. What do you have to do before applying k-means approach and why?

	0
	1
	2
	3

c) * Name one major difference between k-means clustering and hierarchical clustering in terms of reproducibility.


	0
	1
	2
	3










Problem 4 Association Rules (13 credits)

You are provided with an online music store database which contains information of all 25 users and which of the 15 different songs they bought. You applied the apriori algorithm with a minimum support of 0.4 and a minimum confidence of 0.8. The two most often bought songs were “Fortunate Son” by Creedence Clearwater Revival (F) and “Radar Love” by Golden Earring (R). The first (F) was bought by 13 users while the latter (R) was bought by 14 users; 10 users bought both (F) and (R). *Note: Please round to 3 decimals in this exercise*

0  a)* Which of the following association rules will be returned as a result of the apriori algorithm? *Calculate the support as well as the confidence for both rules!*

- 1 
- 2 
- 3 
- 4 
- i) $F \rightarrow R$
 - ii) $R \rightarrow F$
 - iii) Neither
 - iv) Both

0  b) * Assuming all users that bought both songs have bought (F) first and (R) second: How do you interpret the influence of having bought (F) on buying (R)? Calculate the Lift.





c) * In addition to before, you have the information that 15 of the 25 users bought “Thunderstruck” (T) by ACDC. In total 10 users bought both (F) and (T), 13 bought both (R) and (T), and 10 users bought all (F), (R), and (T). Now one more user record is added to the database. After that, the following happens:

Rule R1: $R \rightarrow F, T$; $\text{Conf}(R1)$ is 0.667 (rounded by 3 digits)

Rule R2: $F \rightarrow T$; $\text{Conf}(R2)$ is 0.714

Rule R3: $F \rightarrow R$; $\text{Lift}(R3)$ decreased

Given this information, determine which of the 3 songs the new user has bought. Justify your answer.

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2
<input type="checkbox"/>	3
<input type="checkbox"/>	4
<input type="checkbox"/>	5
<input type="checkbox"/>	6





Problem 5 Neural Networks and Gradient Descent (19 credits)

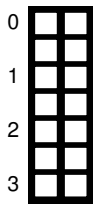
In the following, you will be working with the following simple neural network architecture f that produces an output $\hat{y}_i = f(x_i) \in \mathbb{R}^o$ for each given input data point $x_i \in \mathbb{R}^d$ (where x_i is given as a column vector). The neural network has *no hidden layers* and is thus given by

$$f(x_i) = g(Wx_i + b)$$

where W, b denote the network's weights and biases and g denotes some scalar activation function.

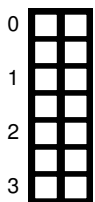
You have access to a dataset X comprising 200 000 houses and apartments ("housing units") that were sold across continental Europe in the previous five years. For each housing unit, the dataset contains a 12-dimensional feature vector with information about the unit. (*You can assume that these features have been adequately preprocessed and are always a valid input.*)

For subproblems you additionally know the sales price y_i in EUR for 100 000 of the housing units. Your goal in subproblems (a) and (b) is to use the neural network in order to perform a regression to predict the price of the remaining 100 000 units.



a)* Based on the information above, determine

1. The input dimension d , the output dimension o
2. A semantic interpretation of each output \hat{y}_i .
3. The dimensions of the weights and biases W and b .
4. The total number of *trainable parameters* of the network.



b) True or false: "By choosing the right activation and loss functions, you can use this neural net to implement ordinary linear regression (OLS)." If true, provide the appropriate activation function g and loss function L . If false, give a counterexample.





In addition to the sales price of the 100 000 labeled data points, assume that you also know how long each of these housing units was listed for sale. You now aim to predict whether a given housing unit was sold within one month of first being listed on the market (binary classification). You choose the *cross-entropy* loss function, and sigmoid activation functions $g(z) = \sigma(z)$.

Reminder: The cross-entropy loss is given by

$$L(y_i, \hat{y}_i) = y_i \cdot \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i).$$

The sigmoid function is given by

$$\sigma(x) = \frac{1}{1 + \exp(-x)},$$

and its derivative is

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)).$$

c)* For a single labeled data point (x_i, y_i) , calculate the gradient $\frac{\partial L}{\partial b}$ of the loss with respect to the bias b of the neural network. *Hint: Use backpropagation. Express your solution in terms of the data (x_i, y_i) and intermediate results from the forward pass, and simplify it as much as possible.*

	0
	1
	2
	3
	4
	5
	6
	7
	8





You split the 100 000 labeled data points into 50 000 points for training and 50 000 points for model validation. You train and evaluate both the "shallow" neural network f from exercise (5a), and a "deep" network which extends the model f by adding three hidden layers with 32 hidden units each. The results of these experiments are listed below, but you forgot which experiment corresponds to which model. After training the neural network using stochastic gradient descent (SGD), you observe the following empirical risk on the training (R_t) and validation (R_v) sets, respectively. (Assume that SGD has converged to a good approximation of the global optimum in both experiments, and further tweaking the optimization method will not yield better results.)

Experiment 1: $R_t < 0.001$, $R_v = 0.135$

Experiment 2: $R_t = 0.117$, $R_v = 0.114$

0	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>

d)* Which experiment belongs to which model? Justify your answer.

0	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>

e)* Are the results satisfactory to use one of these models for prediction on the 100 000 remaining unlabeled housing units? If yes, which model would you choose and why? If no, describe any problems you see and *one* possible way to overcome them?



Additional space for solutions—clearly mark the (sub)problem your answers are related to and strike out invalid solutions.

This image shows a full page of blank graph paper. The grid consists of small, equal-sized squares formed by thin gray lines. There are 20 columns and 20 rows of squares, creating a total of 400 square units. The grid covers the entire area of the page, leaving no margins or other markings.

