UNIVERSIDADE DE COIMBRA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
Departamento de Engenharia Informática

Exame de Recurso de Inteligência Artificial Centrada no Humano

2020/21 – 1º Semestre
English version
MECD

06-01-2021– 90min

<u>Important Note</u>: Fraud denotes a serious lack of ethics and constitutes a conduct that is not admissible in a student and future licensed professional. Any attempt at fraud leads to cancellation of the exam of both the facilitator as the harasser.

-- Write all your answers in the provided space –

<u>Important</u>: For the True vs False questions, you should consider that if you provide an incorrect answer, you receive a penalty of 30%, if you don't provide an answer you get 0%, and if you provide a correct answer you get 100%.

Q 1) (30%)

Consider the problem of predicting how many bikes of a company will be rented depending on the weather and calendar using linear regression. The data is presented in Fig. 1, while the feature characterization is shown in Fig. 2. Fig. 3 presents the estimated weights, standard error of the estimates (SE), and t-statistic, as computed by a Linear Regression toolkit.

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	01/01/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	02/01/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	03/01/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	04/01/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	05/01/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	06/01/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
7	07/01/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	08/01/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	09/01/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
10	10/01/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
11	11/01/2011	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263
12	12/01/2011	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25	1137	1162
13	13/01/2011	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38	1368	1406
14	14/01/2011	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54	1367	1421
15	15/01/2011	1	0	1	0	6	0	2	0.233333	0.248112	0.49875	0.157963	222	1026	1248
16	16/01/2011	1	0	1	0	0	0	1	0.231667	0.234217	0.48375	0.188433	251	953	1204
17	17/01/2011	1	0	1	1	1	0	2	0.175833	0.176771	0.5375	0.194017	117	883	1000
18	18/01/2011	1	0	1	0	2	1	2	0.216667	0.232333	0.861667	0.146775	9	674	683
19	19/01/2011	1	0	1	0	3	1	2	0.292174	0.298422	0.741739	0.208317	78	1572	1650
20	20/01/2011	1	0	1	0	4	1	2	0.261667	0.25505	0.538333	0.195904	83	1844	1927
21	21/01/2011	1	0	1	0	5	1	1	0.1775	0.157833	0.457083	0.353242	75	1468	1543
22	22/01/2011	1	0	1	0	6	0	1	0.0591304	0.0790696	0.4	0.17197	93	888	981
23	23/01/2011	1	0	1	0	0	0	1	0.0965217	0.0988391	0.436522	0.2466	150	836	986
24	24/01/2011	1	0	1	0	1	1	1	0.0973913	0.11793	0.491739	0.15833	86	1330	1416
25	25/01/2011	1	0	1	0	2	1	2	0.223478	0.234526	0.616957	0.129796	186	1799	1985
26	26/01/2011	1	0	1	0	3	1	3	0.2175	0.2036	0.8625	0.29385	34	472	506
27	27/01/2011	1	0	1	0	4	1	1	0.195	0.2197	0.6875	0.113837	15	416	431
28	28/01/2011	1	0	1	0	5	1	2	0.203478	0.223317	0.793043	0.1233	38	1129	1167
29	29/01/2011	1	0	1	0	6	0	1	0.196522	0.212126	0.651739	0.145365	123	975	1098
30	30/01/2011	1	0	1	0	0	0	1	0.216522	0.250322	0.722174	0.0739826	140	956	1096
31	31/01/2011	1	0	1	0	1	1	2	0.180833	0.18625	0.60375	0.187192	42		1501

Figure 1 – Extract (31 first instances) of the bike data set.

- instant: record index
- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Figure 2 – Feature characterization for the bike data set.

	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

Figure 3 – Estimated weight, the standard error of the estimate (SE), and the absolute value of the t-statistic (|t|).

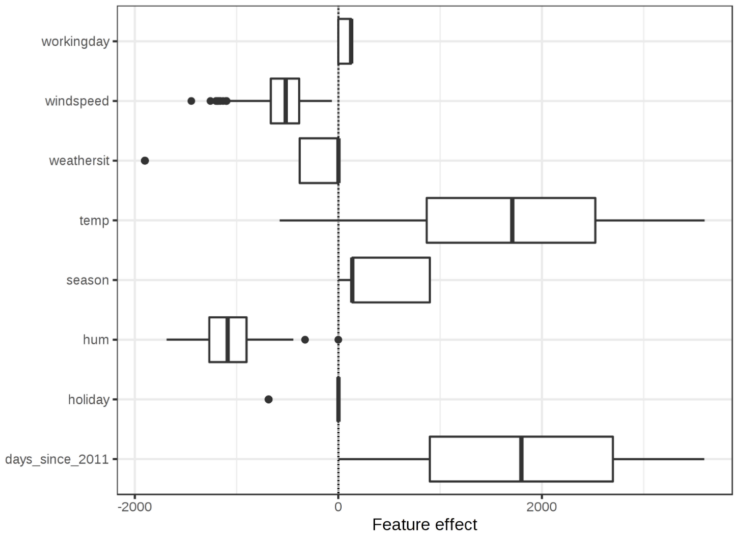

Figure 4 – Effect plot.

a) Compute the prediction for instance 31. How much has each feature of instance 31 contributed to this prediction (consider only the features in Fig. 3 and Fig. 4; provide calculations and mark the local effect of this instance in the feature effect plot in Fig. 4)? Provide also an explanation for the departure of this prediction from the average prediction of 4504 of the complete data set.

b) In this problem, we can say that (classify the following sentences as True or False):

		True?	False?
i.	Prediction is based on both numeric and categorical features.		
ii.	The estimation of missing data is made during the Data Pre-processing step of the Data Mining process.		
iii.	The application of the linear regression is part of the Data Analysis step.		
iv.	The importance of a feature increases with increasing weight.		
v.	The more variance the estimated weight has, the less important the feature is.		
vi.	Linear Regression can be used in model-agnostic methods designed for explaining black box models and does not require any information about the inner workings of the black box model, including its data set and its output predictions.		
vii.	An increase of the *weathersitMisty* by 1 decreases the predicted number of bicycles by 379.4, when all other features remain fixed.		
viii.	An increase of *temp* by 1 degree Celsius increases the predicted number of bicycles by 7.0, when all other features remain fixed.		

Q 2) (20%)

Consider the utility matrix below for the rating of six roads (M1,…,M6) by four drivers/users (A, B, C, and D) in which a Recommender System relies on. What is the prediction for the rating of road M4 by user C, using the typical prediction function and considering users A and B as the neighbour set, with similarities 0.7 and 0.4, respectively?

$$pred(a,p) = \overline{r_a} + \frac{\sum_{b \in N} sim(a,b) * (r_{b,p} - \overline{r_b})}{\sum_{b \in N} sim(a,b)}$$

	M1	M2	M3	M4	M5	M6
A	3	4	2	1	3	5
B	5	4	2	3	1	1
C	2	1	3		5	3
D	1	3	5	2	4	4

1 2 9 0 FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE D COIMBRA UNIVERSIDADE DE COIMBRA FACULDADE DE CIÊNCIAS E TECNOLOGIA *Departamento de Engenharia Informática*	**Exame de Recurso de Inteligência Artificial Centrada no Humano** **2020/21 – 1º Semestre** ***English version*** MECD 06-01-2021– 90min
<u>Important Note</u>: Fraud denotes a serious lack of ethics and constitutes a conduct that is not admissible in a student and future licensed professional. Any attempt at fraud leads to cancellation of the exam of both the facilitator as the harasser.	

-- Write all your answers in the provided space --

Q3) (30%)

You want to develop a classifier for automatically labeling short texts as funny or not. For this purpose, you can easily get large amounts of short texts from social networks like Twitter, but labeling may take time, and funniness is a subjective task.

a) Devise and describe a different approach for **conducting the labelling process** in each of the following scenarios:

(a.1) You have **limited time** but **unlimited funds**.

(a.2) Both your **time** and **funds are limited**.

b) Suppose that, for preliminary experiments, you rely exclusively on the following collection of four documents:

ID	Contents
D1	How do trees get online? They log in!
D2	What do trees do in Autumn? Take a leaf of absence.
D3	What did a tree say to another tree? Trees don't talk.
D4	Where do we get timber from? Tree logs.

(b.1) If the collection is represented by a traditional Vector-Space Model with TF-IDF weighting, how many dimensions / features would be required for representing each document? Consider that each document is previously pre-processed by: tokenization, expanding contractions (*don't -> do not*), removing punctuation signs and stopwords (you can use the following list: *how, do, get, in, what, a, of, to, where, we, from*), and lemmatizing plurals to singulars. Enumerate all the considered **features** and clarify any **assumption** made.

(b.2) For the feature "tree", compute:

- Its **IDF** in the collection (ignore the logarithm) = _____

- Its absolute frequency in D1 (TF) = _____

- Its TF.IDF in D1, considering both of the previous computations = _____

c) Funniness is often caused by **ambiguity** in language. What potential ambiguities can you identify in the presented documents? Explain **one** of them and name the language **knowledge level**(s) it is related to.

1 2 9 0 FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE D COIMBRA UNIVERSIDADE DE COIMBRA FACULDADE DE CIÊNCIAS E TECNOLOGIA *Departamento de Engenharia Informática*	**Exame de Recurso de Inteligência Artificial Centrada no Humano** **2020/21 – 1º Semestre** ***English version*** MECD 06-01-2021– 90min
Important Note: Fraud denotes a serious lack of ethics and constitutes a conduct that is not admissible in a student and future licensed professional. Any attempt at fraud leads to cancellation of the exam of both the facilitator as the harasser.	

-- Write all your answers in the provided space --

Q4) (20%)

Consider the scenario of Jarbas, an intelligent robot that supports the elderly in a Nursing Home. Jarbas is intended to follow the *Ethics Guidelines for Trustworthy AI* set up by the High-Level Expert Group on Artificial Intelligence.

a) (*Distribution task*) The main task performed by Jarbas is the distribution of drugs to each user according to the prescribed doses, at the recommended times. Give examples of acts or decisions that Jarbas must **not** take if it is to conform with the three components of Trustworthy AI (one example for each component)

1.

2.

3.

b) (*Entertainment task*) Consider that Jarbas, in the intervals of its main task, entertains the users by telling them jokes and funny stories. The robot knows about each user's profile and is quite competent in selecting the most convenient gags for each one of them.

Classify the following sentences as True or False (refer to the *Ethics Guidelines* above mentioned):

i. The compliance with the principle of *fairness* is key in the entertainment task (True/False)

-> Why? _____

_____

_____

ii. The compliance with the principle of *prevention of harm* is key in the distribution task (True/False)

-> Why? _____

_____

_____

iii. There may be a tension between the principles of *prevention of harm* and *respect for human autonomy* in the distribution task (True/False)

-> Why? _____

_____

_____

c) Give an example of action/decision that Jarbas must **not** take if it complies with the the requirement of *Diversity, non-discrimination and fairness*:

i. for the distribution task: _____

_____

_____

_____

ii. for the entertainment task: _____

_____

_____

_____