Thursday, March 9, 7:30AM

Final Exam

DATA 119
Winter 2023

Name: _____ **Student ID:** _____

1. Please **PRINT** your name and student ID in the above space.
2. This is a closed book, closed-notes examination. You can have a single two-sided page of notes that you may refer to.
3. Please provide the answers in the space provided. The spaces provided are sufficient for short and clear explanations or justifications.
4. Calculators of any sort are permitted.

Problem	Total Points	Received Points
Portuguese Math		
k-NN Classification		
HIV Test Results		
SQL		
Neural Networks		

Portuguese Math Scores

A dataset including records on 395 students in two Portuguese high schools includes information on the following variables[1]:

- `sex` - student's sex (binary: 0 - female or 1 - male)
- `age` - student's age (numeric: from 15 to 22)
- `Medu` - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- `Fedu` - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- `internet` - Internet access at home (binary: 0 - no or 1 - yes)
- `famrel` - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- `Walc` - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- `absences` - number of school absences (numeric: from 0 to 93)
- `G1` - first period grade (numeric: from 0 to 20)
- `G2` - second period grade (numeric: from 0 to 20)
- `G3` - final grade (numeric: from 0 to 20, output target)

Summary statistics for the numeric variables appear below.

Variable	Minimum	Q1	Median	Q3	Maximum	Mean	Std. Dev.
absences	0	0	2	6	32	3.66	4.64
age	15	16	17	18	22	16.74	1.22
famrel	1	4	4	5	5	3.93	0.96
Fedu	0	1	2	3	4	2.31	1.10
G1	0	10	11	13	19	11.40	2.75
G2	0	10	11	13	19	11.57	2.91
G3	0	10	12	14	19	11.91	3.23
Medu	0	2	2	4	4	2.51	1.13
Walc	1	1	2	3	5	2.28	1.28

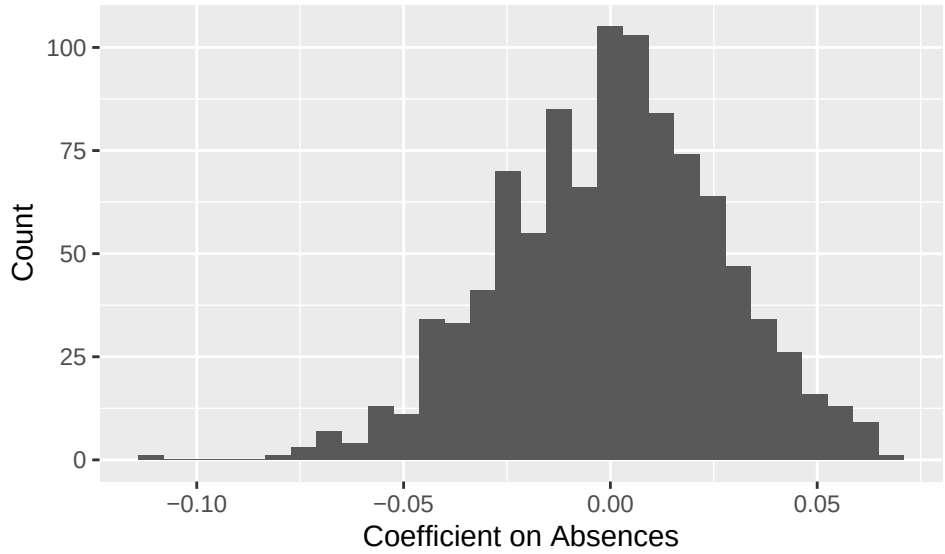
[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Techhnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

In addition, we have access to the correlation matrix for the different variables.

	sex	age	Medu	Fedu	internet	famrel	Walc	absences	G1	G2	G3
sex	1.00	-0.04	0.12	0.08	0.07	0.08	0.32	0.02	-0.10	-0.10	-0.13
age	-0.04	1.00	-0.11	-0.12	0.01	-0.02	0.09	0.15	-0.17	-0.11	-0.11
Medu	0.12	-0.11	1.00	0.65	0.27	0.02	-0.02	-0.01	0.26	0.26	0.24
Fedu	0.08	-0.12	0.65	1.00	0.18	0.02	0.04	0.03	0.22	0.23	0.21
internet	0.07	0.01	0.27	0.18	1.00	0.08	0.06	0.07	0.14	0.15	0.15
famrel	0.08	-0.02	0.02	0.02	0.08	1.00	-0.09	-0.09	0.05	0.09	0.06
Walc	0.32	0.09	-0.02	0.04	0.06	-0.09	1.00	0.16	-0.16	-0.16	-0.18
absences	0.02	0.15	-0.01	0.03	0.07	-0.09	0.16	1.00	-0.15	-0.12	-0.09
G1	-0.10	-0.17	0.26	0.22	0.14	0.05	-0.16	-0.15	1.00	0.86	0.83
G2	-0.10	-0.11	0.26	0.23	0.15	0.09	-0.16	-0.12	0.86	1.00	0.92
G3	-0.13	-0.11	0.24	0.21	0.15	0.06	-0.18	-0.09	0.83	0.92	1.00

(A) (5 points) Find the simple-least squares equation for `G3` as a function of `absences`.

A permutation test can be used to estimate the distribution of the simple linear regression coefficient of the `absences` variable. It results in the below distribution of the linear regression coefficient (summary statistics on the following page).



Min.	2.5^{th}%-ile	Q1	Med.	Q3	97.5^{th}%-ile	Max.	Mean	Std. Dev.
-0.11	-0.05	-0.02	0	0.02	0.05	0.07	0	0.03

(C) (2 points) Comment on the coefficient on `absences` estimated using the permutation test. Would you expect the permutation test to result in a p-value less than 0.05 for the coefficient on `absences`? Why or why not?

(C) (2 points) Comment on the sign of the intercept of this one-parameter linear regression.

(D) (2 points) Is there a valid interpretation of the intercept in the simple linear regression model? Explain.

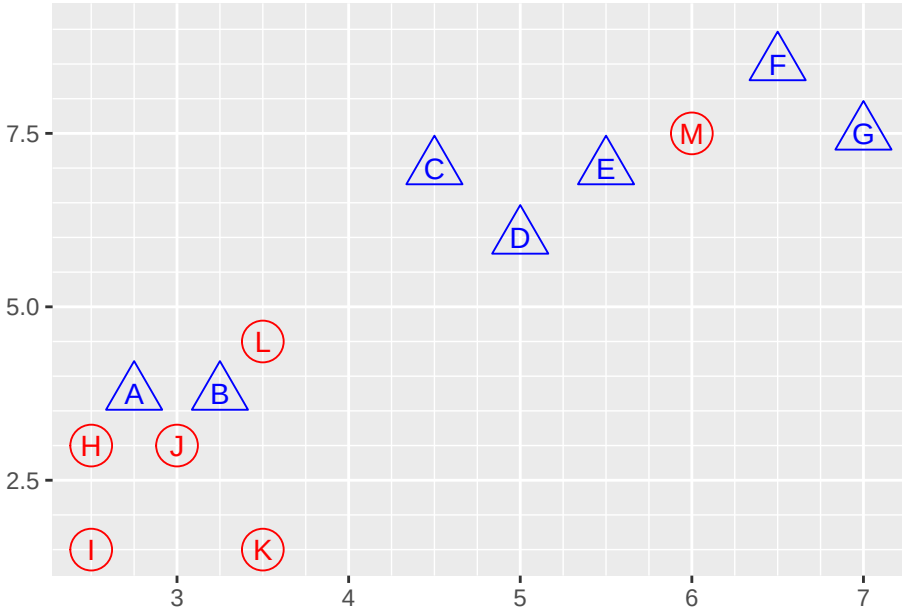
(E) (2 points) What is the proportion of variation in `G3` that can be explained by the variation in `absences`?

(F) (4 points) A table of the multiple linear regression coefficients for predicting final grades appears below. Predict the final grade for a 16 year-old student whose parents have higher education degrees. She has internet access, an excellent family relationship, low weekend alcohol consumption, two recorded absences, a first period grade of 16, and a second period grade of 18.

	Coefficients
(Intercept)	0.21
sex	-0.16
age	-0.01
Medu	-0.04
Fedu	0.03
internet	0.13
famrel	-0.05
Walc	-0.06
absences	0.02
G1	0.15
G2	0.89

kNN Classification

Observations A-G (triangles) belong to Class 1, and observations H-M (circles) belong to Class 2. All of the observations are shown in the plot below.



Point	A	B	C	D	E	F	G	H	I	J	K	L	M
Neighbor 1													
Neighbor 2													
Neighbor 3													

(A) (2 points) Using Euclidean distance, identify the 3 nearest neighbors for point J.

(B) (2 points) Predict the class for point J, based on simple voting with k=3.

(C) (1 point) Let Class 1 be the absence of some event, and Class 2 be the presence of some event. Is Point J a true positive, true negative, false positive, or false negative?

(D) (4 points) Now, predict the class for each point using the three nearest neighbors (if you like, you can use the table above to keep track of your work). Then, fill out the confusion matrix below. Only the matrix will be graded, you can use the righthand side for scratchwork.

		Truth	
		Class 1	Class 2
Predicted	Class 1		
	Class 2		

(E) (2 points) Would increasing the number of neighbors to 5 affect the accuracy of the algorithm? How?

HIV Test Results

In 2008, researchers ran an experiment in Malawi to see whether cash incentives could encourage people to learn the results of their HIV tests.[2] The following data were collected for 2,825 patients.

- `got`: Whether a patient learned the results of their HIV test.
- `distvct`: Distance in kilometers from their residence to the testing center.
- `tinc`: Total incentive received by the patient.
- `any`: Whether the patient received any incentive.
- `age`: The patient's age.
- `hiv2004`: The patient's HIV test results.

The researchers wanted to see which of the variables could predict whether a patient would get their test results. A plot of the model coefficients for a logistic regression model estimated with LASSO using various values of $\log(t)$, along with a table of the intercept at each value of $\log(t)$ and a table of the test set AUC at each value of $\log(t)$.

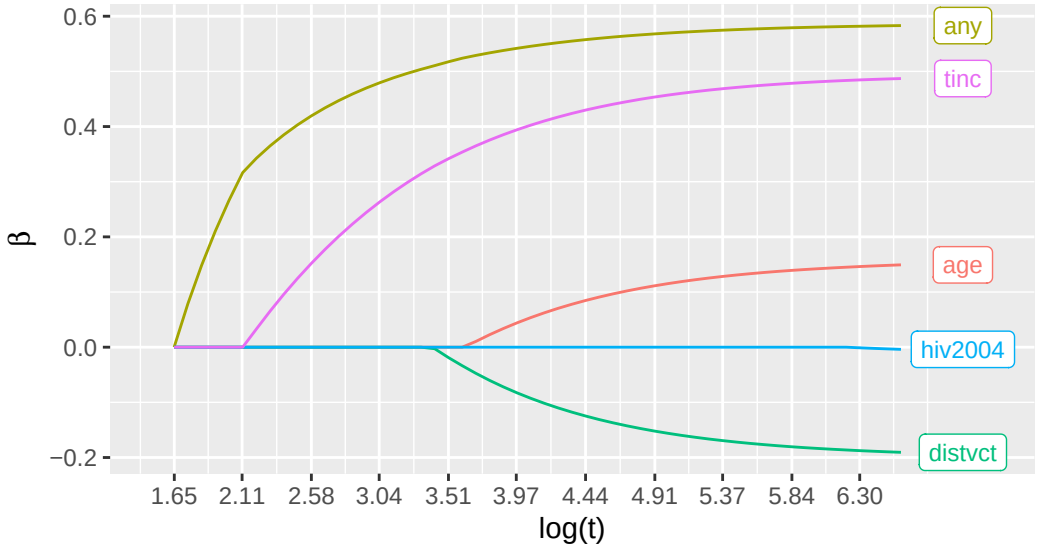


Table 1: Value of the Intercept at $\log(t)$

$\log(t)$	1.65	2.11	2.58	3.04	3.51	3.97	4.44	4.91	5.37	5.84	6.30
Int.	0.81	0.82	0.85	0.88	0.90	0.91	0.92	0.93	0.94	0.94	0.95

Table 2: Value of the Test Set AUC at $\log(t)$

$\log(t)$	1.65	2.11	2.58	3.04	3.51	3.97	4.44	4.91	5.37	5.84	6.30
AUC	0.50	0.67	0.73	0.73	0.74	0.67	0.62	0.61	0.60	0.60	0.59

[2]Thornton, Rebecca L. 2008. 'The Demand for, and Impact of, Learning HIV Status.' American Economic Review 98 (5): 182963.

(A) (1 point) Which value of $\log(t)$ should be used for the LASSO model?

(B) (2 points) What is the effect of large values of $\log(t)$ on the error on the training dataset?

(C) (3 points) Using the value of $\log(t)$ from question 1, write the approximate equation for predicting whether a patient learns their results (two decimal places for the coefficients are fine). Which coefficients are most important?

(D) (2 points) The probability of a 40 year-old HIV-negative patient living 1.8 kilometers from the testing center and receiving a $2 incentive learning their HIV test results is 86.3%. What are the odds that they learn their results?

(E) (3 points) A confusion matrix for this model appears below. Estimate the positive predictive value if the prevalence of HIV positive patients in Malawi is 7%.

		Truth	
		Did Not Learn	Learned
Predicted	Did Not Learn	64	66
	Learned	197	521

SQL

You have three tables in an sqlite database: one table (filehashes) of file paths and md5 hashes, one table (filesizes) of file paths and file sizes, and one table (filehashes2) of a (modified) copy of the set of files on a removable device. We are interested in finding the numbers and the names of files that are duplicates, the numbers and names of files that have been changed, and the sizes of files corresponding to queries such as:

```
sqlite> select * form filehashes limit 4;
path md5
/Users/gornuk/119/lab/lab01.ipynb 277fc2c7454565a1d74182c8c6e20446
/Users/gornuk/119/lab/lab03.ipynb d261e060533c30cc143c4f8f916527e5
/Users/gornuk/119/lab/lab07.ipynb dce13be8c37fd1b353a80602149947cd
/Users/gornuk/119/lab/lab05.ipynb c6b4f0333a8620e86ae7aa0d4c9b52fa

sqlite> select * from filesizes limit 2;
path size
/Users/gornuk/119/lab/lab01.ipynb 20985
/Users/gornuk/119/lab/lab03.ipynb 32196

sqlite> select * from filehashes2 limit 2;
path md5
/Volumes/USB/119/lab/lab01.ipynb 277fc2c7454565a1d74182c8c6e20446
/Volumes/USB/119/lab/baby.csv 304b69e4133a528f6df223a7d5890e89
```

Here are the sizes of the tables:

```
sqlite> select count(*) from filehashes;
count(*)
914

sqlite> select count(*) from filehashes2;
count(*)
849

sqlite> select count(*) from filesizes;
count(*)
918
```

And here are some distinct calculations:

```
sqlite> select count(distinct(path)) from filehashes;
count(distinct(path))
914

sqlite> select count(distinct(md5)) from filehashes;
count(distinct(md5))
827

sqlite> select count(distinct(path)) from filehashes2;
count(distinct(path))
849
```

(A) (1 point) Why is `count(distinct(md5))` smaller than `count(distinct(path))`?

(B) (2 points) Write an SQL expression to count the number of files common between the filehashes and the filesizes tables.

(C) (2 points) Write an SQL expression to report the total of all the file sizes of the files common to the two tables indicated above.

(D) (3 points) Write an SQL expression to retrieve the list of md5s that are present in filehashes more than once and the number of times each is duplicated.

(E) (3 points) Write a command that shows the names of all the file pairs between filehashes and filehashes2 that are duplicated. Sample output below:

```
/Users/gornuk/data119/data/dnam-test.csv /Volumes/USB/119/slides/dnam-test.csv
/Users/gornuk/data119/data/faithful.csv /Volumes/USB/119/faithful.csv
```

```
/Users/gornuk/data119/data/faithful.csv /Volumes/USB/119/lab/faithful.csv
/Users/gornuk/data119/data/faithful.csv /Volumes/USB/119/slides/done/faithful.csv
```

(F) (3 points) Write a command that reports the md5, a path, and the number of copies of files that are duplicated. Sample output below:

```
5698b4a0bef1bbaf70fc6aeb3305bed2 /Users/gornuk/data119/homeworks/Problem-wordle_A.ipynb.  2
5c5ea34ba78f09e4673ec87f393bf16e /Users/gornuk/data119/homeworks/WORK/HW01_Spring22-NEW.ipynb.  2
6079d39997a49fc8d68f9aa26774cada /Users/gornuk/data119/homeworks/imdb/imdb-2010.db 3
6d8c89606b781ebabeed202ebee909e0 /Users/gornuk/data119/homeworks/HW01_Spring22.ipynb 2
```

(G) (3 points) Write a command that reports the md5, a path, and the total file size (Hint: the product of the file size and the number of copies) for each md5 that is present in filehashes more than once.

Neural Networks

Here is a slightly modified version of a neural network model that we ran in class. It took 6 columns of numerical titanic data (`X_train`) and was fit against the indicator variable `Y_train` which was 1 if the passenger survived the ordeal.

```
# define the keras model
model = Sequential()
# add the first layer - note that you need to specify the size of input
# double duty: defining the input or visible layer and the first hidden layer
model.add(Dense(12, input_dim=6, kernel_initializer='normal', activation='sigmoid'))
# add a second layer
model.add(Dense(12, kernel_initializer='normal', activation='sigmoid'))
# this is the last layer which is the output
model.add(Dense(1, activation='linear'))
# compile the model - see below
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
# fit the model
model.fit(X_train, Y_train, epochs=1000, batch_size=50)
```

(A) (3 points) Sketch this neural network model, showing the number of nodes.

(B) (3 points) Estimate the number of parameters (weights) in this neural network (to within 20%).

(C) (2 points) This model has tolerable accuracy on a holdout set: about 83%

```
model.evaluate(X_test, Y_test)
7/7 [==============================] - 0s 3ms/step - loss: 0.1231 - accuracy: 0.8325
```

but the following fact about the model predictions is somewhat unusual:

```
predictions=model.predict(X_test)
7/7 [==============================] - 0s 3ms/step
np.min(predictions), np.max(predictions)
(-0.010760196, 1.1518762)
```

Why is the domain not (0,1) or 0, 1 ? What would you need to change to fix this?