RUBRIC: Conceptual

Streaming Service Shenanigans

You are a data scientist at *NetPrime*, a popular movie and series streaming company. It's very important for your business model that your servers are up at all times, but recently, lots of users have been reporting that they cannot access some movies. You look at the logs, and you notice that lots of your machines are crashing. Therefore, you are modeling the time it takes for a computer system to crash after start-up (`time_to_crash`).

Task 1 (10 points)

(a) Given that `time_to_crash` is numerical, you model it using linear regression. You use your model to predict `time_to_crash` for a new computer system given the brand (`brand_computer_system`) and the age (`age_of_computer`). The model predicts that the time it will take to crash after start-up is −60 seconds. How is this possible?

2 points total

```
We are modeling using a Gaussian link function

Anything about linear regression being able to predict negative values is
accepted
```

(b) Fix the problem, i.e. make sure that your model only predicts `time_to_crash` >= 0. Provide your regression equation and explain your solution.

5 points total

3 for correct family (exponential)

2 for correct regression equation (any variation as long as it is justified)

```
(3 points) The exponential distribution deals with the time between
occurrences of successive events as time flows by continuously.

Poisson is NOT correct (the Poisson distribution deals with the number of
    occurrences in a fixed period of time)


Equation:

Lmer(time to crash ~ brand_computer_system + age_of_computer,
    family='exponential', data=data)
```

(c) Your team is studying the mutual information (MI) between the features `time_to_crash` and `age_of_computer`. Using the same data, one colleague says the MI between these variables is 0.6, while another one says that it is 0.8. Why do they have different values?

3 points total

```
Different number of bins
```


Task 2 (12 points)

Because of these outages, a considerable number of users have canceled their subscriptions. You want to build a classification model. Using the information from the previous month, you want to predict if a user will cancel their subscription in the next month. You have the following features (from the previous month): the number of days in which the user was active, the median hours watched, and the number of different series the user watched.

You want to build a decision tree as a baseline classifier. You take a random sample of 10 points.

(a) Using only the 10 points sample, on which attribute (name the attribute) should you do the first split and why?

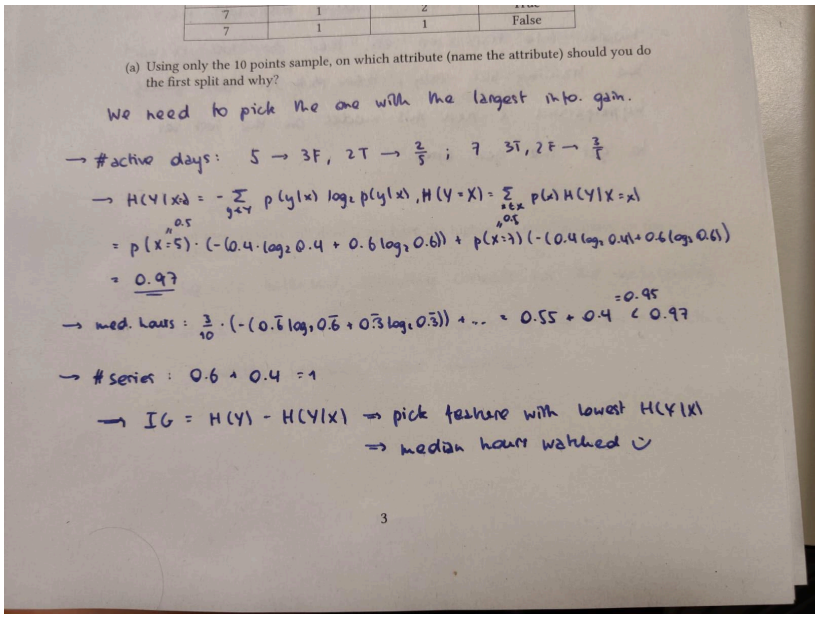
9 points total

3 mentioning IG (information gain)

2 for correct answer -> median hours watched

4 points for computation (correct formula):

2 for active days and median hours and 1 for series



7	1	2	
7	1	1	False

(a) Using only the 10 points sample, on which attribute (name the attribute) should you do the first split and why?

We need to pick the one with the largest info. gain.

→ # active days: 5 → 3F, 2T → $\frac{2}{5}$; 7 3T, 2F → $\frac{3}{7}$

→ $H(Y|x_2) = -\sum_{y \in Y} p(y|x) \log_2 p(y|x)$, $H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x)$

$= p(x=5) \cdot (-(0.4 \cdot \log_2 0.4 + 0.6 \log_2 0.6)) + p(x=7)(-(0.4 \log_2 0.4 + 0.6 \log_2 0.6))$

$= 0.97$

→ med. hours : $\frac{3}{10} \cdot (-(0.\overline{6} \log_2 0.\overline{6} + 0.\overline{3} \log_2 0.\overline{3})) + \ldots = 0.55 + 0.4 = 0.95 < 0.97$

→ # series : $0.6 + 0.4 = 1$

→ IG = $H(Y) - H(Y|X)$ → pick feature with lowest $H(Y|X)$

→ median hours watched ☺

3

3 for showing work

(b) Having built the decision tree, you ask your interns to find the best hyperparameters for the model. They want to tune the following parameters: `min sample split` (2, 3, 4) and `max depth` (2, 3, 4). They use 5-fold cross validation to tune the hyperparameters and assess the generalization error of the model. They obtain a `min_sample_split` of 2 and `max_depth` of 2 as optimal hyperparameters. Here are their results per fold:

- Accuracy Fold 1: 0.90
- Accuracy Fold 2: 0.80
- Accuracy Fold 3: 0.85
- Accuracy Fold 4: 0.80
- Accuracy Fold 5: 0.90

They report that the generalization error of the model is 0.85.

Explain what is wrong with their evaluation procedure and how you would improve their work.

Model selection without nested CV uses the same data to tune model parameters and evaluate model performance. Information may thus "leak" into the model and overfit the data.

3 points total

2 for showing what's wrong (no nested cross-validation or extra set)

1 for improvement

MOOC Mania

You decide to quit your previous job as a streaming service data scientist because you realize that your life calling is in teaching data science instead! You teach a MOOC about data science on the platform CoursEdX.

Your class is wildly popular with thousands of students enrolled. However, you notice that there are a number of bots pretending to take your class and posting random things on the discussion forums. You decide to address this problem by building a spam filter.

Task 3 (7 points)

You continue labeling discussion posts in your course to identify spam. You see that the dataset is severely imbalanced, with 80% of posts with no-spam and 20% of posts with spam.

You consider the following three models:

- *MAJORITY* always predict class no-spam
- *UNIFORM* predicts class no-spam 50% of the time and class spam otherwise
- *PROP* predicts class no-spam 80% of the time and class spam otherwise

(a) Which of these models will most likely have the highest accuracy on the dataset? Explain.

3 points total

1 for correct answer

2 for showing work

Answer accuracy (class A and B):

- **MAJORITY**: 1*80% + 0*20% = 80%
- UNIFORM: 0.5*80% + 0.5*20% = 50%
- PROP: 0.8*80% + 0.2*20% = 68%

(b) Which of these models will most likely have the highest balanced accuracy on the dataset? Explain.

4 points total

2 for correct answer

2 for explanation (why not majority, why not prop)

Answer BAC (class A and B): **ALL**

- MAJORITY: 1/2 + 0/2 = 50%
- UNIFORM: 0.5/2 + 0.5/2 = 50%
- PROP: 0.8/2 + 0.2/2 = 50%

5

Task 4 (2 points)

To deal with the class imbalance of your discussion post data, you decide to use resampling methods instead of a single model with a train-test-val split. It's been a few years and you're a bit confused about the details of the different approaches. Which of the following is true regarding evaluating these models?

Circle the correct answer, no justification required.

(a) Out-of-Bag error estimate uses a separate validation set.

(b) K-fold cross-validation uses a separate validation set.

(c) Out-of-Bag error estimation is more computationally expensive than K-fold cross validation.

(d) Out-of-Bag error estimation provides an unbiased estimate of generalization error. (correct)

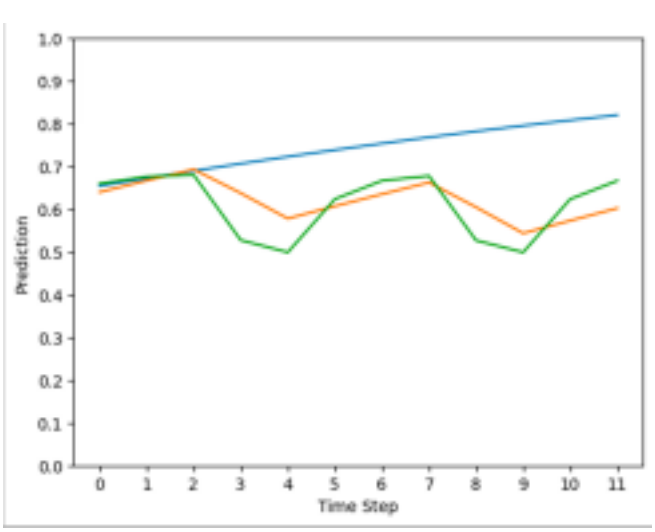
(e) K-fold cross-validation provides an unbiased estimate of generalization error. (not correct)

2 points total

B + D, or either (as the question wording makes it seem like one answer)

6

Now that you are sufficiently happy with bots not taking over your discussion forums, you decide to focus on improving students' performance. You model your student performance data and analyze how they learn skills.



Task 5 (9 points)

You train an AFM, a PFA, and a BKT model on your students' data (their correct or wrong answers on skills taught in the

course). You obtain the following model parameters from the trained models:

You use these three models to predict the probability that a student has learned the skill *s* at time step *t* over the first 12 time steps. The graph shows three models' predictions for a student with observed question interactions [1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1], where 1 represents a correct answer and 0 represents incorrect.

Which model generates each line? Why?

9 points total, 3 for each line

2 for Blue - AFM, 1 for explanation (only increasing)

2 for Orange - PFA (1 for BKT, 0 for AFM), 1 for explanation

2 for Green -BKT (1 for PFA, 0 for AFM), 1 for explanation

Task 6 (15 points)

You realize that your student performance dataset actually has three labels: correct, partially correct, wrong. Since BKT only has correct and wrong options for student answers, you decide to adjust the BKT model to be able to handle these three different outcomes.

(a) What is the predicted probability that a student solves the first question correctly?

7 points total

3 for correct answer *0.52*

4 for showing work with $(1-p\_s -p\_hs) \cdot p\_0+p\_g \cdot (1-p\_0)$

Solution: $p(o\_0=correct)=(1-p\_s -p\_hs) \cdot p\_0+p\_g \cdot (1-p\_0) = 0.7*0.7+0.1*0.3=0.52$

(b) What is the predicted probability that a student has mastered the skill before the second attempt, given that they got their first attempt correct?

8 points total

3 points for $p(s\_0=mastered|o\_0=correct)$

3 points for $p(s1 = mastered|o0 = correct)$

2 points for correct end probability: 0.758

$p(s\_0=mastered|o\_0=correct)$

= (p(o_0=correct | s_0=mastered) · p(s_0=mastered))/(p(o_0=correct)) =
*(1-ps-phs)*p0/0.52 = 0.7*0.7/0.52 = 0.94*

p(s0=not mastered|o_0=correct) = 0.06

p(s1 = mastered|o0 = correct)
= p(s0=mastered|o_0 = correct)(1-pf) + p(s0=not mastered|o0=correct)*pl*
= 0.94(1-pf) + 0.06*pl = 0.758*

9

Task 7 (10 points)

Now that your BKT model is finalized, you want to design interventions for your failing students to get them back on track. You have last year's data from a related class with some features describing each assignment, student performance on each assignment, and their overall performance in the class. Using these, you design a pass-fail prediction LSTM. Before deploying it for all of your students, you want to carefully analyze the model.

Specifically, you are interested in finding which assignments in the course are most predictive of student success. In this way, you can identify which assignments to spend time improving with your limited TA support and which ones are less important. Using your model, how can you design an explainability analysis to help with this? Which method(s) would you use? Justify your decisions.

10 points total

2 points for mentioning LIME or PDP

3 points for looking at weeks and correlating to assignments

3 points for analysis of why this will work

2 points for mentioning "global"

Answer: Global explainability analysis. LIME on all students, features for each assignment, seeing which have the highest importance. PDP on features, examining weeks.

10

Task 8 (10 points)

Your pass-fail model seems to be working well, but you now want to design a time-series model with a separate quiz score prediction for each week to model your students' changing performance and design interventions at the right time in the course. You already have working code for a traditional pass-fail student success prediction LSTM network. You now want to modify it to a time-series prediction of

student quiz performance each week.

State all the changes you need to make to the model architecture below and explain why.

```
def create_model(nb_features, nb_skills, params):

inputs = tf.keras.Input(shape=(None, nb_features), name='inputs')

x = tf.keras.layers.Masking(mask_value=params['mask_value'])(inputs)

x = tf.keras.layers.LSTM(params['recurrent_units'],

                         return_sequences=False,

                         dropout=params['dropout_rate'])(x)

dense = tf.keras.layers.Dense(nb_skills, activation='sigmoid')

outputs = dense(x)

model = tf.keras.models.Model(inputs=inputs, outputs=outputs, name='TimeSeries')

model.compile(loss=tf.keras.losses.binary_crossentropy,

              optimizer=params['optimizer'],

              metrics=[tf.keras.metrics.AUC(), 'binary_accuracy'])

return model
```

10 points total

5 changes, 2 points each

Answer: Time Distributed Layer, change sigmoid to softmax / linear, change loss function from binary cross entropy to MSE, change metrics, return_sequences = True, change the output shape

TrioLingo: Leisurely Language Learning

Now that things are going smoothly with your MOOC course, you decide to pick up language learning as a hobby. You find a very cool app – TrioLingo, and decide to involve all of your friends.

Task 9 (8 points)

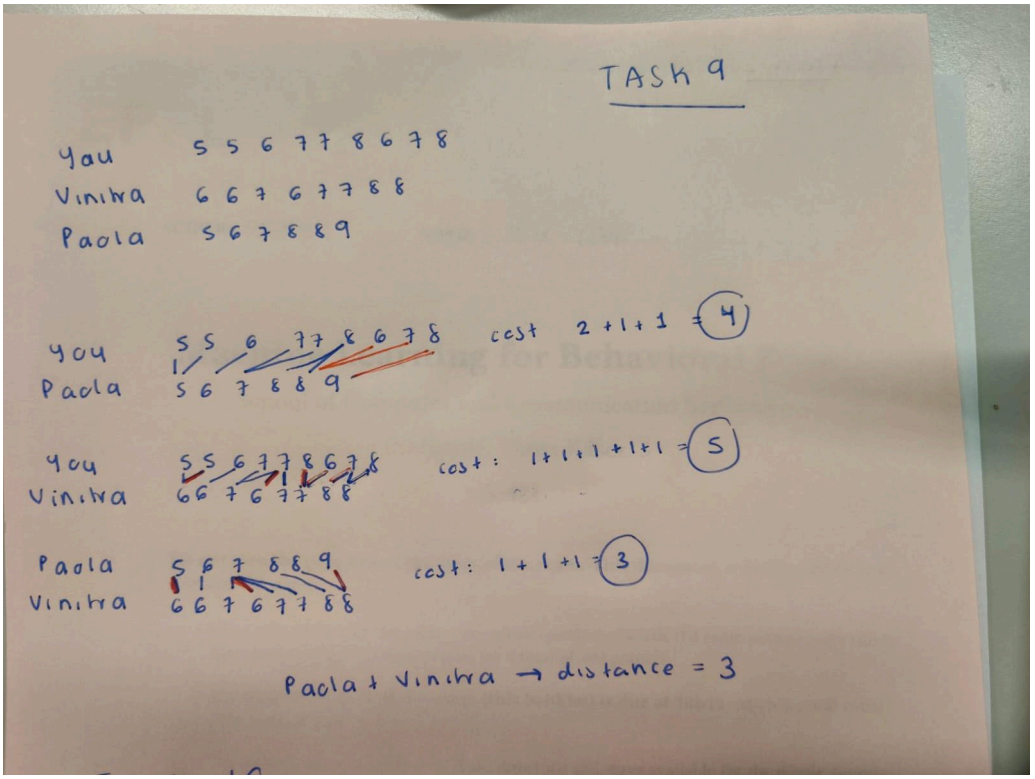
You, Paola and Vinitra are trying to master Italian by the beginning of August so you can go on a summer vacation together. During the month of July, your first exercise was 5 points, second exercise was 5 points and 6,7,7,8,6,7,8 points respectively on the following days over 9 total sessions. Likewise, Paola scored 5,6,7,8,8,9 points over 6 sessions and Vinitra scored 6,6,7,6,7,7,8,8 points over 8 sessions. Using dynamic time warping (minimizing L1 distance), calculate the distance (minimum distance path) between the three TrioLingo participants (You, Paola and Vinitra). Which two TrioLingo learners had the closest study performance and what was their distance? Show your work.

8 points total

2 for correct pair

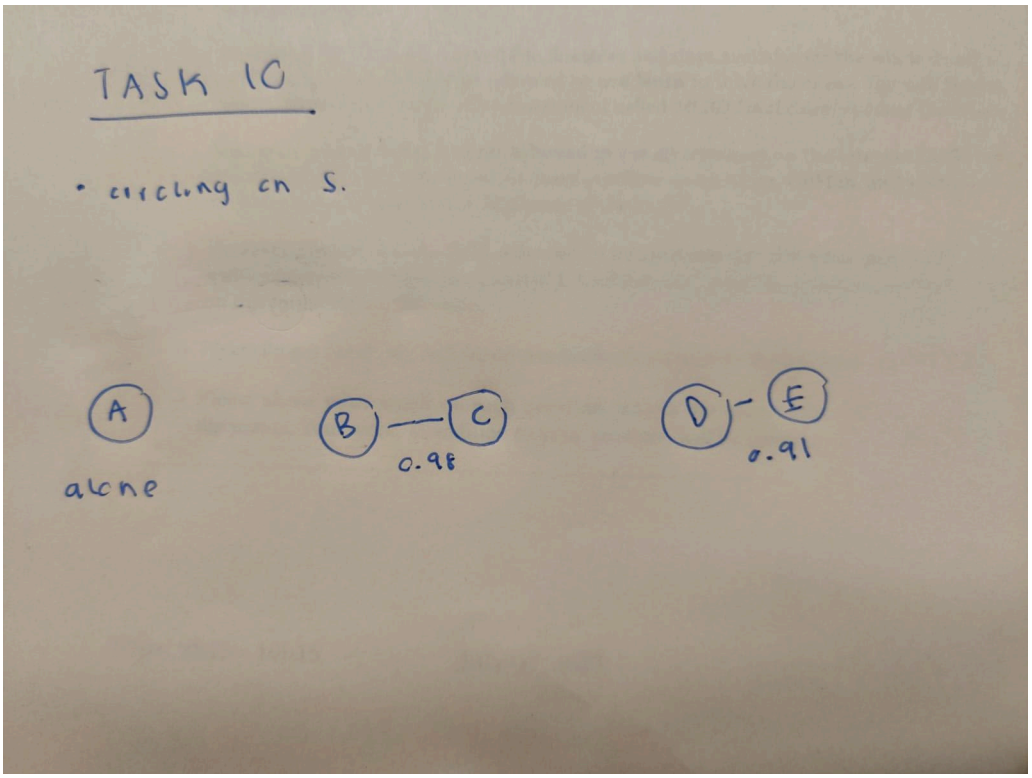
2 for correct distance

4 for work



Task 10 (8 points)

You convince five more of your friends to also learn Italian on TrioLingo with you. You would like to create subgroups based on their Italian level. Matrix ﷺ shows the Italian-level similarity between the five of them: Alex, Bastien, Camille, Dominique and Evan (in that order). For example, Alex and Bastien are 43% similar, and Dominique and Bastien are 65% similar. Draw the resulting similarity graph after applying the mutual 1-nearest graph algorithm. Who will be grouped together? Show your work.



8 points total

3 for grouping A , B-C and D-E

3 for similarity graph (with number)

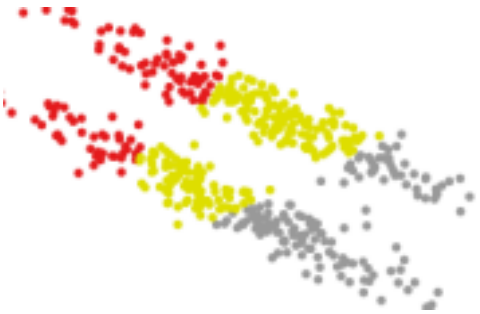
2 for showing work (circling)

Task 11 (6 points)

Your friend, an aspiring data scientist, used the following method to cluster the TrioLingo users:

```
cluster.KMeans(n_clusters=3,init='random',n_init=1)
```

He shows you the following results:



What three pieces of advice would you give him to improve his solution and why?

6 points total

For each advice (x3): 1 for correct advice, 1 for correct reasoning

Answer:
Choose K with a metric (half point for cluster two groups instead of three)
Use spectral clustering
Not random init
Multiple inits

Task 12 (3 points)

Learning a language is harder than expected and you're getting a bit discouraged. You want to see how your peers are performing on TrioLingo to see if you really have a chance of learning Italian in time for the trip. To collect data on EPFL students' use of Triolingo, you stand outside near the Esplanade and ask 200 students for their usage statistics. Afterwards, you design a model on this data and you note that the predictions are unfair for certain cases. What is the most likely kind of bias you identify in your subset? Explain why.

3 points total

1 for representation bias

2 for explanation (sample of students outside are not representative of global triolingo population. Time of day, etc. have factors)