ECON484 Machine Learning
Midterm Exam Part 1
23.03.2022

Student ID		**Question**	**Points**
Student Name and Surname		**1**	____ / 6
This part of the exam is open book. Any printed material (textbooks, slides, notes students prepated, etc) are allowed. However, students cannot exchange material. **Access to online resources and use of mobile phones is not allowed.** Calculations in the exam will not require a calculator. Part 2 of the exam will be completed by code submission through Github. **To qualify for Part 2, students must attend Part 1.** Please type your name on top left corner of each page. Staples tend to get separated from paper. **Exam duration is 50 minutes.**		**2**	____ / 12
		3	____ / 12
		4 (Bonus)	____ / 10
		Total	____ / 30 (40)

Question 1 (True/False, 1 points each, total 6 points)

Please indicate if these statements are true or false.

_**T**_ Computer algorithms are designed to solve a very particular problem given the set of applicable constraints which have been assessed during an analysis (or modeling) phase.

_**T**_ A data structure is a design on how a computer program stores and accesses its data in the computer's memory.

_**T**_ All machine learning methods mimic human or systems behavior to a degree so that we tend to obtain the results that represent an "average" quality of our sample.

_**F**_ When using machine learning in an operational environment, if there is a systemic change in the environment, machine learning never catches up with the environment.

_**F**_ The easiest way to evaluate machine learning behavior is to develop a key performance indicator (KPI) that measures real world success.

_**T**_ Many machine learning problems can be reformulated as typical parameter estimation problems.

Question 2 (12 points)

You are working as a data scientist at an autonomous vehicle company. Your task is to train the vehicle AI about making right or left turns. Here is what you do as development.

1. You find some very good drivers and record their turning behavior in an isolated test drive facility. Then you train the AI with the data you collected.
2. You test your autonomous vehicle individually in the same test drive facility. It makes the turns almost perfectly.

3. You further test the autonomous vehicle by having multiple autonomous vehicles creating a simulated traffic. Some of the vehicles will go straight ahead, some will turn left and some will turn right. This test is also very successful.

After your successful test results, you are given the authority to release the autonomous vehicle into actual traffic for further testing. This is where things go wrong.

1. When your vehicle is turning right, it moves so close to the sidewalk that pedestrians are scared.
2. When your vehicle is turning right, it slows down, so that sometimes human drivers driving behind your vehicle crash into your vehicle.
3. When your vehicle is turning left, it continues to use its assigned lane so perfect that, human drivers in its left lane that turn left but cannot use their lane perfectly crash into your vehicle.

Now you are required to explain what happened.

(a) (5 points) Is there a technical term for what's wrong with your AI?

This is a very good example of over-fitting.

(b) (5 points) How can you (partially) rectify this problem without making any more recordings of professional drivers?

We can rectify this problem by modifying the optimization target for the overall system, ie. introduce some penalty for too perfect results. This can be done by defining the penalty in terms of the weights of the optimization formula.

(c) (2 points) What is the technical term for this corrective technique?

Hyperparameter Optimization

Question 3. (12 points)

On the issue of whether or not politicians make reliable statements, you select **one particular politician from the 1960s** and conduct an analysis on his statements about inflation.

This particular politician has stated 1.200 times that inflation was higher than before, of which 800 times the statement is correct and 400 times the statement is incorrect. He has also stated that the inflation was lower than before 800 times, of which 700 times the statement is correct and 100 times the statement is incorrect.

You want to understand if you used this politician as an estimator fo inflation being higher or lower than before, how would this estimator perform.

(a) (6 points) Construct a Confusion Matrix for this politician's assessments.

		Reality Is	
		High	**Low**
Says	**High**	800	400
	Low	100	700

(b) (6 points) Calculate True Positive Rate, False Negative Rate, and Accuracy.

TPR = TP / (TP + FN) = 800 / (800 + 100) = 0,888

FNR = FN / (TP + FN) = 100 / (800 + 100) = 0.111

Accuracy = (TP + TN) / (P + N) = (800 + 700) / (900 + 1.100) = 0.75

Question 4 (Bonus, 10 points)

Suppose you have this study done for **1.000 politicians from many countries**. And you want to classify them into two groups, **optimists** and **pessimists**. You use the data set of True Positive Rate, False Negative Rate, and Accuracy values to classify.

(a) (4 points) Because you don't know about a lot of classification techniques yet, you decide to use k-nearest neighbor. Using the particular politician above as a hint, what could go wrong with k-NN in this particular problem?

Most politicians are optimists in their speech. Therefore the sample would be overcrowded with optimists. When trying to classify pessimists, their neighborhood would be filled with optimists, hence wrongfully classifying them as optimists.

(b) (4 points) Can you come up with **better indexes** than those you calculated in Question 3 so that your classification problem is better handled? Please try to explain these indexes **very shortly**.

We can define an optimism index as saying inflation is low even when it is high (FNR, already used above), and a pessimism index as saying inflation is high even when it is low (FPR). These could be used instead of or in addition to other indexes. In addition these indexes could be preprocessed (log-scale, etc) so that the boundary between the two classes becomes more evident.

(c) (2 points) Using machine learning terminology, how would you name coming up with new indexes?

This is model rebuilding (modification, update, etc.) because new (derived) data series are used and the dimension changes (increases).