DSCI 552: MACHINE LEARNING FOR DATA SCIENCE

PROBLEM SET 3

Instructors: Dr. Kristina Lerman (lerman@isi.edu) and Dr. Keith Burghardt (keithab@isi.edu)

Deadline: Thursday, March 11, 2021, 10 A.M PDT

You can submit the report on Blackboard and code on GitHub Classroom. As long as the problem set is open, you will be able to upload multiple answers (the last attempt will be graded).

TASK (20 points)

You are a new hire in a mid-size company. You have just completed your second task. You have just handed your report. Your boss took it and went to meet the executive director of the hospital. You circle around the once, expecting that you will be called, when the meeting is done. . . You were not wrong. After the lunch, you were approached by your technical manager. "Our boss wants to see us - it's urgent. Let's go!", he commanded.
Following your manager, you enter the meeting room. The CEO and the senior developer already wait inside.

Your CEO said: "You did a fabulous job! The director of clinical research was impressed. She was in fact, so impressed, that she immediately asked for help with another matter.

There is a new SARS-CoV-2 variant. It seems much more dangerous than other strains. It's a really bad news for all of us. However, there is also a hope. They identified an individuum (they call him patient Z), that seems to be immune to that new variant. They want to study, what makes him resistant to that strain. If they can understand it, they might also be able to propose an updated vaccine.

The situation is serious, and the research would go faster if we could find other people who share the same immunity as the patient Z. You will get a genetic fingerprint of patient Z and a table of genetic fingerprints of all other patients from that hospital. Your job is to identify which patient has the same type of genetic composition as patient Z.

This is a serious situation. Every passing day matters! If you act quick, you can save hundreds. You have two weeks. I want a detailed report describing your main findings, on my desk, on Thursday, March 11, at 10 am."

Hint: This task is related to unsupervised learning. You have to identify the main clusters in your data (you have to decide how many clusters you have and where they are). Next, you have to find which cluster the patient Z belong to. People from that cluster are likely to have the same covid-resistance as patient Z.
Because you do not have any test-set to self-check how good your predictions are-this time we ask you to submit your results to Kaggle (link below). On March 11, we will show you the leaderboard and I will uncover how many cases you were able to identify correctly.
Remember, we do not grade on a curve. You don't have to be better than your collegues to get 100% from that task. As long as you can identify some individuals similar to the patient Z, we will treat that task as completed.

We added some baselines, to help you gauge, if you move into a good direction. If you do everything correctly, you should be able to hit the mid-baseline by a significant margin. However, you can still succeed, even if your score is lower than baseline. What matters is the report and your code. You can still get 100% (or close to 100%) for that assignment, even if your score is low.

Your Technical Manager said: "I looked at the data. Each genetic fingerprint is represented by a vector of 512 numbers. My suggestions for you are:

- Cluster the data using the k-means algorithm (try various values of k).
- Identify the optimal number of clusters. Report that number.
- Visualize the clusters. Because the vectors have dimension 512, you must reduce the dimensionality. You can use the PCA algorithm.
- Find the cluster to which patient Z belong.
- Report, how many people are in that cluster (not counting the patient Z)."

Hint: If you don't know what to do: follow chapters 8 and 9 from "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2019)".

The Senior Developer took you aside and said: "My task will be to maintain your code. However, remember that I'm not a Data Scientist like you - so you have to be very careful when you are writing your code. Write comments and try to explain any nontrivial section."

Hint: You can use Jupyter Notebooks. Remember that you can also add special "markdown cells". You can use it to split your notebook into a few logical parts.

Data

You can find the dataset ps3_patient_zet.npy and ps3_genetic_fingerprints.npy on the Assignments section of Blackboard. The former (a NumPy array format file) contain the genetic fingerprint of patient Z. The latter file contains genetic fingerprints of all other patients.

Report

To help you, I prepared a template. See https://www.overleaf.com/read/vnvhqxkpdhbk. You are encouraged to use the template, but you are free to use other editors or make modifications. Just ensure that the final submission of the report has to be in PDF. Submission of report has to be done on Blackboard. As always submit your code on GitHub classroom and report on Blackboard. Additionally, you can also submit on Kaggle to compare results with your peers. Note that Kaggle competitions in this class as always do not carry extra credit.

Code Submission

We have created a GitHub Classroom where you can create private repositories. We will update the class on Piazza on how to go about uploading your solutions on this platform.

Grading Rules

In order to grade your work, we will role-play the following situation. We will assume, that you are a new-hire in our company. You are asked to provide a comprehensive technical report that illustrates your findings. We will evaluate it from the perspective of three people.

- Your CEO (she would like to hear high level stuff. She will probably only read the conclusions and look at main figure). (4 points for report).

- Your manager (he would like to see a detailed report; he might also look at some parts of the code). (6 points for report and 2 points for code).

- A senior developer (they would like to see the code and won't read the report at all). (8 points for code).

Your final score is: 10 points for report and 10 points for code.

Don't Panic

Don't panic. We understand that this is a large, open ended task. We also understand that this might be the very first technical report that you were asked to write. We are dedicated to help you do your best work all while keeping the standards high. We acknowledge that you have limited time and resources to complete the task. This report doesn't have to be perfect for 100% score.

If you don't know where to start read the Second Chapter of "Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow", 2nd Edition by Aurélien Géron.

If something is not clear, ask your questions on Piazza.

Note: Cite any source you use (even if you adopt/copy a snippet of code). Failure to do so would amount to plagiarism.

Optional Challenge

We also created an optional challenge for you. There is no additional credits for participating in it. However, we encourage you to give it a try. We created a special class competition on Kaggle (https://www.kaggle.com/c/usc-dsci552-section-32416d-spring-2021-ps3). You could participate in this competition and upload your results here to see where you stand. Have fun!