DSCI 552: Machine Learning for Data Science

Problem Set 3

Instructor: Dr. Marcin Abram – email: mjabram@usc.edu

Deadline: Thursday, March 4, 2020 at 10 am PDT

You can submit the solutions (pdfs) on Blackboard. The deadline to submit solutions is always the second Thursday at 10 am (PDT). As long as the problem set is open, you will be able to send multiple answers (the last submitted answer will matter).

Task (20 point)

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 office, 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 4, 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 4, I will show you the leaderboard and I will uncover how many cases you were able to identified correctly.

Remember, we do not grade on a curve. You don't have to be better then your collegues to get 100% from that task. As long as you can indentify some individuas 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 bit the mid-baseline by a significant margin. However, you can still suceed, even if your score is lower then baseline. What matter, 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 suggestion 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: If you 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 – see also post @87 on Piazza.

Learning Objective: You will be able to cluster unlabeled data. You will be able to visualize high-dimensional data. You will be able to find similar instances in highly-dimensional space.

Data

First, enter the competition by accepting the terms at `https://www.kaggle.com/t/1ec453f9af34420ca87cafe28cd29ac0`. Then, you will be able to find the dataset at `https://www.kaggle.com/c/usc-dsci552-section-32415d-spring-2021-ps3/data`. There are two files. The first file, `ps3_patient_zet.npy` (the NumPy array format) contain the genetic fingerprint of patient Z. The second file, `ps3_genetic_fingerprints.npy` contains genetic fingerprints of all other patients.

Report

To help you, I prepared a template for you, see `https://www.overleaf.com/read/qcbcsdscvcdb`. You are encouraged to use the template, but if you prefer to use another editor – or to made some modifications – this is acceptable. Just remember to always export your report to the pdf format at the end.

Report Submission

Submit your report (a pdf file) using Blackboard (check the "Assignments" section on Blackboard page of our class).

Code Submission

To accept the assignment, use the secret link `https://classroom.github.com/a/nN-jMTWP`.

Upload your code to a private GitHub repository linked to our GitHub Classroom, `https://classroom.github.com/classrooms/77759207-usc-dsci552-32415d-spring2021-classroom`.

Please, remember that your GitHub name is linked to your student profile in our GitHub Classroom. If you use Python, you can commit `.py` or `.ipynb` file(s). If you use other programming languages, like `R`, `C++`, `Java`, `Julia`, etc., prepare a description that would explain how to compile (if necessary) and run your code. If you submit more than one file (or one Jupyter Notebook), include `Readme.md` file, where you would explain what each file is about (so we know in which order we should read the files).

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 were 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 the high-level stuff; She will probably read only the abstract and the conclusions, maybe she will also look at the main figure or the main plot). *(4 points for the report)*
- Your manager (he would like to see a detailed report; he might also look at some parts of the code). *(6 points for the report and 2 points for the code)*
- A senior developer (they would like to see your code; they might not read the report at all). *(8 points for the code)*

Your final score is the sum of scores given by each person. Cumulatively, there are 10 points for the report and 10 points for the code.

Each person (the CEO, the manager, the developer) will use the following grade rubrics.

Grade Component	Meets Expectations	Approaches Expectations	Needs Improvement
Completeness			
Clarity and Support			
Validity			

By "Completeness" we mean, that all parts of the question are addressed. By "Clarity" we mean, that the text and the code is written in accessible way. By "Support", we mean that you provide sufficient evidences to back-up your statements. You must cite any source you use (even if you happened to copy or adapt a snippet of a code from the internet – you should still treat it as a citation. You should clearly mark how large is that snipped and provide adequate references).

Don't Panic

Don't panic. We understand, that this is a large, open-ended task. We also understand, that this might be your third technical report that you were tasked to write. We are dedicated to help you do the best work – and while we keep high standards for you (and for us) – in the same time, we acknowledge, that you have limited time and limited resources to complete your task. This report doesn't have to be perfect to give you 100% score.

If you don't know where to start, read the *eighth* and *ninth* chapter of "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow", 2nd Edition by Aurélien Géron. Check also the *Appendix B. Machine Learning Project Checklist* from that book.

If something is not clear, ask questions on Piazza.