WFAIS.IF-XG325.z: Explainable Artificial Intelligence

Problem Set 3

Instructor: Dr. Marcin Abram
E-mail: marcin.abram@alumni.uj.edu.pl

Deadline: Friday, April 25, 2025 at 11.59 pm CEST

As long as the problem set is open, you will be able to send multiple answers (the last submitted answer will matter).

Task (30 point)

You are a new hire in a mid-size company. You have just completed your second task. A moment ago, you handed the report. Your boss took it and went to meet the hospital's executive director. You circle around the office, expecting that you will be called when the meeting is done...

You were not wrong. Only after lunch were you 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 are already 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.

We all thought the pandemic was over. However, there is a new SARS-CoV-2 variant. It seems much more dangerous than previous strains. It's really bad news for all of us. However, there is also 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 (potential) 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 quickly, you can save hundreds. You have two weeks. I want a detailed report describing your main findings on my desk, on Thursday, April 25, at 3 pm PDT."

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 belongs to. People from that cluster are likely to have the same COVID resistance as patient Z.

Because you do not have a test set to self-check how good your predictions are, this time, I ask all of you to submit your results to Kaggle (link below; my secret test set is only there). On April 26, I will show you the leaderboard and reveal how many cases you were able to identify correctly.

Remember, I do not grade on a curve. You don't have to be better than your colleagues to get 100% from that task. As long as you can indentify some individuas similar to the patient Z, I will treat that task as completed.

I added some baselines to help you gauge if you are moving in a good direction. If you do everything correctly, you should be able to beat the mid-baseline by a significant margin. However, you can still succeed, even if your score is lower than this baseline. What matters is the report and your code. You can still get 100% (or close to 100%) for that assignment, even if your Kaggle's 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 belongs.
- 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."

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 a highly-dimensional space.

Data

First, enter the competition by accepting the terms at `https://www.kaggle.com/t/0bf0e39d4333451bb14f6fbd4eed0d9a`. Then, you will be able to find the dataset at `https://www.kaggle.com/competitions/uj-xg-325-xai-spring-2025-ps-3/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 Submission

Your submission should consist of one single PDF file. You should include three separate, self-contained sections:

1. Short, under one page, executive summary for the CEO.
2. A technical discussion (with plots) for the technical manager.
3. A self-contained print-out of the code for the senior developer.

Please do not print out all the codes you have produced. Instead, try to show one final, self-contained, end-to-end example.

Grading Rules

In order to grade your work, we will role-play the following situation. I will assume that you are a new employee. You were asked to provide a comprehensive technical report that illustrates your findings. We will evaluate each subsequent part from the perspective of the following people.

- Your CEO (she would like to hear the high-level stuff; She will only read the executive summary). *(10 points)*
- Your manager (he would like to see a detailed discussion; he is also very keen to see some plots). *(10 points)*
- A senior developer (they would like to see the code; they might not read the rest of the report at all). *(10 points)*

Your final score is the sum of the scores given by each person. Cumulatively, there are 30 points. 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" I mean that all parts of the question are addressed. By "Clarity" I mean that the text and the code are written in an accessible way. By "Support", I mean that you provide sufficient evidence to back up your statements. You must cite any source you use (even if you happened to copy or adapt a snippet of code from the internet – you should still treat it as a citation. You should clearly mark how large that snippet is and provide adequate references). In this class, taking code generated by various LLM-based systems is treated the same as copying the code from the internet. It is allowed as long as you are transparent about that (and you cite the source, i.e., you disclose the provider, the name, and the version of the model, e.g., OpenAI's ChatGPT o1-mini or Anthropic's Claude 3 Haiku, etc.).

Don't Panic

Don't panic. I understand that this is an enormous, open-ended task. I also know this might be only the third technical report you were tasked to write. I am dedicated to helping you do the best work, and while I keep high standards for you, at the same time, I acknowledge that you have limited time and limited resources to complete your task. This report doesn't have to be perfect to be graded at 100%.

If you don't know where to start, read the *third* 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. And as always, if something is not clear, ask questions using the Pegaz's forum.

Learning Objective: You will be able to design a machine learning pipeline. You will be able to create a technical report. You will be able to communicate your findings with several archetypal people.