

Exercise 4 – The Virus Challenge – Clustering

Wet (80%)

Background

In this exercise you will tackle, yet again, the virus classification task. However, with more ML experience and the following tips you can establish better models:

- The features you should focus for virus/risk/spreading labels are:
[DisciplineScore](#), [TimeOnSocialActivities](#), [AgeGroup](#), [StepsPerYear](#), [pcrResult4](#), [pcrResult1](#), [pcrResult12](#), [pcrResult5](#), [pcrResult16](#), [pcrResult14](#), [SyndromeClass](#)
- One model does not rule them all.
 - Use separate models for predicting each of the three labels' components, virus, risk, and spreaders.
- More data is usually better
 - In this exercise you will get much more samples

Assignment

You should submit a process that starts from loading and preparing the data, and up to the completion of the tasks detailed below. This process should include the following:

1. Load the data from [virus_hw5.csv](#) file.
2. Prepare the data.
 - a. Make sure you focus on the selected features as mentioned above.
3. Train and evaluate at least three models
 - a. Use at least one ensemble-based model
4. Select the most accurate model automatically and use it to predict the classes for the new unlabeled data from the "[virus_hw5_unlabeled.csv](#)" file.
5. Output your results to another csv file named "[predicted.csv](#)".
 - a. This file should contain 4 columns only ("PatientID", "Virus", "Risk", "Spreader").
 - b. Use the "[predicted.csv](#)" file from the HW3 section in the webcourse as a reference.

Please submit

1. The Python script file that implements the above
2. A documentation that
 - a. Contains the answers of the dry part below.
 - b. Explains how you address the wet tasks.
 - i. Includes any significant decision you took.
3. A file named "[predicted.csv](#)" as described in (5)

Notes:

- The submission with the most accurate models wins a 15% bonus and an honorable mention in the webcourse news section.
- The other top 10% submissions will receive a 7% bonus and the course staff deep appreciation.

Dry

1. Prove that when running AdaBoost, the distribution is updated such that the error of the chosen weak classifier h_t , w.r.t the updated distribution $D_i^{(t+1)}$, is exactly $\frac{1}{2}$.

That is, prove that $\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i} = \frac{1}{2}$.

Hint: You can fill the missing steps in the following derivation:

$$\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i} = \dots = \frac{\epsilon_t}{\epsilon_t + (1 - \epsilon_t) \exp\{-2w_t\}} = \dots = \frac{1}{2}.$$

2. Consider a fully connected neural network with L linear layers.

Denote the output of the layer by $F_\Theta: \mathbb{R}^d \rightarrow \mathbb{R}^K$, where $\Theta = (\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)})$ is the set of all weights and **no** biases.

As an activation function, we use the ReLU function $\sigma(z) = \max\{0, z\}$.

That is, the network's output is given by:

$$F_\Theta(x) = \mathbf{W}^{(L)\top} h^{(L-1)}(x)$$

where we recursively define the hidden layers:

$$h^{(1)}(x) = \sigma(\mathbf{W}^{(1)\top} x), \quad h^{(\ell)}(x) = \sigma(\mathbf{W}^{(\ell)\top} h^{(\ell-1)}(x))$$

We now scale all weights in Θ by a factor of $\alpha \in \mathbb{R}_{>0}$.

Notice: the ReLU function is positive-homogeneous in the sense that $\sigma(\alpha \cdot z) = \alpha \cdot \sigma(z)$.

- 2.1. Show that the new output holds $F_{\alpha \cdot \Theta}(x) = c \cdot F_\Theta(x)$, and find the appropriate scalar c .

As seen in the lecture, if we want to use the output to classify samples into K classes, we use the

output $F_\Theta(x)$ to induce a discrete *softmax* distribution c_{y_1}, \dots, c_{y_K} , where $c_{y_i} = \frac{\exp\{(F_\Theta(x))_i\}}{\sum_{j=1}^K \exp\{(F_\Theta(x))_j\}}$.

- 2.2. For $\alpha \rightarrow 0$, find the distribution the *softmax* distribution induced by $F_{\alpha \cdot \Theta}(x)$ converges to.

- 2.3. For $\alpha \rightarrow \infty$, find the distribution the *softmax* distribution induced by $F_{\alpha \cdot \Theta}(x)$ converges to.