Instructions on
Project #21.2 "The Next Generation"
and Deliverables to ν-Corp*

November 28, 2025

1 Introduction

In a further effort to mitigate the potential existential risks posed by AGI, also this year, the Government has allocated an undisclosed amount of money to various initiatives, ranging from exploratory research to top-secret interventions. It is important to note that the Government's official stance is that no such risk currently exists.

Project #21 is the one remaining ν-Corp project that also receives funding this year as part of the above-mentioned governmental program. Project #21.2 and various internally financed programs focus on monitoring, deciphering, and utilizing signals intercepted and logs from different strata of the Internet whose threat level cannot be assessed individually. Project #21, its follow-up #21.2, and your assignment specifically, concern a basic, yet essential step in the decryption process of one type of interferential signal.

To maximize the use of human expertise, while minimizing the risk of information leakage (e.g., due to sharing information that has not been authorized to be shared), we devised an overall hierarchical solution strategy akin to divide-and-conquer. Therefore, your sole task is to solve the assigned problem as well as possible and pass the team's findings on to your managing director.

---

*ν-Corp is a subsidiary of E-Corp and established to carry out government-supported research into artificial intelligence and its exploitation for the greater good.

The relationship between your task and other tasks, both classification and otherwise, will not be disclosed. However, rest assured, ν-Corp is also committed to your interests in its selfless support of the advancement of humankind.

The next section provides further details about your assignment. It discloses all we can about the data provided to your team and defines the task at hand. Section 3 describes the deliverables we expect from your team and how its performance will be evaluated. Among others, we expect an executive report of about four pages on your work. The appendix contains a basic LaTeX setup that can be used.

2   Data and Assignment

Your team's unique data comprises a labeled training set and an unlabeled test set. The former contains 4,000 examples in two classes with 2,000 objects in every class. The unlabeled dataset has 50,000 objects and we know that the true prior probabilities in this set are 0.4 for the first class and 0.6 for the second. Every object is described by 72 numerical attributes. The attribute values are recorded using our newly acquired ▆▆▆▆▆▆▆▆ probes from DARPA and calibrated using our in-house ▆▆▆▆▆▆▆▆▆▆▆▆ technology.

For your convenience, one CSV file for the training data and one for the test data have been provided (these can be requested from your managing director prof.dr. Marco Loog by sending him a friendly mail with your group number). In both the training and the test set, every row contains 64 input attributes for one object. In the training set, the first 2,011 rows belong to the first class and the second 1,996 to the second class.

Though it depends on the exact dataset your team is dealing with, it seems safe to say that an AUC of about 0.80 should be easily achieved, at least on the balanced data. The challenge is, however, to perform well in terms of AUC on the imbalanced test set. In short, what primarily matters to this classification challenge is the area under the ROC curve achieved on the unlabeled test dataset *and* your team's estimate of this performance measure.

3 Deliverables

Your team should send in the following three items.

1. A CSV file with 50,000 numbers based on which all test samples can be ordered from more likely to belong to class 1 to more likely to belong to class 2, with smaller values meaning more likely to belong to class 1. As an example, if test sample 15,001 is more likely to belong to class 1 than test sample 3, then row 3 should get a higher numeric value than row 15,001.

   The file your team submits *should only contain these 50,000 numerical values*. Moreover, these numeric values for ordering the data must be provided in the same order as the test data points are provided.

2. A report (pdf; 2050 words max) that describes concisely and clearly a) any results or general findings of interest [600 words max], b) the final classification approach used to classify the test data [600 words max], and c) the way the performance on this test set was estimated and, importantly, what the actual estimate is [600 words max]. The report also leaves some space [250 words max] for important additional issues that, the team feels, need to be mentioned.

   The report should be uploaded as a pdf. A basic LaTeX setup that can be used, is provided in the appendix. This also contains suggested section titles and another brief description of its expected content.

3. The code corresponding to points b) and c) above and to Sections 2 and 3 from your report. It should, in essence, enable ν-Corp to understand the details of your approach and reproduce your results.

   Note: as the datasets contain potentially sensitive and harmful information, you should remove these from all your devices upon completing this assignment.

A   Report

```
\documentclass[12pt,a4paper]{article}
\usepackage[english]{babel}
\usepackage{amsmath}
\usepackage{graphicx}
\usepackage{soul}

\author{Names of team members; s-numbers; project-group number}
\title{TBD}

\begin{document}
\maketitle
% The aim of this document should be to convey a clear and
% concise message, not to use all 2050 words allowed.

\section{Preliminaries (600 words max)}
% Describes initial research, preliminary results, and any
% finding of interest, which led the team to its final choice
% of data pre-/postprocessing and classifier.

\section{Classification Approach (600 words max)}
% Describes the final classification approach, including
% any pre-/postprocessing of the data.  Clearly describe
% how the classifier was trained and how we come from
% an attribute vector to the final indicator of the
% likelihood of the sample belongs to class 1 or 2.

\section{Performance Estimation (600 words max)}
% Describes the way the expected performance was estimated
% and, importantly, clearly identify this estimate!  E.g.
The estimated test performance is \hl{your estimate}

\section{Other Matters (250 words max)}
% Mentions any important issue that needs mentioning.

\end{document}
```