Computer Science Department
Universidad de Ingeniería y Tecnología

Project 1: Classification
Prof. Ariana Villegas

1 Theoretical Part. (30 pts)

1.1 Linear Regression. (15 pts)

We would like to fit a linear regression model to the dataset

$$D = \left\{ \left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(N)}, y^{(N)}\right) \right\}$$

with $x^{(i)} \in \mathbb{R}^M$ by minimizing the ordinary least square (OLS) objective function:

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=1}^{M} w_j x_j^{(i)} \right)^2 .$$

1. We solve for each coefficient w_k $(1 \leq k \leq M)$ by deriving an expression of w_k from the critical point $\frac{\partial J(w)}{\partial w_k} = 0$. What is the expression for each w_k in terms of the dataset $(x^{(1)}, y^{(1)}), \cdots, (x^{(N)}, y^{(N)})$ and $w_1, \cdots, w_{k-1}, w_{k+1}, \cdots, w_M$?
2. How many coefficients (w_k) do you need to estimate? When solving for these coefficients, how many equations do you have? (Give your answers in terms of M or N)

1.2 Logistic Regression. (15 pts)

Consider a dataset D where each data point is represented by a single feature value. Suppose D is separable along this one-dimensional feature, that is, there exists a threshold t such that all points with feature values less than t belong to one class, while all points with feature values greater than or equal to t belong to the other class.

1. If you train an unregularized logistic regression model for infinite iterations on training data that is separable in at least one dimension, the corresponding weight(s) can go to infinity in magnitude. What is an explanation for this phenomenon?
2. How does regularization (such as ℓ_1 and ℓ_2) help correct the problem in the previous question?

2 Applied Part. (70 pts)

1. For this project, you have two options. Review the details of these Kaggle's competitions and pick one:
   (a) Human Activity Recognition Using Smartphones
   (b) EEG of genetic predisposition to alcoholism
2. Analyze the features of the dataset pertinent to the problem you selected in the previous step. Review the following libraries to extract features of time series. Justify your decisions.
   (a) PyTS
   (b) TsFresh
3. Implement two classification methods (e.g. logistic regression, SVM, KNN, decision trees and so on). Justify your decisions.
4. Report the classification metrics: F-Score, Accuracy, Confusion Matrix. Analyze and discuss the results of each method.

Report: Only one member from each team should upload the report. The document must be created in LaTeX and follow this template: Download.

The document structure should follow this outline:

1. Each team member's names must include their respective percentage of participation.
2. Introduction: Project description.
3. Dataset: Exploration and analysis of the dataset.
4. Methodology: Explanation of the model, loss functions, and regularization techniques.
5. Implementation: Include the link to Colab or GitHub where the implementation can be found, avoiding direct code placement in the report. Define a seed to replicate the results. [Optional] Relevant implementation details can also be included (error handling, parallelization, etc.).
6. Experimentation: Present results with graphs and/or tables, avoiding terminal screenshots.
7. Discussion: Interpretation of the obtained results and their relationship with the learned theory.
8. Conclusions: Summary of results, limitations, and recommendations.
9. Contribution Statement: Summary of each team member contribution.

** Avoid using screenshots to display results such as accuracy, F1 score, loss, or error. Instead, ensure all results are properly formatted and presented within the document.*

** The document should be **a maximum of 8 pages** and can include any number of appendices deemed appropriate.*

Library Usage: For preprocessing/methods/metrics other than the required implementation in activity 3, you are free to utilize libraries.

2.1 Rubric

Criteria	**Excellent**	**Good**	**Fair**	**Poor**
Model Design & Justification (15 pts)	Well-suited model with strong justification.	Adequate model with reasonable justification.	Basic design with weak justification.	Inappropriate or unclear model design.
Code & Documentation (15 pts)	Organized, functional code with clear documentation.	Functional code with adequate documentation.	Disorganized code with minimal comments.	Incomplete or non-functional code with poor documentation.
Methodology & Experimentation (15 pts)	Thorough experiments with a solid approach.	Good experiments with some depth.	Basic experiments with limited exploration.	Weak or incomplete methodology.
Results Discussion (15 pts)	Insightful analysis with clear connections.	Good analysis with some insights.	Basic discussion with limited insights.	Minimal or unclear analysis.
Conclusions (5 pts)	Clear, well-supported conclusions.	Reasonable conclusions, somewhat supported.	Vague or general conclusions.	No or unsupported conclusions.
Contribution Statement (5 pts)	Detailed, fair distribution of work.	Clear, but lacks some detail.	Vague or unclear roles.	No statement provided.

Table 1: Rubric for Project