First and last name _____ Student ID _____

FOUNDATIONS OF MACHINE LEARNING

Prof. Tommaso Di Noia

Academic Year 2024/2025

04/12/2024

Exercise

This dataset encompasses details about various workers and their corresponding employment levels, featuring a diverse set of attributes ranging from categorical to continuous. Initialize the data loading process using the suitable Pandas function, and meticulously inspect for any instances of null or duplicated data. Specifically focusing on the "**salary_in_usd**" feature, identify and eliminate outliers while devising a strategy to address any missing values.

Then, use any method you like to encode the categorical features, namely "**work_year**", "**experience_level**", "**employment_type**", "**job_title**", "**employee_residence**", "**remote_ratio**", "**company_location**", and "**company_size**". You may consider to employ the sklearn LabelEncoder class[1].

Following the preprocessing steps, normalize the dataset utilizing the z-score technique to ensure consistent scaling across features. Subsequently, construct a neural network using **PyTorch**, incorporating **2 hidden layers with 5 and 3 neurons**, respectively. Carefully select an appropriate learning rate and normalization value for optimal model training.

Furthermore, assess the model's performance using a relevant evaluation metric, ensuring a comprehensive understanding of its effectiveness in handling the given employment dataset. Finally, find the best hyperparameter combination (namely **lr** and **weight_decay**) using both the **Grid Search** and the **k-fold cross validation** methods.

[1] `https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.LabelEncoder.html`