Course name: Introduction to Machine Learning

Course code: 822047-B-6

Take-home Assignment 1
Academic Year 2022-2023

Sharon Ong and Görkem Saygili
Department of Cognitive Science and Artificial Intelligence
Tilburg University

The Individual Take-home Assignment 1 is worth 100 points and is 20% of your grade. A pass is not required for this component to pass the course as long as the total grade for all components is 5.5 or higher. Because a pass is not required, no resit will be provided for the whole class. Students who fail the assignment may resubmit for a maximum grade of 6.0. Re-submissions must be received before the final exam.

Late submission policy: There is a 5% penalty per day for late submissions. Zero credit is earned if submitted after 7 days past the due date. Start early, most assignments will take longer than you expect.

You are provided with your own dataset in (XXXXX_*pokemon*.csv) where XXXXX is your SSID. This file can be found in under "Files/assignment/individual_files/" on Canvas. These datasets are adapted from a real-world dataset; The Complete Pokemon Dataset obtained from Kaggle [1].

You will create programs to perform regression and submit your code in a Jupyter notebook (*.ipynb). At the top of your file, please reference any code, methods, ideas that are not your own or not provided in this course. Remember that these codes must be publically available and free to use. You do not have to reference the course notebook, the textbook nor the lecture slides. Please also include any special instructions that are required to run your code.

Students are NOT allowed to use any specific (Generative AI) tool to generate code. Using the output of Generative AI tools will typically be deemed plagiarism or accidental plagiarism. If there is any doubt about plagiarism, the case will be referred to the exam committee.

This assignment is due on **22, March 2023**.

The assignment consist of two parts. The completion of all the requirements from Part A and Part B can earn you a maximum of 90 points. In addition, a maximum of 5 points will be awarded for documentation and good coding standards. A maximum of 5 points will also be award for error-free code execution. The grading rubrics can be found on Canvas.

A Regression to predict a pokemon's weight from its height.

A.1 Exploratory Data Analysis Tasks

1. (5 points) Load the dataset. Assign the column "*height_m* " as your feature data and the column *weight_kg* as your target variable (Set X to the contents in the column "*height_m* " and y to the contents in the column "*weight_kg*"). Present a scatterplot showing the relationship between these two variables. Label the x axis as **height (m)** and y axis as **weight (kg)**. Add a title to the plot.

A.2 Regression Tasks

1. (5 points) Split your dataset into 80% training, 10% validation and 10% testing.

2. (5 points) Fit a linear regression model to the data. Evaluate on the test data, how successful the fit is by computing the R-squared score.

3. (10 points) Using a for loop, train polynomial regression models for the degrees of freedom hyperparameter of 2, 3, 4 and 5. For each degree of freedom, fit on the training data and evaluate on the validation set. Evaluate the best performing hyperparameter from the validation set on the test data.

4. (10 points) Using a for loop, train KNN regression models for the following number of nearest neighbours; 1, 3, 6 and 10.(The number of nearest neighbours is your hyperparameter for KNN regression). For each k-nearest neighbours, fit on the training data and evaluate on the validation set. Evaluate the best performing hyperparameter from the validation set on the test data.

5. (5 points) In the same plot, display a scatterplot of your training and test samples. Display with continuous lines, the linear regression, the polynomial and KNN regression solutions with the best hyperparameters. List the best performing regression model.

B Regression to Predict a Pokemon's hit points from a set of features.

B.1 Exploratory Data Analysis Tasks

1. (5 points) Create a feature set without the *name* and *hp* columns. (Assign all the data except the columns *name* and *hp* to X). Assign the *hp* as your target variable.

2. (5 points) Compute the correlation matrix which measures the relationship between all the numerical features and target variable. Display the correlation matrix in a heatmap. Find the feature with the highest correlation with the target variable.

3. (5 points) There is one categorical feature in your dataset, the column - *type*, which shows the pokemon type (e.g. grass, flying). Display the number of pokemons for each type in a histogram.

4. (5 points) Perform the necessary steps to convert the feature *type* to a numerical format for regression tasks.

B.2 Regression Tasks

1. (5 points) Split your dataset into 80% training, 10% validation and 10% testing.

2. (5 points) Fit a linear regression model and a dummy regressor to the data. Evaluate on the test data, how successful the fit is by computing the R-squared score.

3. (15 points) Compare the linear regression analysis with a Ridge regression and Lasso regression with the following six alpha values; 0.001, 0.01, 0.1, 1, 10, 100. Note: The alpha values is your hyperparameters. For each alpha value, fit on the training data and evaluate on the validation set. Evaluate the best performing hyperparameter from the validation set on the test data.

4. (5 points) Create a pandas dataframe to display your results. The pandas dataframe should have three columns. The first column should the name of the regression model, the second is the alpha value and the third is the respective R-squared score. Save that dataframe to a csv file. **Points breakdown**: 1 point for creating the dataframe. 1 point for saving dataframe to csv. 3 points for automatic input the alpha and R-squared scores (you do not manually type these values into the dataframe).

References

[1] Banik, R. (2019). The Complete Pokemon Dataset, https://www.kaggle.com/datasets/rounakbanik/pokemon.