Departamento de Ciencias de la Computación
Universidad de Ingeniería y Tecnología

Project #1: Regression

Prof. Cristian López Del Alamo

---

1. Theory (30pts)

Exercises: Identifying Linear and Nonlinear Functions (10pts)

For each of the following functions $h(x)$, determine whether it is linear or nonlinear.

1. $h(x) = 2x_1 + 3x_2 + 5$
2. $h(x) = x_1 + \log(x_2) + 3$
3. $h(x) = 5x_1 + 2x_2 + x_3$
4. $h(x) = 6x_1x_2 + 4x_3$
5. $h(x) = 2x_1 + 3x_2 + 5\sin(x_3)$

Loss Function for Nonlinear Regression (20pts)

Consider a nonlinear hypothesis $h(x)$:

$$h(x) = g(x_1, x_2, \ldots, x_n; \theta)$$

The loss function $J(\theta)$ is defined as:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\left| h(x^{(i)}) - y^{(i)} \right|^3 + \lambda \sum_{j=0}^{n} \theta_j^2 \right)$$

Where:

- m is the number of training examples.
- $h(x^{(i)}) = g(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}; \theta)$ is the prediction of the model for the i-th example.
- $y^{(i)}$ is the actual value for the i-th example.
- λ is a hyperparameter controlling the regularization.
- θ_j are the parameters of the model.

Work: Find the partial derivatives

2. Practice (70pts)

Project 1

Objective: The goal of this project is to predict the burned area of forest fires in the northeast region of Portugal. To achieve this, you will be provided with a dataset containing 517 instances and 12 features. The 13th feature is the area that you need to predict.

Activities

1. Load the files from the Database. Use seaborn for data visualization. This for Exploratory Data Analysis (EDA) step.
2. Implement multivariate linear or nonlinear regression algorithms.
3. Use Ridge or Lasso regularization techniques.
4. Split the dataset into 70 % for training and 30 % for testing.
5. Apply 10-fold cross-validation methods and, for each fold, calculate the adjusted R-squared. Generate a table showing the adjusted r^2 for each fold and, in the end, provide the average, mean, and standard deviation of the adjusted r^2.
6. Select the most relevant features. You can use Lasso to identify which features have zero coefficients.
7. Display the most relevant features of the model. Use this template for your document: ieee templete

NOTE: Each team must submit a single document with the following structure:

1. Introduction.
2. Explanation of the *regression* methods as well as the considerations taken into account for generating the feature vectors.
3. Experiments: This is the most important part of the project and should be carried out thoroughly. Use graphs and tables to present your findings.
4. Conclusions: Write conclusions based on the results.

Finally, include the link to the GitHub or Colab to verify the source code. It is important that your Colab is accessible; otherwise, the work will be dismissed.

Important:

1. Introduction.
2. Any partial or total copy of the work, or any content taken from ChatGPT or a similar model, will result in a zero grade with no right to appeal.
3. Add the percentage of participation for each team member in the document.
4. The document should not exceed 8 pages.
5. The document should not include pasted code or tables. Use Seaborn to create the images, and use LaTeX to create your own tables.
6. Remember to properly cite any image that is not your own.