جامعة بيرزيت

BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #2

Submission deadline: 25.11.2025

Part 1 – Non-Linear Regression:

A) Generate a set S $= \{(\vec{x}_i, y_i): i = 1, \dots, 25\}$ of 25 data points, where:

- $\vec{x}_i = (1, x_i, x_i^2, \dots, x_i^9)$ for some $x_i \in [0, 1]$ and
- $y_i = \sin(5\pi x_i) + \varepsilon_i$, where $\varepsilon_i \in [-0.3, 0.3]$ is some random noise.

- Apply ridge regression to S using 5 different values of λ, including λ = 0.
- Plot the obtained curves. Which value of λ produces the model with the best generalization performance? Discuss your findings.

B) Using the generated set S in part A, apply non-linear regression (without regularization) using different numbers of RBF basis functions:

- 1 RBF basis function
- 5 RBF basis functions
- 10 RBF basis functions
- 50 RBF basis functions

- For each case, ensure that the centers of the RBFs are evenly spaced across the interval $[0,1]$. Choose a reasonable width for the basis functions.
- Plot the resulting curves for each case, including the true target function $\sin(5\pi x)$ for reference. Which configuration (number of RBFs) results in the model with the best generalization? Discuss your findings.

Part 2 – Logistic Regression:

Use the customer churn dataset from Assignment #1 to apply logistec regression to predict the churn status. Perform the following tasks:

- Apply any necessary data preprocessing steps (this can be taken from Assignment#1, or follow typical preprocessing such as standardization).

- Randomly split the dataset into: 2500 training samples, 500 validation samples, and 500 test samples.

- Train and evaluate the following logistic regression models:
  - Logistic Regression with a linear decssion boundary
  - Logistic Regression with a non-linear decssion boundary using polynomial features of the following degrees: {2, 5, 9}. (hint: you can use `PolynomialFeatures` from `scikit-learn` library)

- For all models, evaluate and report the following metrics on the training, validation, and test sets: Accuracy, Precision, and Recall. Discuss how the model complexity (the degree of the polynomial) affects overfitting and generalization.

- Select the best model based on the performance on the validation set. plot the ROC curve for the selected model and compute the area under the curve (AUC).

Deliverables:

1. **Code**: Submit a Jupyter Notebook (or Python script) containing the code for all parts of the assignment.

2. **Presentation**: Submit a PowerPoint presentation summarizing your findings. The presentation should be supported by **plots and metrics** and be **concise** and suitable for a **7-10 minute presentation**. You will be asked to present and discuss your slides for evaluation.