Shahid Beheshti University
Machine Learning
M.Sc - Fall 2024

Assignment 2

1 Theoretical Questions

1. **Exercise 1:** Can gradient descent get stuck in a local minimum when training a linear regression model? Why?

2. **Exercise 2:** Suppose you are using polynomial regression. You plot the learning curves and you notice that there is a large gap between the training error and the validation error. What is happening? What are three ways to solve this?

3. **Exercise 3:** In a dataset where the number of predictors exceeds the number of observations, what problems arise when applying linear regression, and how might you resolve them?

4. **Exercise 4:** How do outliers in the response variable (target variable) affect the model in least squares regression? Which alternative techniques could help make the model more robust to outliers?

5. **Exercise 5:** Implement Linear Regression with Mean Absolute Error as the cost function from scratch. Compare your results with the Linear Regression module of Scikit-Learn on **Avacado Price** dataset to predict the price.

6. **Exercise 6:**

   How can we use statistical significance tests to determine if one model consistently performs better than another, rather than any differences being due to random chance? Could you give examples of these tests and explain when each one is appropriate for comparing models on metrics like accuracy, F1 score, or error rate?

2 Practical Exercise

In this part, you are going to work with the **News Popularity Prediction**dataset. You will implement a regression model using the **Scikit-Learn** package to predict the popularity of new articles (the number of times they will be shared online) based on about 60 features. You are expected:

- Perform exploratory data analysis on the dataset.
- Propose 5 different hypothesis tests related to the dataset. At least use 3 different tests.
- Try to predict the popularity using linear regression
- Measure your models performance using different meters
- Try Ridge and Lasso regression **(Extra Point)**
- Use various scaling methods and report their effects.
- Add polynomial features and report their effect.
- Try using **GridSearchCV** with **RandomizedSearchCV** to tune your model's hyperparameters. **(Extra Point)**
- Get familiar with and implement the following loss functions from scratch and utilize them with a Linear Regression model and discuss their effect on the performance of the model. **(Extra Point)**
  - Absolute Error
  - Epsilon-sensitive error
  - Huber