CS6364/CS4364

Assignment 1

Question 1 (70pts)

Machine Learning from Scratch: Kaggle Most Streamed Spotify Songs 2023

Objective

The objective of this homework is to implement a machine learning algorithm from scratch using the ”Most Streamed Spotify Songs 2023”. You ***are required to use first principle functions*** and are allowed to use tools such as numpy and pandas for data manipulation, ***but not any machine learning packages such as sklearn, pytorch, tensorflow, etc.***

Dataset

Link: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023

The dataset contains several features about songs in the Spotify library. Your task is to predict a target variable using the features provided in the dataset. The target variable in this dataset could be the "streams" of the song, which represents the total number of streams on Spotify.

Tasks

1. Data Preprocessing

You are required to preprocess the data using first principle functions. This includes handling missing values, outliers, and scaling of the data. You can use numpy and pandas for these tasks.

2. Implement a Machine Learning Algorithm

Choose a machine learning algorithm and implement it from scratch. This could be a simple linear regression model, a decision tree, or even a neural network if you're up for the challenge! You should clearly define the cost function and the optimization algorithm used for learning the parameters of your model.

3. Model Evaluation

Split your dataset into training and testing sets. Train your model on the training set and evaluate its performance on the testing set. Implement evaluation metrics from scratch to assess the performance of your model.

4. Report

Write a report detailing your approach, challenges faced, and learnings from this exercise. The report should also include your results and any visualizations that support your findings.

Deliverables

1. Python code files for data preprocessing, model implementation, and evaluation.
2. A report detailing your approach, challenges faced, learnings, and results.

Evaluation Criteria

Your homework will be evaluated on the following criteria:

1. Correctness of the preprocessing steps.
2. Correctness of the implemented machine learning algorithm.
3. Performance of the model on the testing set.
4. Quality of the report - clarity of writing, presentation of results, and discussion.
5. The report **MUST** contain an analysis of the performance including bias-variance trade-off, how you selected your trained model and how you concluded that your selected model is the best in terms of generalization, things such as overfitting and underfitting.

Remember, the goal of this homework is not just to build a model that works well, but also to demonstrate your understanding of how these algorithms work under the hood. Your report should be 2-5 pages in length.

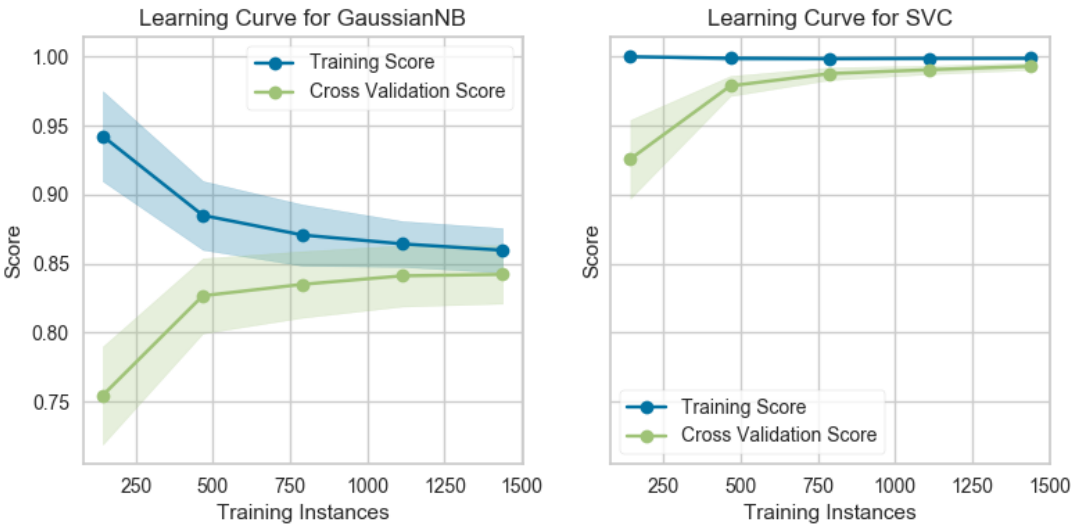
CS6364/CS4364

Assignment 1

Question 2 (30pts)

Bias-Variance Tradeoff

Consider the following scenario: You have trained two machine learning models to perform a classification task. The performance of each model has been evaluated on both the training dataset and an unseen test dataset (Cross Validation Score). The following graph represents the model's performance as the dataset size increases:



Based on the provided graph, please answer the following questions:

A. At which dataset size (approximately) does each model seem to achieve the optimal balance between bias and variance? Please justify your answer.
B. In which regime (high bias, high variance, or optimal) are each model operating at the following dataset sizes:
    a. Small dataset size (e.g., 250 data points)
    b. Large dataset size (e.g., 1000+ data points)

C. How would you modify the model's complexity to improve its performance, if it is operating in the high bias regime? Conversely, what would you do if it is operating in the high variance regime?

D. Do you expect adding more data to improve the performance for each model? Elaborate on your response.

E. Plot a similar plot for a hypothetical binary classification such as above where the model underfits. Draw the curves for both training and validation scores as a function of the Training Instances size.