CSc 447000, Fall 2022

Assignment 3

Due October 12, 2022 at midnight

Like the previous assignment, this assignment involves the Iris Dataset.

As we discussed in class, the dataset includes 150 samples, 50 from each of three species:

- α) Setosa
- β) Versicolor
- γ) Virginica

For each of the samples, the dataset includes four features:

- A) sepal length
- B sepal width
- C) petal length
- D) petal width

In this assignment, you'll be looking at the six possible combinations of two features at a time

1. AB
2. AC
3. AD
4. BC
5. BD
6. CD

and at the entire dataset (which includes 50 samples from each of the three species).

For each of the six possible combinations of two features at a time, apply the following five ML models:

1) Stochastic gradient descent using the modified-Huber loss function
2) Logistic regression
3) Support vector machine with the following kernels
    a. Linear
    b. Polynomial
    c. RBF

Present your findings in the clearest way that you can, comparing the accuracy of 30 feature-model combinations. Including relevant plots is encouraged but focus on the plots that provide the most insight into what's going on.

Discuss your results. Include in your discussion how varying the combination of features and model affects the accuracy of the results.

Please submit you're your work on Blackboard as an ipynb file. Please name your file as

LastName_FirstName_AS03.ipynb

You must discuss your answers and describe how you came up with them. Show your work. Just stating a correct answer won't get you more than half credit.

If you collaborate on this or any other assignment, you must have contributed substantially to anything you submit; just using a current (or past) classmate's work without having contributed substantially to it is not collaboration -- it's cheating.

If you collaborate with anyone you must indicate with whom you collaborated.