Assignment No. 4
EECS 658
Introduction to Machine Learning
Due: 11:59 PM, Thursday, October 14, 2021
Submit deliverables in a single zip file to BlackBoard
Name of the zip file: FirstnameLastname_Assignment4 (with your first and last name)
Name of the Assignment folder within the zip file: FirstnameLastname_Assignment4

Deliverables:

1. Copy of Rubric4.docx with your name and ID filled out (do not submit a PDF)
2. Python source code for CompareFeatureSelectionMethods
3. Screen print showing the successful execution of CompareFeatureSelectionMethods. (Copy and paste the output from the Python console screen to a Word document and PDF it).
4. For Part 2, using the PoV formula and the values from the eigenvalue matrix, show that the program calculated the PoV correctly. (see "Deliverable 4 (PoV) Example" on BlackBoard).
5. Answers to the following questions for CompareFeatureSelectionMethods:
    a. Based on accuracy which dimensionality reduction method, PCA, simulate annealing, or the genetic algorithm worked the best?
    b. For each of the two other methods, explain why you think it did not perform as well as the best one.
    c. Did the best dimensionality reduction method produce a better accuracy than using none (i.e. the results of Part 1)? Explain possible reasons why it did or did not.
    d. Did Part 2 produce the same set of best features as Part 3? Explain possible reasons why it did or did not.
    e. Did Part 2 produce the same set of best features as Part 4? Explain possible reasons why it did or did not.
    f. Did Part 3 produce the same set of best features as Part 4? Explain possible reasons why it did or did not.

Assignment:

- In this assignment, you will use 2-fold cross-validation of the iris data set using the Support Vector Machine (SVM) machine learning model.
- This assignment has four parts.
- In each part (except the first one) you will use different dimensionality reduction methods on the iris data set.
- For each of the parts, the Python program should display (with a label showing the Part number):
    - Confusion matrix
    - Accuracy metric
    - List of features used to obtain the final confusion matrix and accuracy metric.
- Name the program CompareFeatureSelectionMethods
- Part 1:

    - Use the original 4 features: sepal-length, sepal-width, petal-length, and petal-width.
- Part 2:
    - Refer to “PCA Feature Transformation” slides 11-14 and “Python Example” slides 24-30 in the 9/30 lecture
    - Perform PCA on the iris data set as shown in slides 24-30. Do not use the scikit PCA library. You can use the library to check your code.
    - Use PCA to transform the original 4 features (i.e., sepal-length, sepal-width, petal-length, petal-width) into 4 new features (z_1, z_2, z_3, and z_4).
    - To determine the transformed features (see slide 14):
        - $\mathbf{Z} = \mathbf{X}\mathbf{W}^\mathbf{T}$
        - where
        - **X** = X_centered matrix of original features (i.e., 4-by-150 array/matrix of original iris dataset features)
        - **W** = eigenvectors matrix (4-by-4 matrix)
        - **Z** = transformed X_centered matrix of original features (i.e., 4-by-150 array/matrix of transformed iris dataset features, z_1, z_2, z_3, and z_4)
    - Display the eigenvalues and eigenvectors matrices.
    - Select a subset of the transformed features, so that PoV > 0.90.
    - Display the PoV
    - Use the selected subset of transformed features to calculate the confusion matrix and accuracy metric.
- Part 3:
    - Use simulated annealing to select the best set of features from the 4 original features (i.e., sepal-length, sepal-width, petal-length, petal-width) plus the 4 transformed features (z_1, z_2, z_3, and z_4) from Part 2 (for a total of 8 features).
    - Set the iterations = 100
    - Perturb with randomly selected 1 or 2 parameters (because 1-5% of 8 is < 1)
    - c in Pr[accept] = 1
    - Use restart value (x) of 10
    - Print out for each iteration:
        - Subset of features
        - Accuracy
        - Pr[accept]
        - Random Uniform
        - Status: Improved, Accepted, Discarded, or Restart
- Part 4:
    - Use the genetic algorithm we discussed in class to select the best set of features from the 4 original features plus the 4 transformed features from Part 2 (for a total of 8 features).
    - For the initial population use the following sets of features:
        - z_1, sepal-length, sepal-width, petal-length, petal-width
        - z_1, z_2, sepal-width, petal-length, petal-width

        - z_1, z_2, z_3, sepal-width, petal-length
        - z_1, z_2, z_3, z_4, sepal-width
        - z_1, z_2, z_3, z_4, sepal-length
    - Run the algorithm for 50 generations
    - At the end of each generation, print out the features and the accuracy for the 5 best sets of features and the generation number.

Remember:

- Your Programming Assignments are individual-effort.
- You can brainstorm with other students and help them work through problems in their programs, but everyone should have their own unique assignment programs.