DigiPen Institute of Technology

CS 372- CS 596: Assignment # 1

Objectives:

- Show how to use built-in libraries to build K Nearest Neighbor (KNN) and Naïve Bayes classifier.
- Implement KNN from scratch.
- Evaluate KNN & Naïve Bayes Models and analyze the results.
- Compare and analyze according to the results of two algorithms, KNN and Naive Bayes.
- Understand how we apply KNN and Naïve Bayes to real world problems.

Question 1 [44 points]

Note: Use Python to solve this assignment as it has many well-established data science libraries. In this assignment we provided some helpful links to some of the libraries that you will be using.

Instructions

- Download the ***Iris Flower*** dataset from course Moodle and use it for Questions below.
- Load Data from CSV file to Panda's data frame.

[40 points] Exploratory Data Analysis For Iris Dataset.

1. **[20 points]** Check for missing data and data duplication. Upon detection,
    a. **[10 points]** show records with missing values.
        i. **[3 points]** Implement both mean and median imputation.
        ii. **[4 points]** Compare mean/median imputation vs. dropping rows. Which one is better for sepal_width?
    b. **[10 points]** Show duplicated records.
        i. **[4 points]** Explain clearly how you dealt with them. Justify your choice.
        ii. **[3 points]** Explain why you think we have duplicates (what it means).
2. **[12 points]** Draw Box Plot to spot outliers. Create boxplots for each feature grouped by species. Make sure to add necessary labels for each plot. Note that you could create Boxplot graphs using Seaborn library.
    a. **[2 points]** Explain how Box Plot consider outliers in the dataset.
    b. **[3 points]** Analyze and comment on the plot.
    c. **[2 points]** Identify which species has the most outliers and hypothesize why (e.g., measurement errors).
3. **[8 points]** Create scatter plot matrix for Iris flowers descriptive features. Make sure to add necessary labels to each plot such axes, etc.

a. **[3 points]** Explain which two features show the clearest separation between species.
b. **[3 points]** Explain the plot.

4. **[8 points]** Plot Heatmap correlation for descriptive features (no species). Make sure to clearly indicate the name of the features on the heatmap as indicated in the dataset and also display/show the value of the correlation on the heatmap.
    a. Remember that correlation matrix *helps us to quantify and summarize the **direction and strength** of relationships between variables.* The heatmap shall show:
    b. correlation between each pair of descriptive features (X) (in our case flowers attributes) - we need to know correlation between features and avoid multi-correlation features.
        i. Clearly indicate positively and negatively correlated features.
        ii. analyze and comment on the plot.

Split the data

[3 points] In the following parts of question 1, split your data 80/20. Train the model on the 80% fraction and then evaluate the accuracy on the 20% fraction.

1. Make sure that we preserve the proportion of each class in the splits.
2. Make sure that if I run the algorithm again, I will get the same split and also the proportion of the classes are preserved in the split. Remember to shuffle the data before split.

KNN [31 points]

3. **[2 points]** Use StandardScaler to put all descriptive features on the same scale.
4. **[8 points]** In this part we need to Implement KNN using sklearn library (use KNeighborsClassifier).
5. **[8 points]** Experiment with different distance metrics, namely, Euclidean metric and Manhattan distance. Compare between them and comment on your findings.
6. **[5 points]** Choose the best number of neighbors using 5-fold cross validation (using built-in) approach we studied in the class.
7. **[4 points]** K value vs. Accuracy: draw a 2D plot to show the average accuracy of KNN classifier vs different number of k's. Analyze your result and comment on your plot.
8. **[4 points]** Evaluate your model on the test data using Accuracy based on best K found in above. Comment on the result and print misclassified flowers in table format where you show the true label in one column and the predicted table in another column.

Naive Bayes [10 points]

1. **[5 points]** You need to implement Naïve Bayes using the built-in library and use it to predict the Species of iris in the test dataset.
2. **[2 points]** Use Accuracy to assess the performance of your classifier.

3. **[3 points]** Print misclassified flowers in table format where you show the true label in one column and the predicted table in another column and comment on your output.

Question 2 [56 points]

Problem description: Predict the onset of diabetes based on diagnostic measures.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Some information about the data :

- Number of pregnancies.
- Body Mass Index (BMI)
- Insulin level.
- Age.
- Glucose.
- Blood Pressure.
- Diabetes Pedigree Function.
- Outcome (Target Feature).

Instructions

1. Download the ***Pima Indians Diabetes Database*** data from course Moodle.
2. Load Data from CSV file to Panda's data frame.

[35 points] Exploratory Data Analysis.

Note: Leverage medical domain knowledge to provide meaningful processing and analysis.

1. **[10 points]** Check for missing data and data duplication. Upon detection,
   A. **[5 points]** show records with missing tuples and explain clearly how you dealt with them. Justify your choice.
   B. **[5 points]** Show duplicated records and explain clearly how you dealt with them. Justify your choice.
2. **[5 points]** Plot scatter matrix and analyze the distribution of each feature.
3. **[8 points]** Outliers Detection & Analysis
   A. [4 points] Draw Box Plot to visually identify potential outliers in each feature.
   B. **[4 points]** Document decisions on how to handle outliers based on your analysis and reasoning.
4. **[4 points]** Plot Heatmap correlation.
   A. Clearly indicate all positively and negatively correlated features.

5. **[8 points]** Identify features containing unrealistic zero values.
    A. Choose a method/ approach to handle those values.
    B. Justify.
    C. Evaluate and document effectiveness through comparative analysis:
        i. Generate histograms and density plots before and after imputation.
        ii. Document changes in descriptive statistics.

Splitting Data [2 point] Split your data 80/20. Train the model on the 80% fraction and then evaluate the accuracy on the 20% fraction.

i. Make sure that we preserve the proportion of each class in the splits.
ii. Make sure that if I run the algorithm again, I will get the same split and also the proportion of the classes are preserved in the split. Remember to shuffle the data before split.

Question 2 Part 1 [27 points]

1. **[7 points]** You need to implement KNN **from scratch**:
    a) **[5 points]** Implement Euclidean Distance from scratch (please, don't use built-in library).
    b) **[2 points]** Use Min-Max to normalize the dataset.

2. **[20 points] Tune the number of nearest neighbors k**
    - What is the optimal value of k? To figure it out:
        - **[10 points] Implement N-fold-cross validation from scratch**:
            - Train KNN classifier using 5-fold-cross validation using various values of k.
                a. Explain/ justify the choice of the range of K values.
                b. Choose k that obtained the highest accuracy.
        - Next, evaluate the classifier using the optimal value of K that we found using 5-fold cross validation on the test set (20% that we did not use in the 5-fold cross validation) and obtain the final results.
    - **[3 points]** In case of tie, KNN algorithm prefers the neighbor with closer distance to the query.
    - **[2 points]** K value vs. Accuracy: draw a 2D plot to show the accuracy of KNN classifier vs different number of k's.
    - **[5 points]** Evaluate your model on the test data using Accuracy based on best K you found using 5-fold-cross validation. Comment on the result and print misclassified records in table format where you show the true label in one column and the predicted label in another column.

Question 2 Part 2 [10 points]

- **[5 points]** You need to implement Gaussian Naïve Bayes using **built-in library** and use it to predict diabetes on PIMA dataset. Make sure to use the same training and testing dataset you used for KNN in Question 2 Part 1, so we are able to compare the performance of the models.
- **[5 points] Evaluation:**
    - **[3 points]** print misclassified records.
    - **[2 points]** Calculate the accuracy.

[12 points] Model Evaluation

Compare, comment, and analyze the result of all the classifiers you built thus far Naïve Bayes & KNN using built-in library (you implemented in Question 1) with KNN from scratch implementation in Question 2 on ***Pima Indians Diabetes Database*** data. To compare the performance of your classifier based on:

I. **[6 points]** execution time.
II. **[6 points]** Accuracy

Submit the following:

Final Submission:

Your notebook should include two types of cells: Code cells and markdown cells. Use the first to write code and comments that explain what the code is doing. And the later can be used to add text for analysis, summary, and conclusion.

a. **Code.**
    - Name your file: Assignment1_GroupName.ipynb. such as:
        - Assignment1_Group1.ipynb.
    - **[10 points]** Your code should contain:
        - appropriate comments to facilitate understanding.
        - Clear execution Instruction.
        - Write the name of the student who wrote each piece of code.
    - **[5 points]** In each cell you need to add the name of the students who wrote each piece of code and analysis.

b. [5 points] Summary at the end:

Make sure to include a full explanation of the results and analysis. Moreover, include:

- A high-level description of how your code works.
- The evaluation matrices you used such as accuracies you obtain under various settings.

- Discussion:
    - Explain which options work well and why.
    - If all your evaluation matrices are low, tell us what you have
    - tried to improve them and what you suspect is failing,
    - challenges, and explain your output.
- Conclusion and summary

c. Make sure to include a file for any use of AI generative tools and properly cite the use. Refer to the course syllabus for more information.