Assignment

In this assignment, we will implement k-nn and logistic regression classifiers and analyze feature importance in predicting accuracy.

For the dataset, we use "heart failure clinical records data set at UCI: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

Dataset Description: From the website: "This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features."

These 13 features are:

1. age: age of the patient (years)
2. anaemia: decrease of red blood cells or hemoglobin (boolean)
3. high blood pressure: if the patient has hypertension (boolean)
4. creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
5. diabetes: if the patient has diabetes (boolean)

6. ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
7. platelets: platelets in the blood (kiloplatelets/mL)
8. sex: woman or man (binary)
9. serum creatinine: level of serum creatinine in the blood (mg/dL)
10. serum sodium: level of serum sodium in the blood (mEq/L)
11. smoking: if the patient smokes or not (boolean)
12. time: follow-up period (days)

target death event: if the patient deceased (DEATH_EVENT = 1 or "+") during the follow-up period (boolean) or survived (DEATH_EVENT = 0 or "-")

We will focus on the following subset of four features:

1. f_1: creatinine phosphokinase
2. f_2: serum creatinine
3. f_3: serum sodium
4. f_4: platelets

and try to establish the importance of features in predictions.

Question 1:

1. load the data into Pandas dataframe. Extract two dataframes with the above 4 features: df_0 for surviving patients (DEATH_EVENT = 0) and df_1 for deceased patients (DEATH_EVENT = 1)

2. for each dataset, construct the visual representations of correponding correlation matrices M_0 (from df_0) and M_1 (from df_1) and save the plots into two separate files

3. examine your correlation matrix plots visually and answer the following:

   (a) which features have the highest correlation for surviving patients?

   (b) which features have the lowest correlation for surviving patients?

   (c) which features have the highest correlation for deceased patients?

   (d) which features have the lowest correlation for deceased patients?

   (e) are results the same for both cases?

4. for each class and for each feature f_1, f_2, f_3, f_4, compute its mean $\mu()$ and standard deviation $\sigma()$. Round the results to

2 decimal places and summarize them in a table as shown below:

class	$\mu(f_1)$	$\sigma(f_1)$	$\mu(f_2)$	$\sigma(f_2)$	$\mu(f_3)$	$\sigma(f_3)$	$\mu(f_4)$	$\sigma(f_4)$
0								
1								
all								

5. examine your table. Are there any obvious patterns in the distribution of in each class

Question 2:

1. split your dataset X into training X_{train} and $X_{testing}$ parts (50/50 split). Using "pairplot" from seaborn package, plot pairwise relationships in X_{train} separately for class 0 and class 1. Save your results into 2 pdf files "survived.pdf" and "not-survived.pdf"

2. visually examine your results. Come up with three simple comparisons that you think may be sufficient to predict a survival. For example, your classifier may look like this:

```
# assume you are examining a patient
# with features f_1,f_2,f_3 and f_4
```

```
# your rule may look like this:
if (f_1 > 4) and (f_2 > 8) and (f_4 < 25):
    x = "survive"
else:
    x = "not_survive"
```

3. apply your simple classifier to X_{test} and compute predicted class labels

4. compare your predicted class labels with true labels in X_{test}, compute the following:

   (a) TP - true positives (your predicted label is $+$ and true label is $+$)

   (b) FP - false positives (your predicted label is $+$ but true label is $-$

   (c) TN - true negativess (your predicted label is $-$ and true label is $-$

   (d) FN - false negatives (your predicted label is $-$ but true label is $+$

   (e) TPR = TP/(TP + FN) - true positive rate. This is the fraction of positive labels that your predicted correctly. This is also called sensitivity, recall or hit rate.

   (f) TNR = TN/(TN + FP) - true negative rate. This is the

fraction of negative labels that your predicted correctly. This is also called specificity or selectivity.

5. summarize your findings in the table as shown below:

TP	FP	TN	FN	accuracy	TPR	TNR

6. does your simple classifier gives you higher accuracy on identifying "fake" bills or "real" bills" Is your accuracy better than 50% ("coin" flipping)?

Question 3 (use k-NN classifier using sklearn library)

1. take $k = 3, 5, 7$. For each k, generate X_{train} and X_{test} using 50/50 split as before. Train your k-NN classifier on X_{train} and compute its accuracy for X_{test}

2. plot a graph showing the accuracy. On x axis you plot k and on y-axis you plot accuracy. What is the optimal value k^* of k?

3. use the optimal value k^* to compute performance measures and summarize them in the table

TP	FP	TN	FN	accuracy	TPR	TNR

4. is your k-NN classifier better than your simple classifier for any of the measures from the previous table?

Question 4: One of the fundamental questions in machine learning is "feature selection". We try to come up with the least number of features and still retain good accuracy. The natural question is whether some of the features are important or can be dropped.

1. take your best value k^*. For each of the four features $f_1, \ldots, f_4$, generate new X_{test} and X_{train} and drop that feature from both X_{train} and X_{test}. Train your classifier on the "truncated" X_{train} and predict labels on X_{test} using just 3 remaining features. You will repeat this for 4 cases: (1) just f_1 is missing, (2) just f_2 is missing, (3) just f_3 missing and (4) just f_4 is missing. Compute the accuracy for each of these scenarios.

2. did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?

3. which features, when removed, contributed the most to loss

of accuracy?

4. which features, when removed, contributed the least to loss of accuracy?

Question 5 (use logistic regression classifier using sklearn library)

1. Use 50/50 split to generate new X_{train} and X_{test}. Train your logistic regression classifier on X_{train} and compute its accuracy for X_{test}

2. summarize your performance measures in the table

TP	FP	TN	FN	accuracy	TPR	TNR

3. is your logistic regression better than your simple classifier for any of the measures from the previous table?

4. is your logistic regression better than your k-NN classifier (using the best k^*) for any of the measures from the previous table?

Question 6: We will investigate the change in accuracy when removing one feature. This is similar to question 4 but now we use logistic regression.

1. For each of the four features $f_1, \ldots, f_4$, generate new X_{train} and X_{test} and drop that feature from both X_{train} and X_{test}. Train your logistic regression classifier on the "truncated" X_{train} and predict labels on "truncated" X_{test} using just 3 remaining features. You will repeat this for 4 cases: (1) just f_1 is missing, (2) just f_2 missing, (3) just f_3 missing and (4) just f_4 is missing. Compute the accuracy for each of these scenarios.

2. did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?

3. which features, when removed, contributed the most to loss of accuracy?

4. which features, when removed, contributed the least to loss of accuracy?

5. is the relative significance of features the same as you obtained using k-NN?