Assignment

In this assignment, we will implement k-nn and logistic regression classifiers to detect "fake" banknotes and analyze the comparative importance of features in predicting accuracy.

For the dataset, we use "banknote authentication dataset" from the machine Learning depository at UCI: `https://archive.ics.uci.edu/ml/datasets/banknote+authentication`

Dataset Description: From the website: "This dataset contains 1,372 examples of both fake and real banknotes. Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images."

There are 4 continuous attributes (features) and a class:

1. f_1 - variance of wavelet transformed image
2. f_2 - skewness of wavelet transformed image
3. f_3 - curtosis of wavelet transformed image
4. f_4 - entropy of image

5. class (integer)

In other words, assume that you have a machine that examines a banknote and computes 4 attributes (step 1). Then each banknote is examined by a much more expensive machine and/or by human expert(s) and classified as fake or real (step 2). Step 2 is very time-consuming and expensive. You want to build a classifier that would give you good results after step 1 only.

We assume that class 0 are good banknotes. We will use color **"green"** or "+" for legitimate banknotes. Class 1 are assumed to be fake banknotes and we will use color **"red"** or "−" for counterfeit banknotes. These are "true" labels.

Question 1:

1. load the data into Pandas dataframe and add a column "color". For each class 0, this should contain "green" and for each class 1 it should contain "red"

2. 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:

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

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								

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 "good_bills.pdf" and "fake_bills.pdf"

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

```
# assume you are examining a bill
# 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 = "good"
else:
    x = "fake"
```

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:

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

TP	FP	TN	FN	accuracy	TPR	TNR

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

1. take $k = 3, 5, 7, 9, 11$. 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?

5. consider a bill x that contains the last 4 digits of your BUID as feature values. What is the class label predicted for this

bill by your simple classifier? What is the label for this bill predicted by k-NN using the best k^*?

Question 4: One of the fundamental questions in machine learning is "feature selection". We try to come up with a 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 missing, (3) just f_3 missing and (4) just f_4 is missing. Compute the accuracy for each of these scenarious.

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

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

4. which feature, 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?

5. consider a bill x that contains the last 4 digits of your BUID as feature values. What is the class label predicted for this bill x by logistic regression? Is it the same label as predicted by k-NN?

Question 6: We will investigate 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 scenarious.

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

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

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

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