CSC 315 / 615
Project 2
Due 03 / 08 / 2023

Counterfeit Money Analysis

Overview

Your goal is to write a program to analyze and classify the Banknote Authentication dataset using matplotlib, numpy, and pandas. Banknote Authentication is a dataset contains the imaging features for Real and Counterfeit banknotes, including variance, skewness, kurtosis, and entropy to compare the real and counterfeit

You are also going to plot the average of the imaging features by plotting an "X" on each plot for the average value. The average value for each variety is called the "cluster center".

Next, you are going to create a simple classifier to "guess" the legitimacy based on it's features. Imagine if you had an unknown bank note and you measured the imaging features in the dataset. You can compare these features with the "centroids" for the real and counterfeit bank notes, and whichever "centroid" is closest is the predicted label.

For extra credit, we can calculate the Gaussian probabilities of drawing the unknown bank note from either the real or counterfeit distributions. Whichever event (real or counterfeit) is more probable is the predicted label of the bank note.

Be sure to review the sample output first, then read these requirements.

R1. Your program must display the Banknote dataset using six different color coded scatterplots **(a-d 40 pts)**

R.1.a. Scatter plots must have correct titles and axies **10 pts**

The six different color-coded scatter plots must be titled

	x-axis	y-axis
Plot 1	variance	vs skewness (accuracy %)
Plot 2	variance	vs kurtosis (accuracy %)
Plot 3	variance	vs entropy (accuracy %)
Plot 4	skewness	vs kurtosis (accuracy %)
Plot 5	skewness	vs entropy (accuracy %)

Plot 6 kurtosis vs entropy (accuracy %)

R.1.b. Scatter plots must display each bank note using a different color points **10 pts**

		Predicted	
		Real	**Counterfeit**
Label	**Real**	`#8888ff`	`#ccccff`
	Counterfeit	`#ffcccc`	`#ff8888`

This color scheme ensures that the scatter plots exhibit the same colors are the specified output, with real bank notes in light blue, and counterfeit bank notes in light red. Also correct predictions are given a more prominent shade than incorrect predictions.

More information on hexadecimal color codes is available here
Link How to Read Hex Color Codes

R.1.c. Scatter plots must display correct axis titles **10 pts**

For a scatterplot titled attribute1 vs attribute2

For example, the scatterplot variance vs entropy

The x-axis shows variance whereas the y-axis shows entropy

R.1.d. Appearance of scatter plots must match the sample output **10 pts**

The scatter plots must visually appear similar to the scatter plots provided.

R.2 You must calculate and display the attribute centroids in the plots **(a-c 20 pts)**

R.2.a You must calculate the centroids for the real and counterfeit (fake) bank notes **10 pts**

$$\vec{\mu}_r = \begin{bmatrix} \mu_{r0} \\ \mu_{r1} \\ \mu_{r2} \\ \mu_{r3} \end{bmatrix} = \begin{bmatrix} \text{mean of variance for real bank notes} \\ \text{mean of skewness for real bank notes} \\ \text{mean of kutorsis for real bank notes} \\ \text{mean of entropy for real bank notes} \end{bmatrix}$$

$$\vec{\mu}_f = \begin{bmatrix} \mu_{f0} \\ \mu_{f1} \\ \mu_{f2} \\ \mu_{f3} \end{bmatrix} = \begin{bmatrix} \text{mean of variance for counterfeit bank notes} \\ \text{mean of skewness for counterfeit bank notes} \\ \text{mean of kutorsis for counterfeit bank notes} \\ \text{mean of entropy for counterfeit bank notes} \end{bmatrix}$$

R.2.b You must display the centroids with a black X within the plot **5 pts**

The “centroids” must appear as a black X within the center of each cluster on each plot.

R.2.c You must display real or counterfeit centered underneath the centroid **5 pts**

See sample output plots that display “real” or “counterfeit” under the centroid X

R.3 Bank note classification based on distance to centroid **30 pts**

Now we must create a classifier based on the distance to the centroid.

Imagine if you have an unknown bank note **x** and you wish to determine if this bank note is real or fake. You image the bank note, and measure the variance, skewness, kurtosis, and entropy of the banknote as follows.

$$\vec{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \text{variance of bank note x} \\ \text{skewness of bank note x} \\ \text{kurtosis of bank note x} \\ \text{entropy of bank note x} \end{bmatrix}$$

You can determine how similar bank note **x** is to the centroids of the real and counterfeit bank notes using Euclidian distance as follows,

$$\text{dist}(x, \mu_r) = \sqrt{(x_0-\mu_{r0})^2 + (x_1-\mu_{r1})^2 + (x_2-\mu_{r2})^2 + (x_3-\mu_{r3})^2}$$

$$\text{dist}(x, \mu_f) = \sqrt{(x_0-\mu_{f0})^2 + (x_1-\mu_{f1})^2 + (x_2-\mu_{f2})^2 + (x_3-\mu_{f3})^2}$$

If **x** is closer to the real centroid, then we predict it to be real. If it is closer to the fake centroid, then we predict it to be counterfeit.

$$\text{if} \quad \text{dist}(x, \mu_r) < \text{dist}(x, \mu_f)$$

predict x is real

otherwise

predict x is fake (counterfeit)

R.4. You must display the correct classification accuracy within the title of each plot 10 pts

For distance-based classifier, accuracy should be 70.70%

For extra-credit Gaussian classifier, accuracy should be 84.77%

E.X. For extra credit, you must use a Gaussian probability density classifier. + 10 pts

Define the standard deviations for the real and fake samples as follows,

$$\vec{\sigma_r} = \begin{bmatrix} \sigma_{r0} \\ \sigma_{r1} \\ \sigma_{r2} \\ \sigma_{r3} \end{bmatrix} = \begin{bmatrix} \text{standard deviation of variance for real bank notes} \\ \text{standard deviation of skewness for real bank notes} \\ \text{standard deviation of kutorsis for real bank notes} \\ \text{standard deviation of entropy for real bank notes} \end{bmatrix}$$

$$\vec{\sigma_f} = \begin{bmatrix} \sigma_{r0} \\ \sigma_{r1} \\ \sigma_{r2} \\ \sigma_{r3} \end{bmatrix} = \begin{bmatrix} \text{standard deviation of variance for counterfeit bank notes} \\ \text{standard deviation of skewness for counterfeit bank notes} \\ \text{standard deviation of kutorsis for counterfeit bank notes} \\ \text{standard deviation of entropy for counterfeit bank notes} \end{bmatrix}$$

We may now use the Gaussian distribution to model the probability of any bank note **x** being drawn from the **real** and **counterfeit** distributions based on only a single feature as a time as follows,

$$p(\text{real}, x_0) = N(x_0, \mu_{r0}, \sigma_{r0}) = \frac{1}{\sigma_{r0}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_0-\mu_{r0}}{\sigma_{r0}}\right)^2}$$ real probability based on variance alone

$$p(\text{real}, x_1) = N(x_1, \mu_{r1}, \sigma_{r1}) = \frac{1}{\sigma_{r1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu_{r1}}{\sigma_{r1}}\right)^2}$$ real probability based on skewness alone

$$p(\text{real}, x_2) = N(x_2, \mu_{r2}, \sigma_{r2}) = \frac{1}{\sigma_{r2}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu_{r2}}{\sigma_{r2}}\right)^2}$$ real probability based on kurtosis alone

$$p(\text{real}, x_3) = N(x_3, \mu_{r3}, \sigma_{r3}) = \frac{1}{\sigma_{r3}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_3-\mu_{r3}}{\sigma_{r3}}\right)^2}$$ real probability based on entropy alone

$$p(\text{fake}, x_0) = N(x_0, \mu_{f0}, \sigma_{f0}) = \frac{1}{\sigma_{f0}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_0-\mu_{f0}}{\sigma_{f0}}\right)^2}$$ fake probability based on variance alone

$$p(\text{fake}, x_1) = N(x_1, \mu_{f1}, \sigma_{f1}) = \frac{1}{\sigma_{f1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu_{f1}}{\sigma_{f1}}\right)^2}$$ fake probability based on skewness alone

$$p(\text{fake}, x_2) = N(x_2, \mu_{f2}, \sigma_{f2}) = \frac{1}{\sigma_{f2}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu_{f2}}{\sigma_{f2}}\right)^2}$$ fake probability based on kurtosis alone

$$p(\text{fake}, x_3) = N(x_3, \mu_{f3}, \sigma_{f3}) = \frac{1}{\sigma_{f3}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_3-\mu_{f3}}{\sigma_{f3}}\right)^2}$$ fake probability based on entropy alone

As the bank note **x** has four features $\mathbf{x_0, x_1, x_2, x_3}$, we can assume feature independence so the probability of drawing a **real (fake)** bank note **x** with all four features simultaneously is equal to the product of the probabilities of drawing a **real (fake)** bank note based on each feature individually.

$$p(\text{real}, x) = p(\text{real}, x_0)\ p(\text{real}, x_1)\ p(\text{real}, x_2)\ p(\text{real}, x_3)$$

$$p(\text{fake}, x) = p(\text{fake}, x_0)\ p(\text{fake}, x_1)\ p(\text{fake}, x_2)\ p(\text{fake}, x_3)$$

Once you calculate **p(real, x)** and **p(fake, x)** you can predict whether **x** is a real or counterfeit bank note by comparing the probabilities as follows,

```
if p(real, x) > p(fake, x)
    predict x is real
else
    predict x is fake (counterfeit)
```

Additional Deductions

D.1	**If program fails to run, but TA can easily fix the issue**	**-10 pnts**
D.2	**If program fails to run, but TA cannot easily fix the issue**	**-50 pnts**
D.3	**If student forgets to write their name in the comment**	**-10 pnts**

Note: 615 students receive an automatic -10 pnt deduction, which can be made up by completing the extra credit.

Note: Students who complete the extra credit (Gaussian classifier) are not required to implement the basic distance classifier.