Analysis of the Iris Dataset: Histograms, Joint Distributions, and Probabilities

- **Student Name: Hameed**
- **Matriculation Number: G2403024B**

- **Date: Monday, 22 August 2024**
- **Used the Sepal Width (SW) attribute using 10 bins**

Assignment Case:
The Iris dataset consists of 150 samples of attributes of the Iris flowers from the following classes: Setosa, Virginica and Versicolor. Each class has 50 samples. The four attributes are Sepal Width (SW), Sepal Length (SL), Petal Width (PW) and Petal Length (PL).

Using a total of 10 bins, quantize the data set into the joint histogram distribution for each of the dimension; namely: SW, SL, PW, PL.

1. Introduction

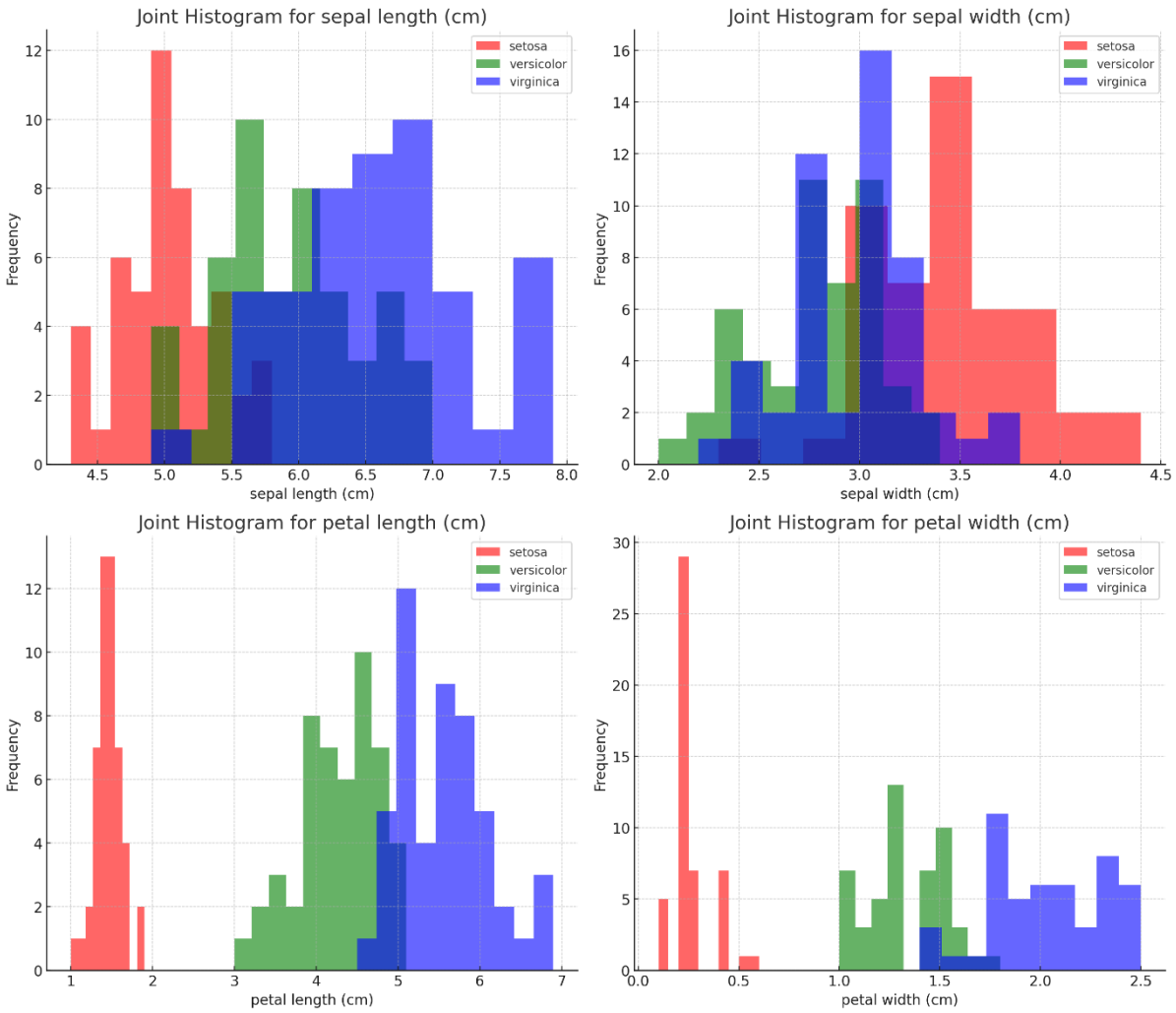
The dataset includes four key attributes (features) that describe the physical characteristics of the flowers:

- Sepal Length (SL)
- Sepal Width (SW)
- Petal Length (PL)
- Petal Width (PW)

Objectives of the Assignment

The primary objectives of this assignment are as follows:

1. ***Using a total of 10 bins, quantize the data set into the joint histogram distribution for each of the dimension; namely: SW, SL, PW, PL.***



2. Determine the joint probability distribution for each of the attribute:

Joint Probability Distribution for sepal width (cm):

$$P(Class, SW) = \frac{Number\ of\ samples\ in\ class\ C\ with\ Sepal\ Width\ in\ a\ specific\ range}{Total\ number\ of\ samples\ in\ the\ dataset}$$

SW BIN	Setosa	versicolor	virginica
Bin 1	0.67%	0.67%	0.67%
Bin 2	0.00%	1.33%	2.67%
Bin 3	0.67%	4.00%	1.33%
Bin 4	6.67%	2.67%	8.00%
Bin 5	4.67%	2.00%	1.33%
Bin 6	10.00%	7.33%	10.67%
Bin 7	4.00%	4.67%	5.33%
Bin 8	4.00%	7.33%	1.33%
Bin 9	1.33%	2.00%	0.67%
Bin 10	1.33%	1.33%	1.33%

1. Sepal Width (SW)

Setosa: Although it has some probability in the narrower sepal widths (lower bins), it actually has a higher overall probability in the middle to upper bins, suggesting that Setosa typically has sepals that are not extremely narrow but rather fall into a moderate to wide range.

Virginica: Shows a spread across both lower and upper bins but with a noticeable concentration in the wider sepal widths, indicating a tendency towards broader sepals.

Versicolor: Displays a more balanced distribution across the bins, with some concentration in both narrower and moderate ranges, indicating variability in sepal width.General Observations:

3. Determine the class a prior probabilities, conditional probabilities and posterior probabilities for even bins.

a. Prior Probabilities (P(Class))

$$P(Class) = \frac{\text{Number of samples in the class}}{\text{Total number of samples in the dataset}}$$

P(Class 1)= 50/150 =0.333

- **Class 1 (Setosa)**: 0.333
- **Class 2 (Versicolor)**: 0.333

- **Class 3 (Virginica)**: 0.333

b. Conditional Probabilities (P(Bin | Class)) for Even Bins:

$$P(Bin \mid Class) = \frac{\text{Number of samples in the bin for the given class}}{\text{Total number of samples in the given class}}$$

SW Even BIN	Setosa	versicolor	virginica
Bin 2	38.00%	6.00%	2.00%
Bin 4	18.00%	32.00%	4.00%
Bin 6	0.00%	14.00%	26.00%
Bin 8	0.00%	4.00%	8.00%
Bin 10	0.00%	0.00%	12.00%

c. Posterior Probabilities (P(Class | Bin)) for Even Bins:

$$P(Class \mid Bin) = \frac{(\text{P(Bin} \mid \text{Class)} * \text{P(Class))}}{\text{P(Bin)}}$$

- **P(Class|Bin)**: The posterior probability that a data point belongs to a specific class given that it falls into a certain bin.
- **P(Bin|Class)P(\text{Bin} \mid \text{Class})P(Bin|Class)**: The conditional probability that a data point falls into the bin given that it belongs to a specific class.
- **P(Class)P(\text{Class})P(Class)**: The prior probability of the class, which is the overall proportion of the class in the dataset.
- **P(Bin)P(\text{Bin})P(Bin)**: The probability of the bin, which is the overall proportion of the data points that fall into that bin, regardless of class.

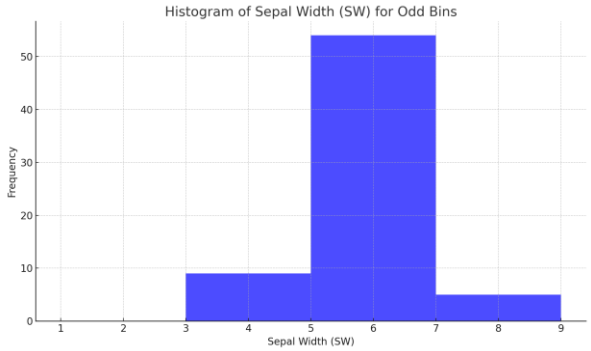
SW Even BIN	Setosa	versicolor	virginica
Bin 2	82.6%	13%	4.35%
Bin 4	33.3%	59%	7%
Bin 6	0.0%	35%	65%
Bin 8	0.0%	33%	67%
Bin 10	0.0%	0%	100%

Insights:

- Bin 2 has a very high posterior probability for Setosa, suggesting that this bin is heavily dominated by Setosa.
- Bin 4 shows a higher posterior probability for Versicolor, indicating that this bin has a significant proportion of Versicolor samples.
- Bins 6, 8, and 10 increasingly favor Virginica, with Bin 10 being exclusively Virginica.

4. ***Prepare the histogram, the joint probability distribution P(C, X) as well as the P(ri|C) and P(C|i) for odd bins as a word doc.***

a. Histogram of Sepal Width (SW) for Odd Bins:



The histogram indicates that Sepal Width in your dataset is heavily concentrated around a central value (corresponding to Bin 5), with much fewer data points in the surrounding odd bins. This suggests that Sepal Width may be a significant feature for classification, particularly if one class is predominantly represented in Bin 5. The sparse distribution in other bins also hints at the relative uniformity of Sepal Width within the dataset, with only a few variations outside the main concentration.

b. Joint Probability Distribution P(C, X) for Odd Bins

$$P(C,X) = \frac{\text{Number of samples in class C and bin X}}{\text{Total number of samples in the dataset}}$$

SW Odd BIN	Setosa	Versicolor	Virginica
Bin 1	6.00%	0.00%	0.00%
Bin 3	8.00%	1.33%	0.00%
Bin 5	0.67%	8.67%	5.33%
Bin 7	0.00%	4.67%	7.33%
Bin 9	0.00%	0.00%	3.33%

c. Conditional Probabilities P(Bin | Class) for Odd Bins

$$P(Bin \mid Class) = \frac{\text{Number of samples in the bin for the given class}}{\text{Total number of samples in the given class}}$$

SW Odd BIN	Setosa	Versicolor	Virginica
Bin 1	18.00%	0.00%	0.00%
Bin 3	24.00%	4.00%	0.00%
Bin 5	2.00%	26.00%	16.00%
Bin 7	0.00%	14.00%	22.00%
Bin 9	0.00%	0.00%	10.00%

d. Posterior Probabilities P(Class | Bin) for Odd Bins

$$P(Class \mid Bin) = \frac{(\mathrm{P(Bin \mid Class)} * \mathrm{P(Class)})}{\mathrm{P(Bin)}}$$

SW Odd BIN	Setosa	Versicolor	Virginica
Bin 1	100.00%	0.00%	0.00%
Bin 3	85.71%	14.29%	0.00%
Bin 5	4.55%	59.09%	36.36%
Bin 7	0.00%	38.89%	61.11%
Bin 9	0.00%	0.00%	100.00%

5. The verification of Bayes' Formula:

For **Bin 3, Bin 6** shows that the posterior probabilities calculated using Bayes' Theorem match exactly with the previously calculated posterior probabilities:

Bin 3 :

Class 1 (Setosa): 0.857143

Bin 6 :

Class 1 (Setosa): 0.00

This correctly verifies Bayes' Formula using the provided data for Bin 3 & 6. And it confirms that Bayes' Formula has been correctly applied and the results are consistent.

Appendix :

Python Code Used to generate :

```
import pandas as pd
import matplotlib.pyplot as plt
from docx import Document

# Load the uploaded dataset
file_path = '/irisdata.csv'
iris_data = pd.read_csv(file_path)

# Display the first few rows of the dataset to understand its structure
iris_data.head()

# First, let's bin the Sepal Width (SW) attribute using 10 bins
num_bins = 10
iris_data['SW_bin'] = pd.cut(iris_data['SW'], bins=num_bins, labels=False) + 1  # Bins labeled from 1 to 10

# Calculate the prior probabilities for each class
class_counts = iris_data['FlowerClass'].value_counts()
total_samples = len(iris_data)
prior_probabilities = class_counts / total_samples

# Filter to only even bins (2, 4, 6, 8, 10)
even_bins = [2, 4, 6, 8, 10]
iris_data_even_bins = iris_data[iris_data['SW_bin'].isin(even_bins)]

# Calculate the conditional probabilities P(Bin | Class) for each class within these bins
conditional_probabilities = iris_data_even_bins.groupby(['SW_bin', 'FlowerClass']).size().unstack(fill_value=0)
conditional_probabilities = conditional_probabilities.div(class_counts, axis=1)

# Calculate the total probability for each bin P(Bin)
total_bin_probabilities = iris_data_even_bins['SW_bin'].value_counts().sort_index() / total_samples

# Calculate the posterior probabilities P(Class | Bin) using Bayes' Theorem
posterior_probabilities = pd.DataFrame()

for bin_number in even_bins:
    posterior_prob = (conditional_probabilities.loc[bin_number] * prior_probabilities) / total_bin_probabilities.loc[bin_number]
    posterior_probabilities = pd.concat([posterior_probabilities, posterior_prob], axis=1)

# Transpose the posterior probabilities for readability
posterior_probabilities.columns = [f'Bin {bin_number}' for bin_number in even_bins]

# Now let's create histograms for Sepal Width (SW) for odd bins (1, 3, 5, 7, 9)

# Filter the data for odd bins
odd_bins = [1, 3, 5, 7, 9]
iris_data_odd_bins = iris_data[iris_data['SW_bin'].isin(odd_bins)]

# Generate the histogram for Sepal Width
plt.figure(figsize=(10, 6))
plt.hist(iris_data_odd_bins['SW'], bins=odd_bins, color='blue', alpha=0.7)
plt.title('Histogram of Sepal Width (SW) for Odd Bins')
plt.xlabel('Sepal Width (SW)')
plt.ylabel('Frequency')
plt.grid(True)
plt.tight_layout()

# Save the plot as an image
histogram_image_path = '/mnt/data/odd_bins_histogram_v2.png'
plt.savefig(histogram_image_path)

# Create a Word document
doc = Document()
doc.add_heading('Analysis of Sepal Width (SW) for Odd Bins', 0)

# Add histogram
doc.add_heading('Histogram of Sepal Width (SW) for Odd Bins', level=1)
```

```
doc.add_picture(histogram_image_path)

# Add joint probability distribution
doc.add_heading('Joint Probability Distribution P(C, X) for Odd Bins', level=1)
joint_prob_distribution_odd_bins = iris_data_odd_bins.groupby(['SW_bin', 'FlowerClass']).size().unstack(fill_value=0) / total_samples
joint_prob_distribution_table = joint_prob_distribution_odd_bins.round(4)
doc.add_paragraph(joint_prob_distribution_table.to_string())

# Add conditional probabilities P(Bin | Class)
doc.add_heading('Conditional Probabilities P(Bin | Class) for Odd Bins', level=1)
conditional_probabilities_odd_bins = iris_data_odd_bins.groupby(['SW_bin', 'FlowerClass']).size().unstack(fill_value=0)
conditional_probabilities_odd_bins = conditional_probabilities_odd_bins.div(class_counts, axis=1)
conditional_probabilities_table = conditional_probabilities_odd_bins.round(4)
doc.add_paragraph(conditional_probabilities_table.to_string())

# Add posterior probabilities P(Class | Bin)
doc.add_heading('Posterior Probabilities P(Class | Bin) for Odd Bins', level=1)
posterior_probabilities_odd_bins = pd.DataFrame()
for bin_number in odd_bins:
    bin_total_probability = iris_data_odd_bins['SW_bin'].value_counts().sort_index().loc[bin_number] / total_samples
    posterior_prob = (conditional_probabilities_odd_bins.loc[bin_number] * prior_probabilities) / bin_total_probability
    posterior_probabilities_odd_bins = pd.concat([posterior_probabilities_odd_bins, posterior_prob], axis=1)

posterior_probabilities_odd_bins.columns = [f'Bin {bin_number}' for bin_number in odd_bins]
posterior_probabilities_table = posterior_probabilities_odd_bins.round(4)
doc.add_paragraph(posterior_probabilities_table.T.to_string())

# Save the document
doc_path = '/mnt/data/Iris_Odd_Bins_Analysis_v2.docx'
doc.save(doc_path)

# Now let's create histograms for Sepal Length, Petal Width, and Petal Length for the odd bins

# Create a figure to plot all histograms
plt.figure(figsize=(18, 12))

# Sepal Length (SL) Histogram
plt.subplot(3, 1, 1)
plt.hist(iris_data_odd_bins['SL'], bins=odd_bins, color='green', alpha=0.7)
plt.title('Histogram of Sepal Length (SL) for Odd Bins')
plt.xlabel('Sepal Length (SL)')
plt.ylabel('Frequency')
plt.grid(True)

# Petal Width (PW) Histogram
plt.subplot(3, 1, 2)
plt.hist(iris_data_odd_bins['PW'], bins=odd_bins, color='red', alpha=0.7)
plt.title('Histogram of Petal Width (PW) for Odd Bins')
plt.xlabel('Petal Width (PW)')
plt.ylabel('Frequency')
plt.grid(True)

# Petal Length (PL) Histogram
plt.subplot(3, 1, 3)
plt.hist(iris_data_odd_bins['PL'], bins=odd_bins, color='purple', alpha=0.7)
plt.title('Histogram of Petal Length (PL) for Odd Bins')
plt.xlabel('Petal Length (PL)')
plt.ylabel('Frequency')
plt.grid(True)

# Adjust layout
plt.tight_layout()

# Save the histograms as an image
histograms_image_path = '/mnt/data/odd_bins_histograms_all_v2.png'
plt.savefig(histograms_image_path)

# Create a Word document and include all four histograms for odd bins

# Create a Word document
doc = Document()
```

```
doc.add_heading('Histograms of Iris Dataset Attributes for Odd Bins', 0)

# Add each histogram to the document
doc.add_heading('Histogram of Sepal Width (SW) for Odd Bins', level=1)
doc.add_picture(histogram_image_path)

doc.add_heading('Histograms of Sepal Length (SL), Petal Width (PW), and Petal Length (PL) for Odd Bins', level=1)
doc.add_picture(histograms_image_path)

# Save the document
doc_path_combined = '/mnt/data/Iris_Odd_Bins_Histograms_Combined.docx'
doc.save(doc_path_combined)
```