Overview of Data

The four input features are petal width (PW), petal length (PL), sepal width (SW), and sepal length (SL). The Parzen window empirical distributions for each of the features are seen in Fig. 1 for the three classes: *Setosa, Versicolor*, and *Verginica*.

As seen in the figure, the *Verginica* is well separated from the other two classes, and the petal features appear to be more discriminating. The distance properties between the three classes can be seen in Fig. 2. As could be predicted by the 1-dimensional empirical distributions, *Setosa* and *Versicolor* have poor separation, while *Verginica* and *Versicolor* have excellent separation.
Optimal Bayes Classifier

The Bayes classifier assumes a distribution for the class likelihood, \( f(x|c_i) \), the likelihood of a feature vector \( x \) arising given that it comes from class \( c_i \). The natural choice in this case is the multivariate Gaussian distribution. The parameters for this distribution are found for each class from the training data. These parameters are the covariance matrix \( \Sigma_i \) and the mean vector \( \mu_i \). In \( k \) dimensions, the likelihood takes the form,

\[
 f(x|c_i) = \frac{1}{(2\pi)^{k/2} |\Sigma_i|^{1/2}} \exp\left( -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right)
\]

We choose the class \( c_i \) that gives the maximum a posteriori probability (MAP). This is the probability of the observation \( x \) arising from class \( c_i \) given that observation. Bayes’ rule provides the MAP in terms of the likelihood, which we have a model for.

\[
 f(c_i|x) = \frac{f(x|c_i) f(c_i)}{f(x)}
\]

Since the denominator \( f(x) \) is common to all classes, and we can assume the prior distribution on the classes \( f(c_i) \) is the same, the MAP rule reduces to the maximum likelihood (ML) rule. The classification problem is now phrased as follows,

\[
 c_{MAP} = c_{ML} = \arg \max_{c_i} f(x|c_i)
\]

The confusion matrix counts the decisions made by the classifier. The rows represent the true class and the columns represent the decision.

<table>
<thead>
<tr>
<th>True class</th>
<th>Setosa</th>
<th>Versicolor</th>
<th>Virginica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setosa</td>
<td>47</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Versicolor</td>
<td>2</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>Virginica</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

As can be seen from the confusion matrix, 5 errors were made out of the 150 samples, which is an error rate of 3.33%.

Fig. 3 depicts the decision boundary for the optimal Bayes classifier. Each plot is a 3-dimensional slice of the 4-dimensional input space at the given value of sepal length. The feature names are abbreviated. An observation is allocated to the plot with the sepal length closest to its own value.
The Bayes error provides a measure of the error associated with the overlap in likelihood functions. That is, even in the ideal case, where the functions come directly from a multivariate Gaussian distribution, the classifier is not perfect. The Bayes error is found by integrating the distribution on the wrong side of the threshold, but this is complicated in higher dimensions. Instead, we generate a large number of multivariate Gaussian random vectors with the same
means and covariances as our data. Then we classify these and determine the error rate. Generating 100,000 vectors for each set of class parameters yielded a Bayes error of \textbf{2.80\%}.

**Fischer’s Linear Discriminant**

Fischer’s linear discriminant is a 2-class method, which provides a vector \( w \) upon which the projection of features from each class have maximum separation. Given a binary class \( c_i \), its mean vector \( \mu_i \), and its covariance matrix \( \Sigma_i \), the optimal \( w \) is,

\[
    w = (\Sigma_0 + \Sigma_1)^{-1}(\mu_0 - \mu_1)
\]

Features to be classified are projected onto \( w \) and a threshold \( \delta \) is applied. This threshold is often the mean of the projections of the class means,

\[
    c = \frac{w^T(\mu_0 + \mu_1)}{2}
\]

Since we have 3 classes, we will conduct classification in a pairwise manner. This means that we will have a \( w \) between each of the three pairs of classes. A feature will first be contested between two classes. The winner of this first round then contests the feature with the third class. The second round winner owns the feature.

We must choose three thresholds for the three discriminant functions. Fig. 4 contains the projections for each pair of classes.
Figure 4 – Projections of features onto each of the three discriminants. The threshold between projected means is shown.

The confusion matrix for the Fischer discriminant classifier is shown below.

<table>
<thead>
<tr>
<th></th>
<th>Setosa</th>
<th>Versicolor</th>
<th>Verginica</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True class</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Setosa</td>
<td>48</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Versicolor</td>
<td>2</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>Verginica</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

In this case, the Fischer discriminant made 1 fewer error than the Bayes classifier, for an error rate of **2.67%**. This demonstrates that the Gaussian model of the feature distributions is not ideal.

Fig. 5 depicts the decision boundary for classification with Fischer’s linear discriminant. These plots are analogous to Fig. 3 for the Bayes classifier.
Conclusions

Fischer’s linear discriminant classifier (operating pairwise between classes) outperformed the Bayes classifier with a Gaussian likelihood, though only by a single sample. The linear classifier also outperformed the Bayes error. This shows that while the Gaussian assumption is not
perfect, the main issue is the lack of separation between the *Setosa* and *Versicolor* features. This causes a lot of overlap of the 4-dimensional likelihood functions, leading to the relatively high Bayes error. From the confusion matrix, we see that *Verginica* was perfectly separated by both methods.