- t test
- F test
- Z test,

2. Use a random data and evaluate the scores using your own code.

3. Run the LDA code on fishiries data set.

Zunera Zahid

Contents

Test Statistics Null & Alternate Hypothesis

Null Hypothesis:
- There is no significant difference in a set of observations: "Two sample means are equal"

Alternate Hypothesis:
- "Two sample means are not equal"

1. Test statistic is created to choose if we are rejecting the null hypothesis or not.
2. This null hypothesis is going to fail fail when the test statistic lies inside a particular region of the test statistic.

Test statistic	Associated test	Sample size	Information given	Distribution	Test question
z-score	z-test	Two populations or large samples (n > 30)	• Standard deviation of the population (this will be given as σ) • Population mean or proportion	Normal	Do these two populations differ?
t-statistic	t-test	Two small samples (n < 30)	• Standard deviation of the sample (this will be given as *s*) • Sample mean	Normal	Do these two samples differ?
f-statistic	ANOVA	Three or more samples	• Group sizes • Group means • Group standard deviations	Normal	Do any of these three or more samples differ from each other?

Z-test

Null Hypothesis:
- Sample mean is same as the population mean

Alternate:
- Sample mean is not the same as the population mean

1. A z-test is used when:
   a) Your sample size is greater than 30. Otherwise, use a t test.
   b) Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.
   c) Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.
   d) Your data should be randomly selected from a population, where each item has an equal chance of being selected.
   e) Equal sample sizes:Sample sizes should be equal if at all possible.
2. In a z-test, we need to compare two given sample means.
3. To calculate a z-score we need the population **mean** and the population **standard deviation**.
4. The sample follows a Gaussian distribution.
5. A one sided z-test will have one critical boundary, while a two sided z-test will have two critical boundaries.
6. A z-test is used when the population parameters like standard deviation are known.
7. Using the below formula we can calculate the z-statistic:

$$z = (x - \mu) / (\sigma / \sqrt{n})$$

Where:
x= sample mean
σ / √n = standard deviation of population

8. A one-sample z-test allows for us to see if a particular piece/group of data is actually from a larger population of data.
9. If the p-value is lower than 0.05, reject the hypothesis or else accept the null hypothesis.
10. If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis
11. Example:
    a. Example:Comparing the average engineering salaries of men versus women.
    b. Example: Comparing the fraction defectives from 2 production lines.

T-test

1. Situations to use a T-Test:
    a) Data is independent
    b) Approximately normally distributed data
    c) Similar amount of variance within the groups being compared (homogenous)
2. A t-test is a type of inferential statistic which is used to determine if there is a significant difference between the means of two groups which may be related in certain features.
3. It is mostly used when the data sets would follow a normal distribution and may have unknown variances.
4. T test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.
5. A t-test is used to see if two separate samples have the same mean value. However, T-Tests differ from z-tests because we use them when the mean and standard deviations of the population are not known.
6. The T-test is used to compare the mean of two given groups.
7. The sample follows the Gaussian distribution.
8. A t-test is used when parameters like the standard deviation of the population are not known.
9. We can calculate the t statistics by the given formula

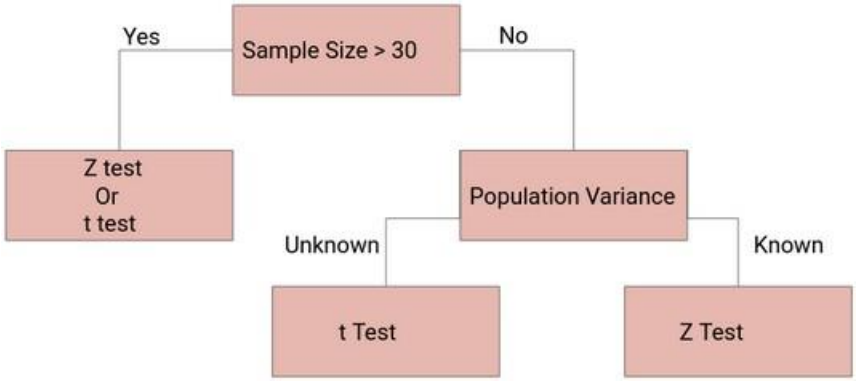
    t = (x1 — x2) / (σ / √n1 + σ / √n2)

    Where
    x1 = sample 1 mean
    x2 = sample 2 mean
    n1 = sample 1 size
    n2 = sample 2 size

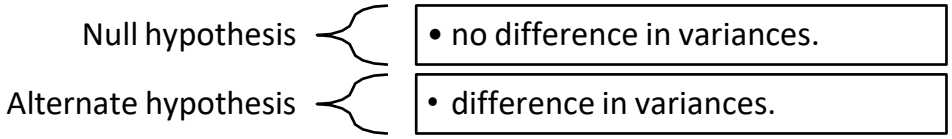
10. There are three T-Test types:
    a) **Paired T-Test:** If the groups come from a single population
    b) **Two Sample T-Test (Independent):** If the groups come from different populations
    c) **One Sample T-Test:** If there is one group being compared against a standard value
11. Example :Measuring the average diameter of shafts from a certain machine when you have a small sample.

Deciding between Z Test and T Test

1. If the sample size is large enough, then the Z test and t-Test will conclude with the same results. For a large sample size, Sample Variance will be a better estimate of Population variance so even if population variance is unknown, we can use the Z test using sample variance.
2. Similarly, for a Large Sample, we have a high degree of freedom. And since t-distribution approaches the normal distribution, the difference between the z score and t score is negligible.



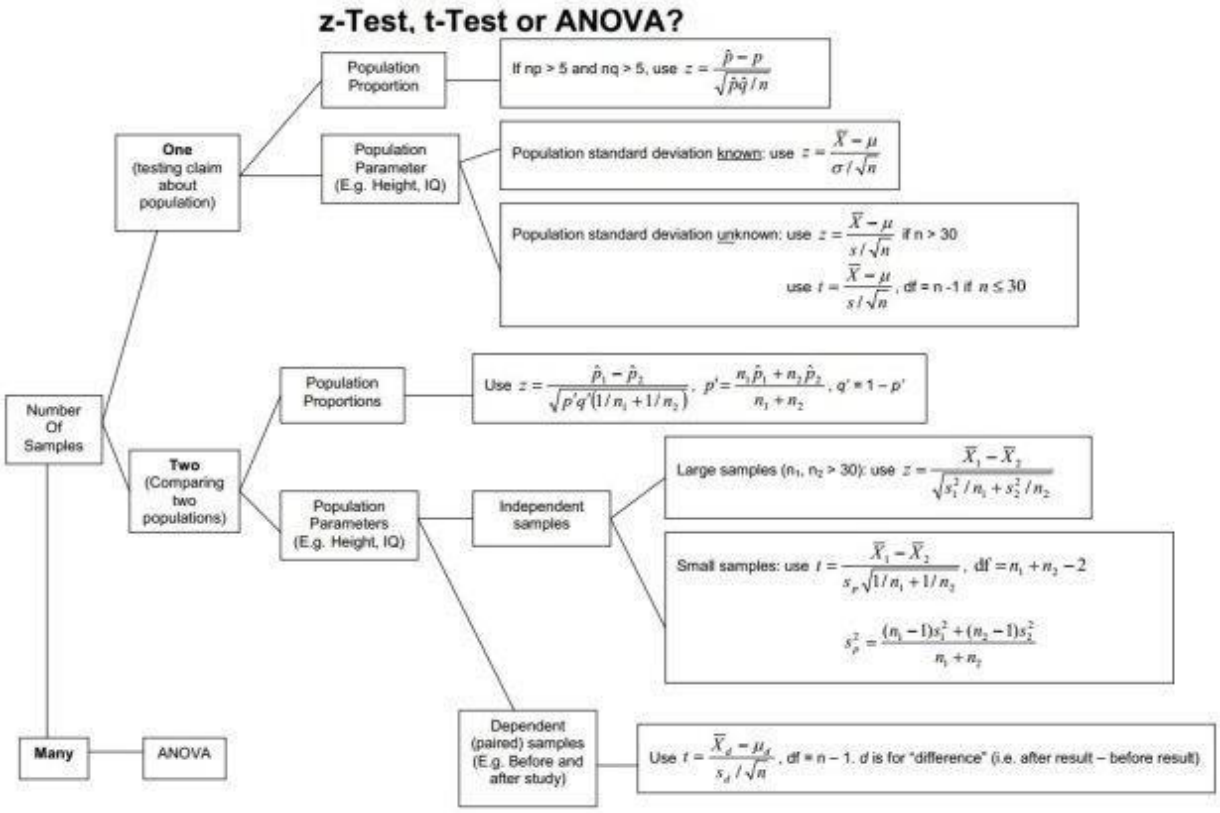
F-test



1. F-Test is the most often used when comparing statistical models that have been fitted to a data set to identify the model that best fits the population.
2. Researchers usually use it when they want to test whether two independent samples have been drawn from a normal population with the same variability.
3. The t-test works well when dealing with two groups, but sometimes we want to compare more than two groups at the same time. For example, if we wanted to test whether voter age differs based on some categorical variable like race, we have to compare the means of each level or group the variable. We could carry out a separate t-test for each pair of groups, but when you conduct many tests you increase the chances of false positives.
4. The analysis of variance or ANOVA is a statistical inference test that lets you compare multiple groups at the same time.

   F = Between group variability / Within group variability
5. Unlike the z and t-distributions, the F-distribution does not have any negative values because between and within-group variability are always positive due to squaring each deviation.
6. One Way F-test(Anova) :- It tell whether two or more groups are similar or not based on their mean similarity and f-score.
7. Two Way F-test :- Two way F-test is extension of 1-way f-test, it is used when we have 2 independent variable and 2+ groups. 2-way F-test does not tell which variable is dominant. if we need to check individual significance then Post-hoc testing need to be performed.
8. Example: Comparing the variability of bolt diameters from two machines

Deciding between Z Test T Test F Test



Evaluation Summary

- Z Test
  - used for testing the mean of a population versus a standard, or comparing the means of two populations, with large (n ≥ 30) samples whether you know the population standard deviation or not.
  - It is also used for testing the proportion of some characteristic versus a standard proportion, or comparing the proportions of two populations.
- T-test
  - A t-test is used for testing the mean of one population against a standard or comparing the means of two populations if you do not know the populations' standard deviation and when you have a limited sample (n < 30).
  - If you know the populations' standard deviation, you may use a z-test.
- F-test
  - An F-test is used to compare 2 populations' variances. The samples can be any size.
  - It is the basis of ANOVA.

Code that uses random Data to calculate T test values

While writing the code for t test, the code was written from scratch using these steps

1. function to calculate t test for independent data is defined
2. function to calculate t test for dependent data is defined
3. First we calculate mean_value
4. Then we calculate standard error
5. Then we calculate t statistical values
6. Then we calculate degree of freedom
7. Then we calculate critical values
8. Then we calculate p_value
9. Open google collab and run the code

```
# importing required libraries

from numpy.random import seed
from math import sqrt
from numpy.random import randn
from scipy.stats  import  t
from scipy.stats import sem
from numpy import mean
# function to caluclate t test for independent values
#first we calculate mean_value,then standard error,then t statistical
#then degree of freedom,then critical value in the end p_value is computed
def independent_samples_t_test(sample1, sample2, alpha_values):
  mean_value1, mean_value2 = mean(sample1), mean(sample2)
  standard_err1, standard_err2 = sem(sample1), sem(sample2)
  standard_err = sqrt(standard_err1**2.0 + standard_err2**2.0)
  t_stat = (mean_value1 - mean_value2) / standard_err
  degrees_of_freedom = len(sample1) + len(sample2) - 2
  critical_value = t.ppf(1.0 - alpha_values, degrees_of_freedom)
  p_value = (1.0 - t.cdf(abs(t_stat), degrees_of_freedom)) * 2.0
  return t_stat, degrees_of_freedom, critical_value, p_value

# function to caluclate t test for dependent values
#first we calculate mean_value,then standard error,then t statistical
#then degree of freedom,then critical value in the end p_value is computed
def dependent_samples_t_test(sample1, sample2, alpha_values):
  mean1, mean2 = mean(sample1), mean(sample2)
  n = len(sample1)
  d1 = sum([(sample1[i]-sample2[i])**2 for i in range(n)])
  d2 = sum([sample1[i]-sample2[i] for i in range(n)])
  sd = sqrt((d1 - (d2**2 / n)) / (n - 1))
  standard_err = sd / sqrt(n)
  t_stat = (mean1 - mean2) / standard_err
  degrees_of_freedom = n - 1
  critical_value = t.ppf(1.0 - alpha_values, degrees_of_freedom)
  p_value = (1.0 - t.cdf(abs(t_stat), degrees_of_freedom)) * 2.0
  return t_stat, degrees_of_freedom, critical_value, p_value
```

```
# Data Generation
seed(8)
sample1 = 45 * randn(99) + 71
sample2 = 45 * randn(99) + 72
# calculate the t test for independent samples
alpha_values = 0.05
t_stat, degrees_of_freedom, critical_value, p_value = independent_samples_
t_test(sample1, sample2, alpha_values)
print('calculate the t test for independent samples t=%.3f, degrees_of_fre
edom=%d, critical_value=%.3f, p_value=%.3f' % (t_stat, degrees_of_freedom,
 critical_value, p_value))
# using critical_values as parameter
if abs(t_stat) <= critical_value:
  print('Null hypothesis (means are equal) Accepted')
else:
  print('Null hypothesis (means are equal) Rejected')
# using p values as parameter
if p_value > alpha_values:
  print('Null hypothesis (means are equal) Accepted')
else:
  print('Null hypothesis (means are equal) Rejected')

# applying t test for dependent data
alpha_values = 0.05
t_stat, degrees_of_freedom, critical_value, p_value = dependent_samples_t_
test(sample1, sample2, alpha_values)
print('calculate the t test for dependent samples t=%.3f, degrees_of_freed
om=%d, critical_value=%.3f, p_value=%.3f' % (t_stat, degrees_of_freedom, c
ritical_value, p_value))
#  interpret  via  critical  value
if abs(t_stat) <= critical_value:
  print('Null hypothesis (means are equal) Accepted')
else:
  print('Null hypothesis (means are equal) Rejected')
# with the help of p values
if p_value > alpha_values:
  print('Null hypothesis (means are equal) Accepted')
else:
  print('Null hypothesis (means are equal) Rejected')
```

Code that uses random Data to calculate Z test values

While writing the code for Z test, the in-built python functions and libraries were used
Open google collab and run the code

```
import  math
import numpy as np
from numpy.random import randn
from statsmodels.stats.weightstats import ztest
from numpy.random import seed
from math import sqrt
from numpy.random import randn
from scipy.stats  import  t
from scipy.stats import sem
from numpy import mean


# seed the random number generator
seed(8)
# generate data
sample1 = 45 * randn(99) + 71
mean_value1 = mean(sample1)
alpha_value =0.05
null_mean =100

# print mean and sd
ztest_Score, p_value= ztest(data,value = null_mean, alternative='larg
er')
print('Using z test value of mean=%.2f stdv=%.2f ztest_Score=%.2f' %
(np.mean(sample1), np.std(sample2), ztest_Score))

if(p_value < alpha_value):
  print("Null Hypothesis Rejected")
else:
  print("Rejecting NUll Hypothesis Failed")
```

Code that uses random Data to calculate F test values

While writing the code for f test, the code was written using these steps

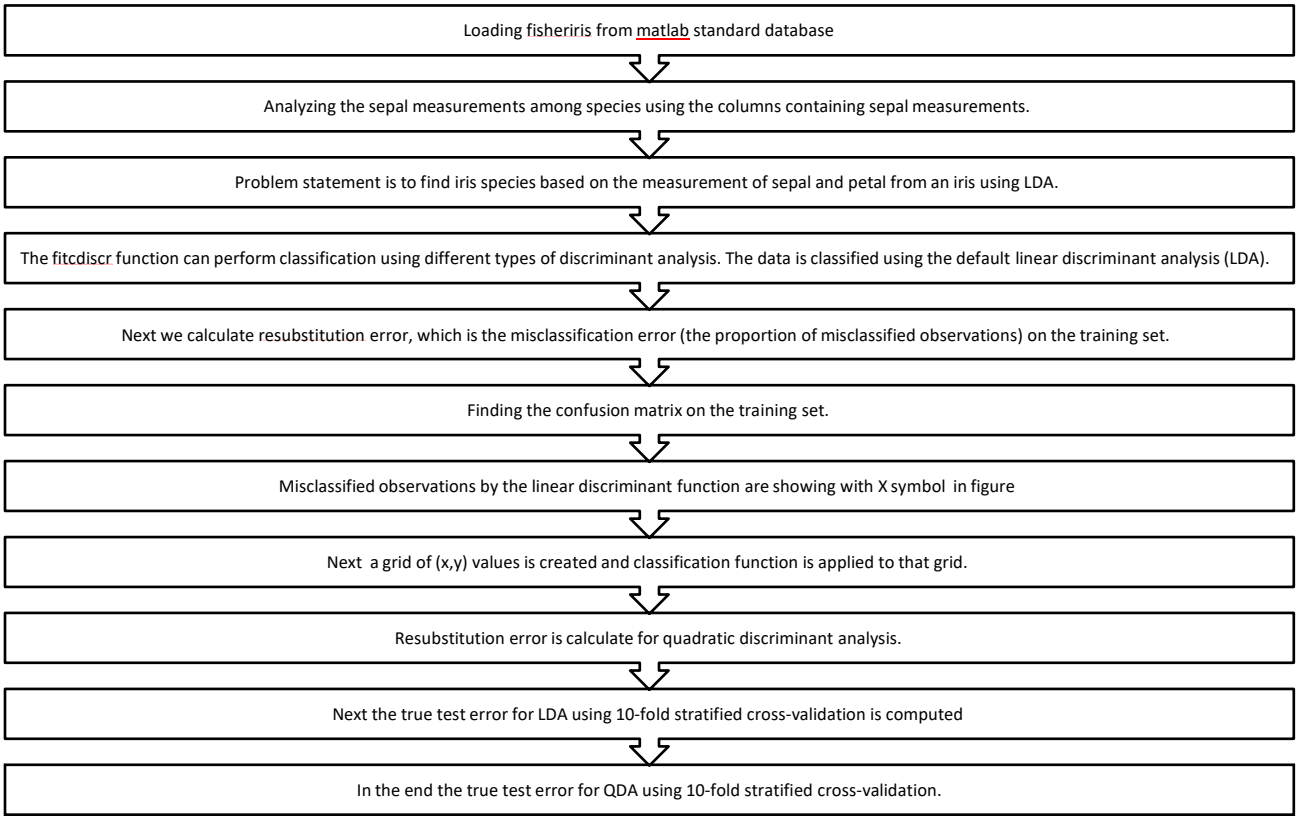
1. Function for calculating F test values is defined
2. Then I calculated Statistical f test value
3. Then numerator of degrees of freedom is declared
4. Then denominator of degrees of freedom is declared
5. Then p_value is computed

```
from numpy.random import seed
from numpy.random import randn
from scipy.stats  import  t
from scipy.stats import sem
from numpy import mean
import numpy as np
import scipy
import  math
import numpy as np
from numpy.random import randn
from statsmodels.stats.weightstats import ztest
from numpy.random import seed

#Function for calculating F test values
#first I calculated Statistical f test value
# then numerator of degrees of freedom is declared
# then denominator of degrees of freedom is declared
#p_value is computed

def f_test(sample1, sample2):
    x = np.array(sample1)
    y = np.array(sample2)
    f_value = np.var(x, ddof=1)/np.var(y, ddof=1)
    degree_of_freedom_numerator = x.size-1
    degree_of_freedom_denominator = y.size-1
    p_value = 1-
scipy.stats.f.cdf(f_value, degree_of_freedom_numerator, degree_of_freedom_
denominator)
    return f_value, p_value
# Data generation
seed(8)
# generate two independent samples
sample1 = 45 * randn(99) + 71
sample2 = 45 * randn(99) + 72
#invoking ftest
f,p=f_test(sample1, sample2)
print('calculations using F test the f test f=%.3f, p=%.3f' % (f, p))
```

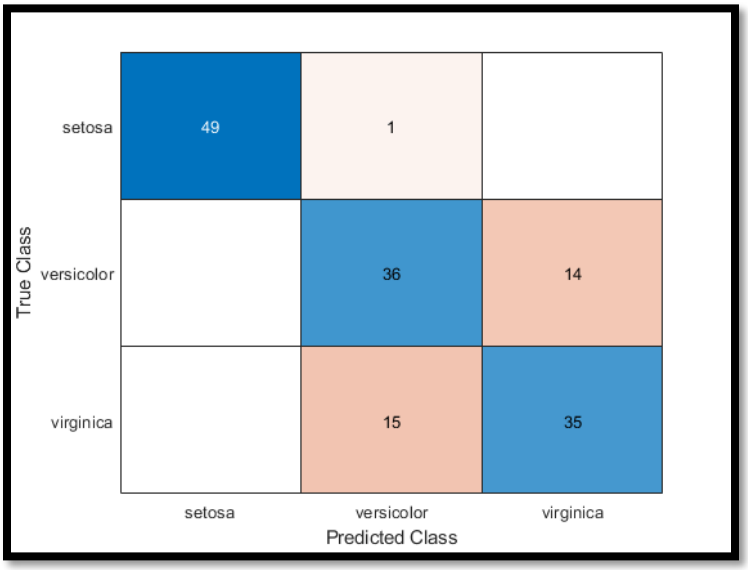
Run LDA code on fisheriris data set

Linear and Quadratic Discriminant Analysis Steps:

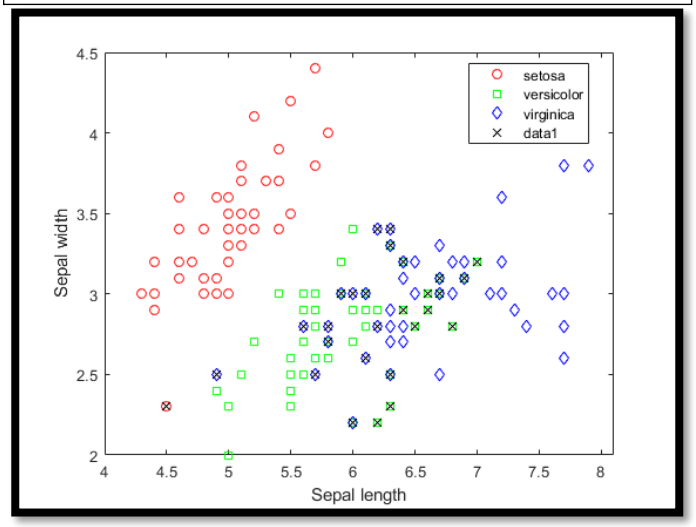
1. Loading `fisheriris` from matlab standard database
2. Analyzing the sepal measurements among species using the columns containing sepal measurements.
3. Problem statement is to find iris species based on the measurement of sepal and petal from an iris using LDA.
4. The `fitcdiscr` function can perform classification using different types of discriminant analysis.
5. The data is classified using the default linear discriminant analysis (LDA).
6. Next we calculate resubstitution error, which is the misclassification error (the proportion of misclassified observations) on the training set.
7. Finding the confusion matrix on the training set.
8. Misclassified observations by the linear discriminant function are showing with X symbol in figure
9. Next a grid of (x,y) values is created and classification function is applied to that grid.
10. Resubstitution error is calculate for quadratic discriminant analysis.
11. Next the true test error for LDA using 10-fold stratified cross-validation is computed
12. In the end the true test error for QDA using 10-fold stratified cross-validation.



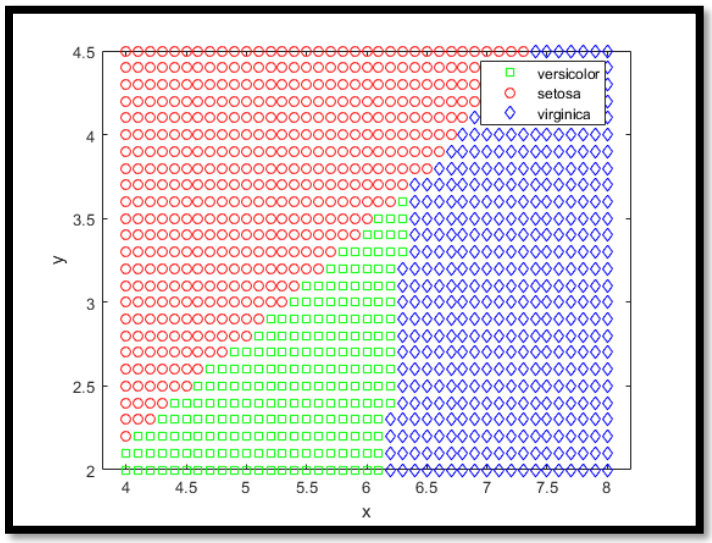
```
load fisheriris
f = figure;
gscatter(meas(:,1), meas(:,2), species,'rgb','osd');
xlabel('Sepal length');
ylabel('Sepal width');
N = size(meas,1);


lda = fitcdiscr(meas(:,1:2),species);
ldaClass = resubPredict(lda);
ldaResubErr = resubLoss(lda)
figure
ldaResubCM = confusionchart(species,ldaClass);
```



```
figure(f)
bad = ~strcmp(ldaClass,species);
hold on;
plot(meas(bad,1), meas(bad,2), 'kx');
hold off;
```




```
[x,y] = meshgrid(4:.1:8,2:.1:4.5);
x = x(:);
y = y(:);
j = classify([x y],meas(:,1:2),species);
gscatter(x,y,j,'grb','sod')
```



```
qda = fitcdiscr(meas(:,1:2),species,'DiscrimType','quadratic');
qdaResubErr = resubLoss(qda)
rng(0,'twister');
cp = cvpartition(species,'KFold',10)
cvlda = crossval(lda,'CVPartition',cp);
ldaCVErr = kfoldLoss(cvlda)
cvqda = crossval(qda,'CVPartition',cp);
qdaCVErr = kfoldLoss(cvqda)
```

Evaluation of LDA for Fisheriris data:

Summary of LDA for Fisheriris data:

1. The LDA cross-validation error has the same value as the LDA resubstitution error on this data.
2. QDA has a slightly larger cross-validation error than LDA.
3. Simple models may get comparable, or better performance than a more complicated model.
4. For some data sets, the regions for the various classes are not well separated by lines. When that is the case, linear discriminant analysis is not appropriate. Instead, quadratic discriminant analysis (QDA) can be used for such data.