Creating Classification Rules to Distinguish Between Cherry Tree and Pear Tree Leaves

Presented to Dr. Steven Vamosi

Professor of Ecology and Evolutionary Biology

University of Calgary

Prepared by Jana Osea

Undergraduate Student in Statistics

University of Calgary

March 14, 2021

Contents

Summary

Introduction

In many fields of study, classifying items or individuals as belonging to one of two or more population or groups is an integral part of analysis. In many cases, it is often the goal of a research or study to classify each sample item correctly. In fields like finance and medical research, correctly classifying items can imply stopping transaction fraud or discovering deadly tumors. Hence, it is important to understand how classification procedures work.

Background

As the importance of classification is becoming more evident, Professor Steven Vamosi, a Doctor in Ecology and Evolutionary Biology has entrusted an undergraduate student in Statistics at the University of Calgary, Jana Osea, with the task to develop a classification rule to distinguish between cherry and pear tree leaves. This allows Dr. Vamosi a simple method to classify between the two species and for us demonstrate our knowledge of classification.

Goal

Using width and length measurements taken from cherry and pear tree leaves, our *goal* is to create a classification method to distinguish between the two species.

Data Generation Process

Data Source

- Cherry leaves: Figure 7.2 of the pdf file printed on a standard A4 paper
- Pear leaves: Figure 7.3 of the pdf file printed on a standard A4 paper

Data Input

In an Microsoft Excel Sheet (2020), I prepared 3 empty columns with the following headers: species, width, and length.

For each leaf a new row with 3 columns is recorded in the excel sheet that contains the species, width, and length measured according to the procedure outlined below. After recording each value, I saved the data as a csv file named "data.csv". In addition, the full raw data can be found in A1 in the appendix.

- species: If the leaf is part of figure 7.2, then species contains string input "cherry." If the leaf is part of figure 7.3, then species contains string input "pear."
- width: Measured the widest part of each leaf using a straight ruler to the closest millimeter
- length: Measured from the bottom tip to the top tip using a straight ruler of each leaf to the closest millimeter

Methods

Overview of Methods

After inputting the entire data set, I imported the csv file into my program. I made 2 classifications: (1) with equal variance assumption and (2) with no equal variance assumption. Densities and lambda values were calculated for each leaf and visualizations of classifcations were made. 3 new leaf measurements were provided and classified according to the first classification. In addition, misclassifications of each method were recorded.

Software and Packages

I used R version 4.0.3 (2020-10-10) (R Core Team (2020)) to perform all my classification programming. I also used the following R package to help me visualize and aid my density calculations

- ggplot2 (H. Wickham (2016))
- gridarrange (Baptiste Auguie (2017))
- mtvnorm (Alan Genz, et. al (2020))

First Classification

Assumptions

The first classification rule assumes the following:

1. For each species k = cherry or pear, the distribution of the width and length measurements follow a bivariate normal distribution as follows

   $$\begin{pmatrix} X_k \\ Y_k \end{pmatrix} \sim N_2\left(\mu_k, \Sigma\right)$$

   where

   $$\begin{aligned} X_k &= \text{ width (mm) of the } k \text{ species} \\ Y_k &= \text{ length (mm) of the } k \text{ species} \\ \mu_k &= \begin{pmatrix} \mu_{kx} \\ \mu_{ky} \end{pmatrix} \text{ of the } k \text{ species} \\ \Sigma &= \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \end{aligned}$$

   Hence, the density of a leaf given the x width and y length according to the k = cherry or pear species is given by

   $$f_k(x, y \mid \mu_k, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left[\frac{-1}{2} \begin{pmatrix} x - \mu_{kx} \\ y - \mu_{ky} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x - \mu_{kx} \\ y - \mu_{ky} \end{pmatrix}\right]$$

   where

   $$|\Sigma| = \text{ determinant of the covariance matrix.}$$

2. The covariance matrix Σ of the k = cherry or pear species is the same with possible differences in the mean vectors μ_k.

Parameter Estimation

Given the data collected, for $k =$ cherry or pear, we estimate the unknown parameters μ_k and Σ as $\hat{\mu}_k$ and $\hat{\Sigma}$ by using the method of moments as follows.

$$\hat{\mu}_k = \begin{pmatrix} \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki} \\ \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki} \end{pmatrix} \quad \text{and} \quad \hat{\Sigma} = \frac{1}{2}\left(\hat{\Sigma}_{\text{cherry}} + \hat{\Sigma}_{\text{pear}}\right)$$

where

$$\begin{aligned} x_{ki} &= \text{ width (mm) of the } i\text{-th leaf for the } k\text{-th species} \\ y_{ki} &= \text{ length (mm) of the } i\text{-th leaf for the } k\text{-th species} \\ n_k &= \text{ number of total leaves gathered for the } k\text{-th species} \\ \hat{\Sigma}_k &= \begin{pmatrix} \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}^2 - \left(\frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}\right)^2 & \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}y_{ki} - \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}\frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki} \\ \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}y_{ki} - \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}\frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki} & \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki}^2 - \left(\frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki}\right)^2 \end{pmatrix} \end{aligned}$$

Using the formulas above and the sample data, the estimated mean and covariance matrix are shown below.

$$\hat{\mu}_{\text{cherry}} = \begin{pmatrix} 36.44 \\ 79.56 \end{pmatrix}$$

$$\hat{\mu}_{\text{pear}} = \begin{pmatrix} 41.19 \\ 70 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}$$

Hence using the estimated parameters, the density formula for a given x width and y length for the $k =$ cherry or pear species is shown below.

$$f_{\text{cherry}}(x, y \mid \hat{\mu}_{\text{cherry}}, \hat{\Sigma}) = \frac{1}{2\pi\sqrt{\left|\begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}\right|}} \exp\left[\frac{-1}{2}\begin{pmatrix} x - 36.44 \\ y - 79.56 \end{pmatrix}^T \begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}^{-1} \begin{pmatrix} x - 36.44 \\ y - 79.56 \end{pmatrix}\right]$$

$$f_{\text{pear}}(x, y \mid \hat{\mu}_{\text{pear}}, \hat{\Sigma}) = \frac{1}{2\pi\sqrt{\left|\begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}\right|}} \exp\left[\frac{-1}{2}\begin{pmatrix} x - 41.19 \\ y - 70 \end{pmatrix}^T \begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}^{-1} \begin{pmatrix} x - 41.19 \\ y - 70 \end{pmatrix}\right]$$

Classification Rule

Let n be the total number of observations in the training data, i be any integer from 0 to n, and (x_i, y_i) be the i-th observation where x_i is the width measurement and y_i is the length measurement. The respective λ values for each observation (x_i, y_i) is as follows.

$$\lambda_i = \frac{f_{\text{cherry}}(x_i, y_i | \hat{\mu}_{\text{cherry}}, \hat{\Sigma})}{f_{\text{pear}}(x_i, y_i | \hat{\mu}_{\text{pear}}, \hat{\Sigma})}$$

$$= \frac{\frac{1}{2\pi\sqrt{\left|\begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}\right|}} \exp\left[\frac{-1}{2}\begin{pmatrix} x - 36.44 \\ y - 79.56 \end{pmatrix}^T \begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}^{-1} \begin{pmatrix} x - 36.44 \\ y - 79.56 \end{pmatrix}\right]}{\frac{1}{2\pi\sqrt{\left|\begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}\right|}} \exp\left[\frac{-1}{2}\begin{pmatrix} x - 41.19 \\ y - 70 \end{pmatrix}^T \begin{pmatrix} 22.51 & 31.28 \\ 31.28 & 120.5 \end{pmatrix}^{-1} \begin{pmatrix} x - 41.19 \\ y - 70 \end{pmatrix}\right]}$$

The following classification rule described below is used to determine whether observation (x_i, y_i) belongs to a certain species. The lambda values and classifications of all the observations is recorded in A2 in the appendix.

- if $\lambda_i > 1$, then observation (x_i, y_i) is a cherry leaf
- if $\lambda_i < 1$, then observation (x_i, y_i) is a pear leaf
- if $\lambda_i = 1$, then observation (x_i, y_i) is undetermined

Classification Errors

There is a total of 4 misclassifications. This is evident in table 1 as well as the rey circles and orange triangles in figure 2.

Table 1. Observation Points Where Misclassification Occurs

	species	width (mm)	length (mm)	classification
5	cherry	37	67	pear
7	cherry	47	88	pear
25	pear	40	90	cherry
27	pear	32	63	cherry

Figure 1. Scatter Plot of Observation Data

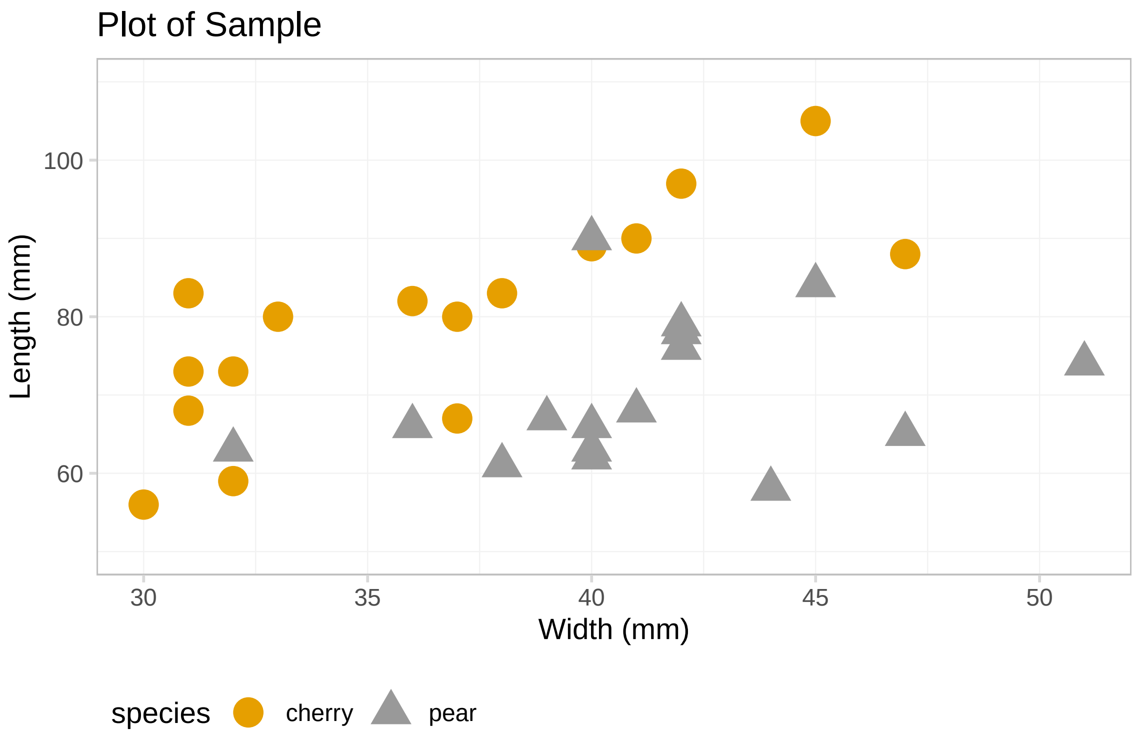
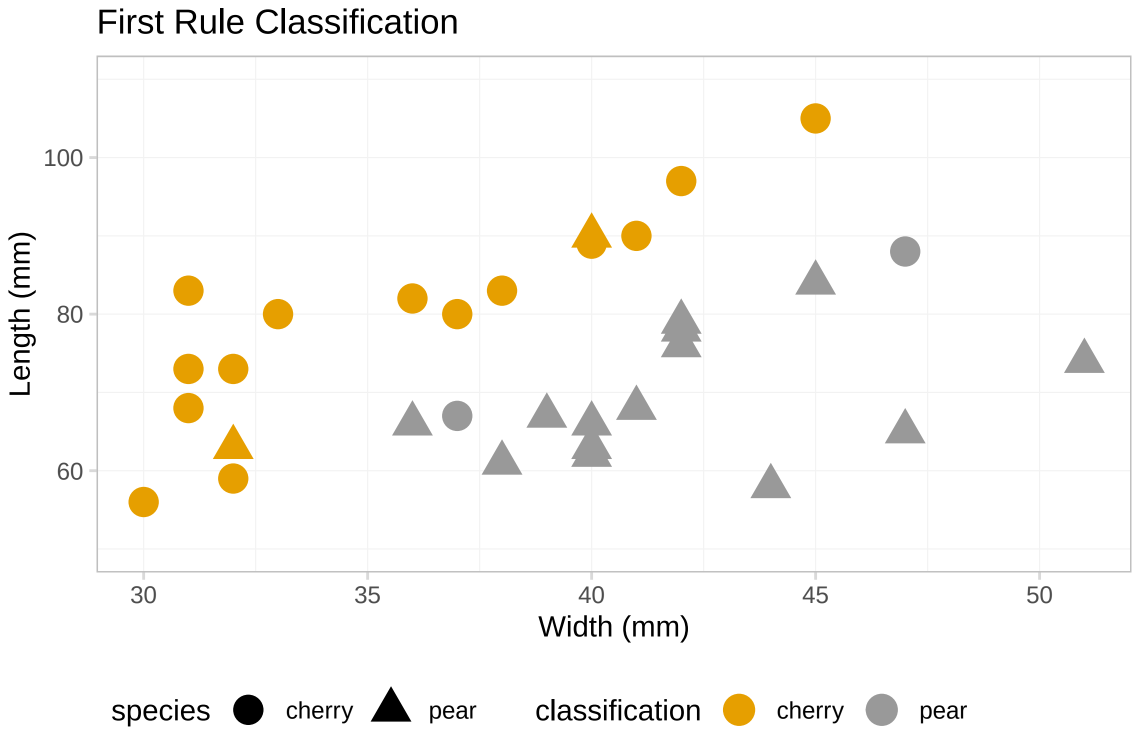


Figure 2. Scatter Plot of First Classification Rule



New Classification

Using the classification rules outlined above, we are tasked to classify new leaves with the following meassurements

id	width (mm)	length (mm)
u	32	82
v	38	52
w	52	76

We calculate λ using the formula stated above and get the following resuts

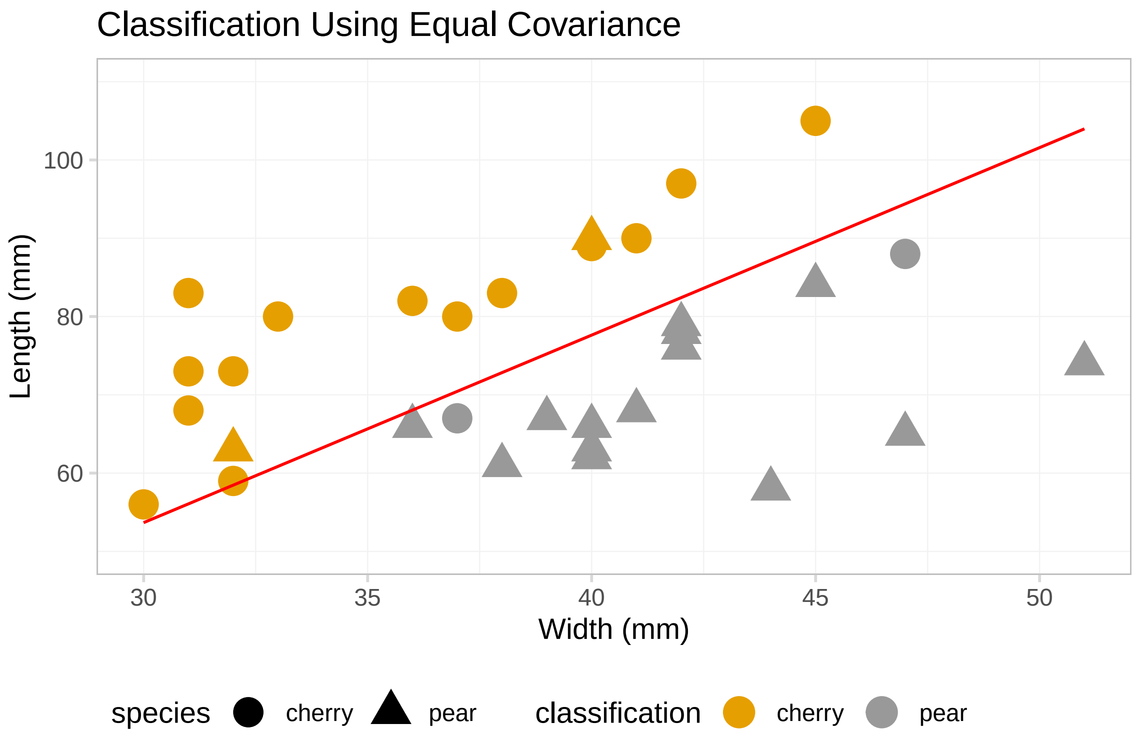
id	lamda	classification
u	139.61	cherry
v	0.01	pear
w	0.71	pear

Decision Boundary

Using the derivation of the decision boundary found in A4 in the appendix, the decision boundary for the given in the formula below.

$$y = 2.4x - 18.18$$

Figure 3. Scatter Plot with Decision Boundary



Furthermore, I predicted all the possible points in the observation space to the closest 0.1 mm using the first classification rules and got the following results in the plot below. From figure 4, it is clear that the decision boundary is a linear equation that divides the observation space into 2 distinct regions.

Figure 4. Grid Plot of a Predicted Points in the Observation Space

Second Classification

Assumptions

The assumptions are similar to the first classification. However, the only difference is that the covariance matrices are no longer equal. Instead, for each species $k =$ cherry or pear, the distribution of the width and length measurements follow a bivariate normal as follows

$$\begin{pmatrix} X_k \\ Y_k \end{pmatrix} \sim N_2\left(\mu_k, \Sigma_k\right)$$

where

$$\Sigma_k = \begin{pmatrix} \sigma_{kx}^2 & \sigma_{kxy} \\ \sigma_{kxy} & \sigma_{ky}^2 \end{pmatrix} \text{ covariance matrix of the } k \text{ species}$$

Hence, the density of a leaf given the x width and y length according to the $k =$ cherry or pear species is given by

$$f_k(x, y \mid \mu_k, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma_k|}} \exp\left[\frac{-1}{2}\begin{pmatrix} x - \mu_{kx} \\ y - \mu_{ky} \end{pmatrix}^T \Sigma_k^{-1} \begin{pmatrix} x - \mu_{kx} \\ y - \mu_{ky} \end{pmatrix}\right]$$

where

$$|\Sigma_k| = \text{ determinant of the } k \text{ species covariance matrix.}$$

Parameter Estimation

For $k =$ cherry or pear, $\hat{\mu}_k$ is the same as derived in the first classification. $\hat{\Sigma}_k$ is no longer the pooled covariance matrix. For each $k =$ cherry or pear the estimation of $\hat{\Sigma}_k$ is the method of moments covariance matrix. The derivation of this can be found in the first classification parameter estimation.

$$\hat{\Sigma}_{\text{cherry}} = \begin{pmatrix} 27.12 & 52.07 \\ 52.07 & 162.87 \end{pmatrix}$$

$$\hat{\Sigma}_{\text{pear}} = \begin{pmatrix} 17.9 & 10.5 \\ 10.5 & 78.12 \end{pmatrix}$$

Hence using the estimated parameters, the density formula for a given x width and y length for the $k =$ cherry or pear species is shown below.