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 dataset can be found in Appendix A.

- 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 leaves measurements follow a bivariate normal distribution as follows

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

   where

$$\begin{aligned} X = &\ \text{width (mm)} \\ Y = &\ \text{length (mm)} \\ \mu_k = &\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \text{where } k = \text{cherry or pear} \\ \Sigma = &\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \end{aligned}$$

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

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

   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

Second Classification

Conclusion

References

- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org./
- H. Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Torsten Hothorn (2020). mvtnorm: Multivariate Normal and t Distributions. R package version 1.1-1. URL http://CRAN.R-project.org/package=mvtnorm
- Alan Genz, Frank Bretz (2009), Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195., Springer-Verlag, Heidelberg. ISBN 978-3-642-01688-2

Appendix A

Full Data Table

```
##    species width length
## 1   cherry    30     56
## 2   cherry    32     59
## 3   cherry    31     68
## 4   cherry    41     90
## 5   cherry    37     67
## 6   cherry    45    105
## 7   cherry    47     88
## 8   cherry    31     73
## 9   cherry    32     73
## 10  cherry    33     80
## 11  cherry    36     82
## 12  cherry    40     89
## 13  cherry    42     97
## 14  cherry    37     80
## 15  cherry    31     83
## 16  cherry    38     83
## 17    pear    44     58
## 18    pear    41     68
## 19    pear    38     61
## 20    pear    40     62
## 21    pear    42     78
## 22    pear    40     66
## 23    pear    40     63
## 24    pear    47     65
## 25    pear    40     90
## 26    pear    42     76
## 27    pear    32     63
## 28    pear    45     84
## 29    pear    42     79
## 30    pear    36     66
## 31    pear    39     67
## 32    pear    51     74
```