BIOS512_hw5

September 18, 2025

1 Homework 5

This homework requires `wine.csv`, and the `tidyverse` and `Rtsne` packages. Install them if you haven't already!
See the following link for how to add new packages to Binder: https://github.com/rjenki/BIOS512?tab=readme-ov-file#adding-packages-to-installr-later.
For readability and easier processing, please make each question part a different code chunk.

[1]:
```
library(tidyverse)
library(Rtsne)
```

```
  Attaching core tidyverse packages                tidyverse
2.0.0
  dplyr     1.1.2       readr     2.1.4
  forcats   1.0.0       stringr   1.5.0
  ggplot2   3.4.2       tibble    3.2.1
  lubridate 1.9.2       tidyr     1.3.0
  purrr     1.0.1
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
  Use the conflicted package
(<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

1.1 Question 1

a) Import your data.

b) Check out the columns present using one of R's data frame summary.

c) Get summary statistics on the numeric variables.

1.2 Question 2

a) Scale and center your data *Hint:* Use a `mutate()` statement across all columns **except class** with `function(x) as.numeric(scale(x))`.

b) Based on what you saw in the summary statistic table from the imported data, why would scaling and centering this data be helpful before we perform PCA?

1.3 Question 3

a) Perform PCA

b) How much of the total variance is explained by PC1? PC2? What function do we use to see that information?

c) Why are we doing PCA first?

d) What is the rotation matrix? Print it explicitly. *Hint:* Check the notes for a simple way to do this!

e) Plot PC1 vs. PC2, using the wine class as labels for coloring. *Hint:* You'll first need a data set with only PC1 and PC2, then add back the class variable from your scaled data set with a `mutate()` statement. Then, you can use `color = factor(class)` in your `ggplot` statement.

f) What do you see after plotting PC1 vs. PC2? What does this mean in context of wine classes?

g) Give an example of data where PCA would fail. You can describe the data or do a simulation. *Hint:* Our notes have a few examples!

h) Explain the difference between vector space and manifold, and how these terms apply to what we did/will do with T-SNE.

1.4 Question 4

a) Perform T-SNE Set `seed = 123`.
Hint: Subset your PCA results to PC1–PC10, add the class variable back in, remove duplicates, then perform T-SNE.

b) Plot the results in 2D *Hint:* Convert your T-SNE results to a tibble and add back the class variable from your scaled data set using a `mutate()` statement. Then, you can use `color = factor(class)` in your `ggplot` statement.

c) Why didn't we stop at PCA?

d) What other types of data does this workflow make sense for?