CAMB 698 Final Paper

Vincent Wu

2018-12-09

Contents

Chapter 1

Prerequesites

This book will feature code and graphs produced using the R statistical language. Below are packages that will be used throughout this book.

```
library(tidyverse)
library(usedist)
library(qiimer)
library(reshape2)
library(ggplot2)

library(ape)
library(Rtsne)
```

The below code will load in the data that will be used in the next sections.

```
load("data/poop_across_penn1.Rdata")
```

```
# Create vendor/mice dataframe
s_vendor_all <- s %>%
  filter(grepl("ARC Vendor Experiment", Experiment)) %>%
  rename(Vendor = Mouse_Source_Vendor) %>%
  mutate(SubjectID = factor(paste("Mouse", Mouse_Number))) %>%
  mutate(SampleType = trimws(as.character(SampleType))) %>%
  arrange(SampleType, Vendor, SubjectID)

# Identify and remove suspicious samples
suspect_SampleIDs <- c("Tac.33.CE.Day1", "Env.13.Stool.Day0")

# Set final dataframe
s_vendor <- s_vendor_all %>%
  droplevels() %>%
  filter(!(SampleID %in% suspect_SampleIDs))
```

Chapter 2

Background

From the initial raw data to visualizing the results, data processing and analysis are fundamental aspects of conducting scientific research. With the recent advances of high throughput technologies (i.e. the many "-omics", multiparametric flow cytometry, etc.), the accompanying data is often high-dimensional and difficult for manual efforts to analyze. The development of machine learning methods aims to help with this problem, thus helping to make sense of the data to draw meaningful and accurate conclusions.

The purpose of this paper is to introduce and explore the different machine learning methods that are commonly used in the biological sciences. To help demonstrate the methods and to maintain the same context across all of the methods, a single dataset (will be referred to as the PAP dataset) was provided by the PennCHOP Microbiome Program (courtesy of Dr. Kyle Bittinger). Mice were purchased from vendors with the purpose of assessing whether the mice have different phenotypes from different vendors. The fecal microbiota was sequenced from the mice, resulting in a distance matrix (how distant each mouse's microbiome was from the other mice sampled) as well as the relative abundance of bacterial species for each mouse. Additionally, metabolites as well as different immune phenotypes were assessed for each mouse. This high-dimensional dataset is representative of datasets that are seen with microbiome research.

Chapter 3

Dimensionality reduction

3.1 Introduction to dimensionality reduction

Dimensionality reduction is a necessary tool for working with high-dimensional data, which can be seen from this dataset. Each of the diversity distances against each of the mice is a single variable, as is each of the relative abundances for different species. A variable is a dimension in this scenario – creating a dataset with hundreds of dimensions. While rich and informative, this dataset would be difficult to visualize beyond two dimensions. Without going too far into the mathematics behind these methods, dimensionality reduction creates a means to observe the data from a different perspective and assists in reducing the computational burden for the machine learning techniques that will be discussed.

3.2 PCA and PCoA

Principal components analysis (PCA) and principal coordinates analysis (PCoA) are common techniques used both in and outside of the biological sciences. PCA makes new variables that are linear combinations of the original variables. PCoA is similar in concept but takes in a distance matrix (such as the one used for our dataset) to transform into new coordinates where the axes of this coordinate system are not correlated with each other. The power of PCA and PCoA is that all new variables have no correlation with each other and can explain all the covariance from the original data. The data points in PCA or PCoA space can be easily visualized as seen in the below graph. It is common to display the variance explained by each of the principal component axes as a means to show how well the principal components can explain the variance in the original data.

3.2.1 PCoA example

```
# get unweighted unifrac distances
uu <- dist_subset(uu, s_vendor$SampleID)

# run pcoa
pc <- pcoa(uu)

# create dataframe for ggplot2
pc_df_uu <- cbind(s_vendor, pc$vectors[s_vendor$SampleID,1:3])

# calculate variance coverage by axis
```

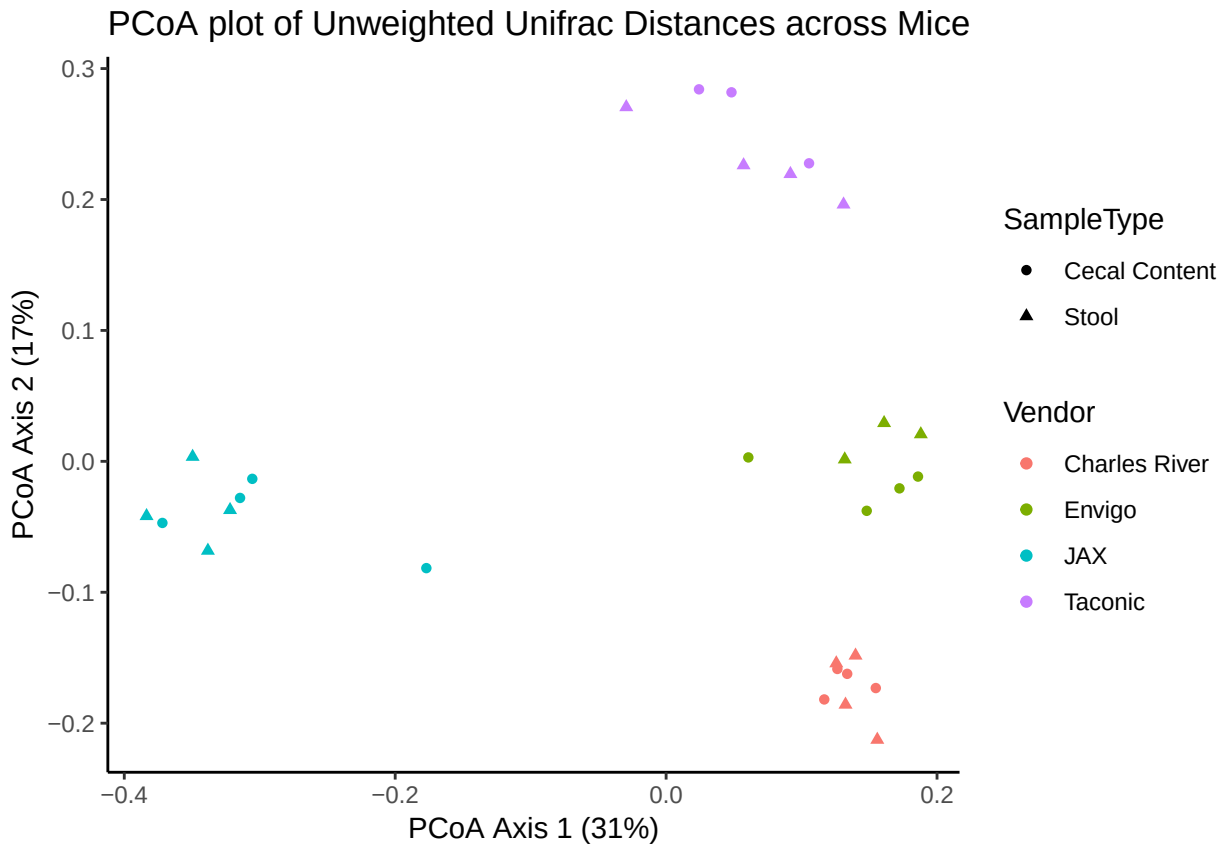

Figure 3.1: PCoA plot

```
pc_pct <- round(pc$values$Relative_eig * 100)

# finish setting up dataframe
pc_df_uu <- pc_df_uu %>%
  mutate(Label = ifelse(SampleID %in% suspect_SampleIDs, SampleID, ""))

# make fig
ggplot(pc_df_uu, aes(x = Axis.1, y = Axis.2)) +
  geom_point(aes(color = Vendor, shape = SampleType)) +
  geom_text(aes(label = Label)) +
  labs(
    title = "PCoA plot of Unweighted Unifrac Distances across Mice",
    x = paste0("PCoA Axis 1 (", pc_pct[1], "%)"),
    y = paste0("PCoA Axis 2 (", pc_pct[2], "%)")) +
  theme_classic()
```