DATA 1010
Problem Set 11
Due 30 November 2018 at 11 PM

Problem 1

Show that for each $\alpha \in [0,1]$, there exists $t \in [0,\infty]$ such that such that the likelihood ratio classifier h_t is the function $h : \mathcal{X} \to \mathcal{Y}$ which minimizes

$$L(h) = \alpha\,\mathbb{P}(h(X) = +1 \text{ and } Y = -1) + (1-\alpha)\mathbb{P}(h(X) = -1 \text{ and } Y = +1).$$

(a) Identify the relationship between α and its corresponding t value. (For simplicity, assume that $\mathcal{X}$ is finite.)

Hint: write $L(h)$ as a sum over the elements $x \in \mathcal{X}$. For each x, consider the resulting contribution to that sum if $h(x) = +1$, and similarly for $h(x) = -1$. Classify each x according to which of the two contributions is smaller.

(b) Determine and explain the motivation for this problem.

Problem 2

(a) Consider the coordinates of n points in $\mathbb{R}^p$, organized into an $n \times p$ matrix A. Suppose that `U, Σ, V = svd(A .- mean(A, dims=1))`, and explain why `V[:,1:k]'` is the matrix which maps each point in $\mathbb{R}^p$ to its coordinates in the subspace of $\mathbb{R}^p$ spanned by the columns of `V[:,1:k]`.

(b) Plot an image of the *third* principal component for the MNIST dataset. Identify a digit which you think should predominantly have a large or small dot product with this image, and make a scatter plot of which shows the dot product with the first principal component on the x-axis and the dot product with the third principal component on the y-axis. Check whether your prediction was accurate.

(c) What do you think the 100th principal component might look like, compared to the first few? Display it and check your prediction.

Problem 3

(a) Write an R function called `makelabels` which takes a vector of postive integers and returns a vector of strings with "label" prepended to the string representation of each integer:

R
```
makelabels(c(4,5,7)) == c('label4','label5','label7')
```

(b) Write an R function called `numzeros` which accepts a vector as an argument and returns the number of zeros in the vector.

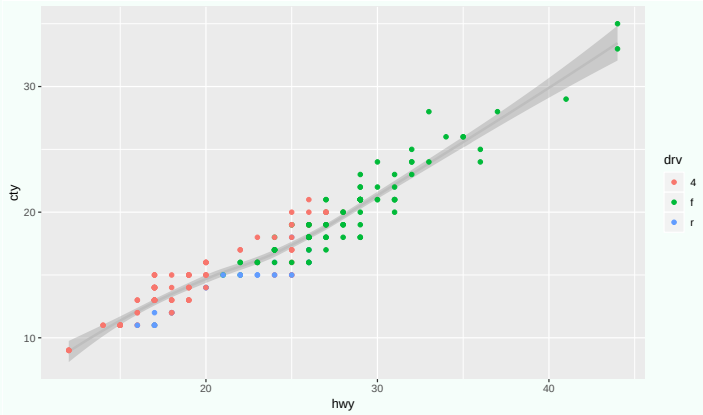
R
```
numzeros(c(-1,0,2,3,0,1)) == 2
```

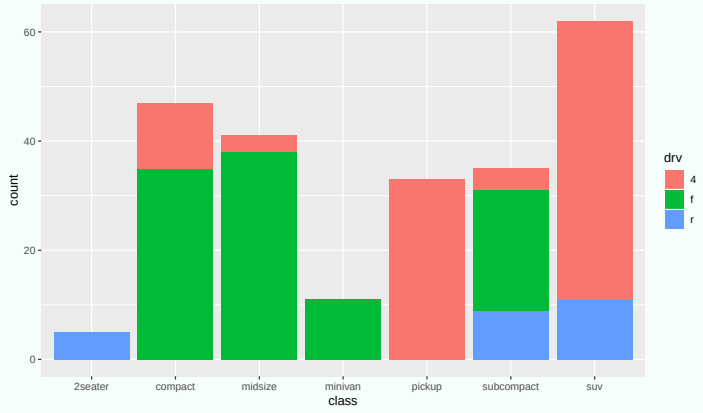
(c) Write an R function called `numincreasing` which accepts a vector as an argument and the number of components of that vector which are greater than the immediately preceding component.

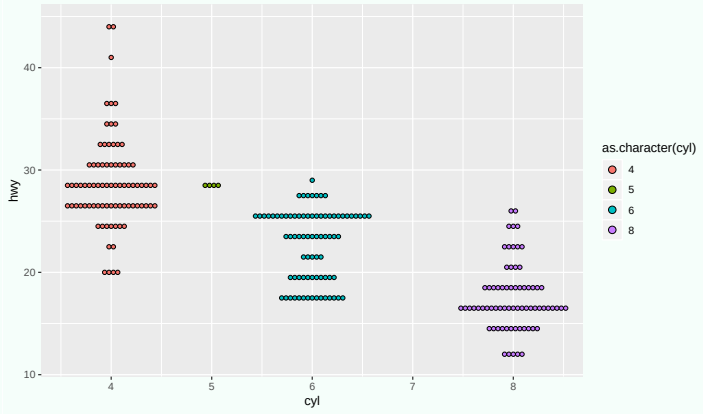
R
```
numincreasing(c(-1,0,2,3,0,1)) == 4
```

Problem 4

Use `ggplot2` to reproduce each of the following graphs. The dataset used is `mpg`, which is automatically loaded when you run `library(tidyverse)`.







Problem 5

Write `dplyr` code to perform each of the following operations on the `mpg` dataset. We say "average mpg" to mean the $\frac{1}{2}$ times the sum of the highway and city mpg recorded for each vehicle.

(i) Return a dataframe containing only the Audis with an average mpg of at least 24.

(ii) Return a dataframe with all of the cars sorted in decreasing order of average miles per gallon.

(iii) Return a dataframe with just the `trans` and `hwy` columns for all of the Volkswagens.

(iv) Return a dataframe with a new column containing each vehicle's average miles per gallon.

(v) Return a dataframe showing the average highway miles per gallon and average city miles per gallon for each manufacturer.