Homework 5

Stats 102B Lec 1 and 2

Spring 2024

General Guidelines

Show all your work, including any and all relevant code and output. Any and all collaboration must adhere to **Level 1** collaboration described in the Stats 102B Collaboration Policy.

Please submit your homework as a single file, in PDF format only. Name your assignment file with the convention `123456789_stats102b_hw0.pdf`, where `123456789` is replaced with your UID and `hw0` is updated to the actual homework number. Include your first and last name and UID in your assignment as well.

All R code is expected to follow the Tidyverse Style Guide: https://style.tidyverse.org/. If you scan or take a picture of any written work, please make sure the resolution is high enough that your work is clear and legible. Submissions with severe style or formatting issues may receive a penalty. Any submission that cannot be properly read will not be graded.

Please read in the `pokemon.csv` from Bruin Learn. Standardize all of the numeric variables (generation is not numeric).

Question 1

(a)

Subset the data frame to create two sub data frames: `gen1` which contains all of the Pokémon from generation 1 and `gen2` which contains all of the generation 2 Pokémon. Display the first 6 rows of each.

(b)

Your task is to build a k-nearest neighbors classifier that predicts whether a generation 2 Pokémon is legendary or not from its total points and hp (hit points) based on data from generation 1 Pokémon.

Write a function `knn(trainx, trainy, testx, k)` that implements the k-NN method using the Euclidean norm with the following parameters

- `trainx` is a data frame of the training observations
- `trainy` is a vector of training labels
- `testx` is a data frame of the testing (new) observations
- `k` is the number of neighbors

The output should be a vector of the predicted labels corresponding to the new observations.

(c)

Run your k-NN function for the above scenario with $k = 5$. Compute the misclassification rate. Some starter code has been provided for you.

```r
trainx <- gen1[, c("total_points", "hp")]
testx <- gen2[, c("total_points", "hp")]
trainy <- gen1$is_legendary
```

Note: Never copy-paste code from a PDF.

(d)

Use 6-fold cross validation with the misclassification rate as your error to determine which $k \in \{3, 7, 11, 21, 51\}$ is optimal. You can (and should) borrow code from the previous homework.

Hint: You should only be looking at generation 1 data, and the code may take a while to run…

(e)

Even if you achieved a very small misclassification rate, it may not tell the entire story. Another common way to display misclassification is with a **confusion matrix**, which is a two-way frequency table that summarizes the frequencies of the true classes of the observations against their predicted classes from a classification model. For example, in this question, a confusion matrix for a certain model could be written

	Predicted Non-Legendary	Predicted Legendary
Actually Non-Legendary	#	#
Actually Legendary	#	#

Repeat part (c) now using the k that gave you the smallest cross-validation error. Create a confusion matrix displaying your misclassification frequencies and explain your results.

Hint: `table()` works great here.

Question 2

In this problem, you will implement the k-means algorithm using a variety of distances.

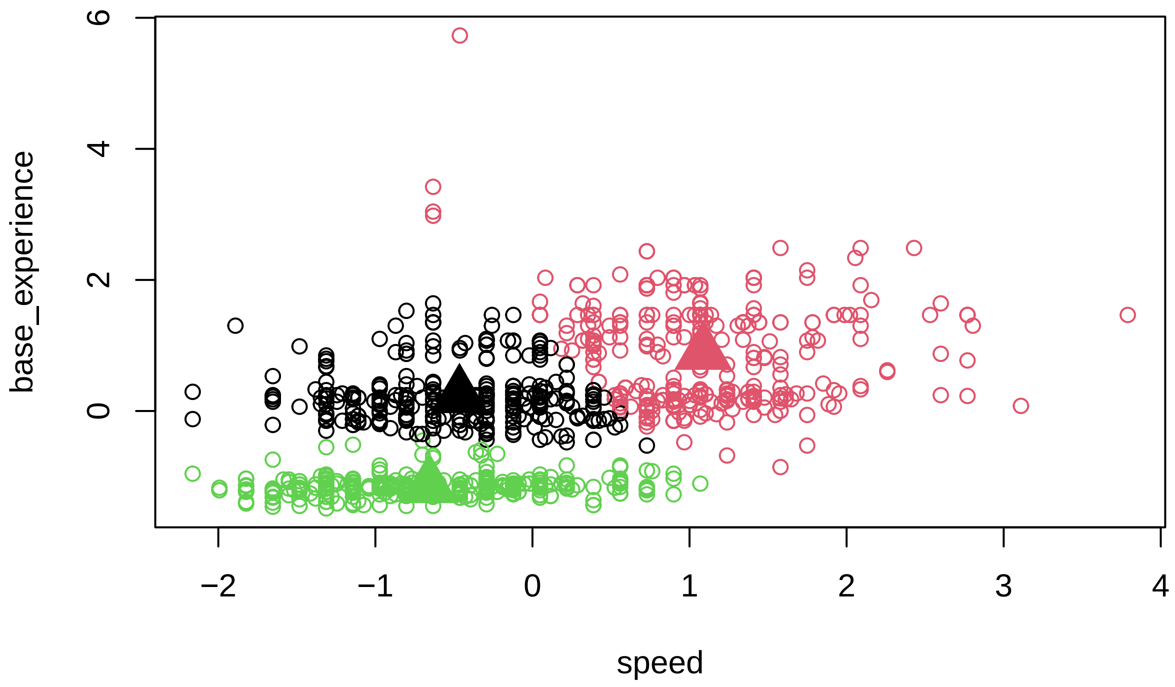
(a)

Using the Pokémon dataset, create a plot of `base_experience` against `speed`. Speed shold be on the x-axis and base experience should be on the y-axis. Comment on your plot.

(b)

Implement the k-means algorithm with the Euclidean distance to create clusters of the data you plotted above. Color each point according to their final cluster and plot the centroid as a larger triangle of the same color. The first of these plots has been provided below for you to check your work. Comment on your plot in relation to the chosen distance.

Hint: Write this as generally as possibly for any distance/metric function, as this is the first of many of these tasks.



(c)

Repeat part (b) with the Manhattan distance. Comment on your plot in relation to the chosen distance.

(d)

Repeat part (b) with the Mahalanobis distance. Comment on your plot in relation to the chosen distance.

(e)

Now consider the following distance defined below

$$d(x, y) = \sqrt{(x - y)^T Q (x - y)},$$

where

$$Q = \begin{bmatrix} 1 & -2 \\ -2 & 10 \end{bmatrix}.$$

Repeat part (b) with this distance. Comment on your plot in relation to the chosen distance. Your response should be specific to the different components of Q.

Question 3

(a)

In class, we mentioned centering our predictors before performing PCA, but what if we standardized them instead?

Consider a data set with two, uncorrelated, standardized predictors. Show that the loadings will be

$$\left\{ \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \right\}.$$

Furthermore, show that if the data is positively correlated, the first loading will be $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ while if the data is negatively correlated, the first loading will be $\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$. Explain what your derivations mean.

Hint: Is the correlation matrix in the room with us?

(b)

Using the Pokémon dataset, illustrate the above phenomenon by implementing PCA for the standardized attack and defense columns. Create two plots similar to lecture with the first plot showing the data in the original axes and the second plot showing the data in the principal component axes. For each plot, draw the loadings vectors (i.e., principal component directions) as arrows.

Note: Use the `attack` and `defense` columns. The columns `sp_attack` and `sp_defense` refer to special attack and special defense, which are different measures.

(c)

Now run PCA for the `height_m` through `base_experience` columns (i.e., all of the numeric columns in the dataset). Provide the first 6 elements of the first principle component.

(d)

Verify your results with the `prcomp()` function.