# Individual Assignment 3

@cderamond - MBA UTFSM

[**Permalink**](https://github.com/cderamond/mba_UTFSM_BI/blob/5095bb74780a26386f6d904160381725b96144d7/Ind3/p3.md) **Github**

## Questions

1. A data analyst working in your group reports to you that he is developing a k-NN model to predict the yearly sales to potential new customers, and that he is currently using a k-fold cross-validation procedure to settle on the best value for k. Explain this to your boss.

* k-fold is a procedure that aims to avoid bias on sampling by retraining and re-testing the model k times over the dataset split on equal k batches. Is a rotation exercise done over the set k times, always leaving 1 “k-fold” in which to test the model.

1. How do we prevent Euclidean distance from being biased by attributes with wide ranges of values?

* Use a method for weighting distance or standarization.

1. In a sentence or two, explain how k (in “k-NN”) is related to model complexity.

* k is the number of points to be randomly set in the space, as the models are complex or have elevated number of dimensions, the calculation gets trickier.

1. A data analyst working for you reports that she has decided to use a standard OLS linear regression model instead of a weighted k-NN regression model, even though the k-NN regression model seems to be providing slightly higher accuracy on the holdout test set, because she believes that you will have to explain the model in great detail in order to get support from important stakeholders within the organization. What does she mean? Do you agree with her?

* It could mean she has low number of observations, higher accuracy might be due to overfitting. Normally if the data indicates a linear function, simplicity is best and OLS might be better. On the other hand, k-NN is easy to explain, in concept and the explainability of the model: you have a different regression calculated on the average of the k-neighbors.

1. Why would we want to perform some type of feature selection before building a k-NN model? Name one feature selection method for categorical target variables and one for numeric target variables.

* For categorical values a random tree classifier might be good choice. For numerical values any kind of dimensional reduction (PCA for example), should do.

1. A data analyst reports to you that the team is building a “similarity-moderated k-NN regression model, using Manhattan distance instead of Euclidean distance.” What does that (whole phrase) mean?

* The distance is calculated as a ~taxi cab~ distance, it being the absolute sum of the orthogonal decomposition from the euclidean distance. In other words, imagine how you travel around in a city.

1. You report to your boss that your data science team is building a k-means cluster model to help better understand whether or not your customers occur in natural groupings. He laughs and says, “I don’t know what a k-means cluster model is. Matter of fact, I don’t even know what a cluster model is.” How would you further explain it to him in simple terms?

Normally the explanation is: > Imagine you put plot all the data, now you will have different shapes delimited by those points that are closer to each other usual center of mass problem. > The hypothesis behind this, is that those neighbors will share characteristics.

1. Explain the k-means clustering algorithm by writing out the steps as bullet points or a numbered list.

* **Steps for k-NN**
  1. pick k value
  2. k random points are set on the space
  3. assign each point to the closest point (distance calculation may vary)
  4. recompute new cluster centers iterate on steps 2 and 3 until convergence.

1. How can supervised learning be used to develop good cluster descriptions?

* a supervised model for classication will create and weight relevant classifiers.

1. What is the difference between a characteristic description (of a cluster) and a differential description?

* Differential tries to estimate a gradient among the k-neighbors, whereas characteristic is normal distance based, which does not really account for the underlying function.

1. What is the difference between the R functions lapply() and vapply()? Structure your answer around the advantages and disadvantages of each.

apply is a lambda function, it executes a given function over an array. lapply is apply to a list (complex Structure). vapply lets you select the data type of the output.

dat <- data.frame(x1 = 1:5,   
 x2 = 2:6,  
 x3 = 3)  
 apply(dat,1,sum)

## [1] 6 8 10 12 14

vapply(dat, length, integer(1))

## x1 x2 x3   
## 5 5 5

1. Explain what the following R code is doing, and try to guess why we would do this. Assume that you know that the churn.csv dataset has 20,000 rows. churn\_data <- read.csv(‘churn.csv’) idx <- sample(1:nrow(churn\_data), 16000) churn\_train <- churn\_data[idx,] churn\_test <- churn\_data[-idx,]

This is standard procedure to read a CSV data source and split it into randomized train and test set

churn\_data <- read.csv('churn.csv') # read data  
 idx <- sample(1:nrow(churn\_data), 16000) #split index with 16k samples from the length of dataset  
 churn\_train <- churn\_data[idx,] # define train set  
 churn\_test <- churn\_data[-idx,] # define test set

1. Which R function would you use to randomly draw some numbers from the standard normal distribution?

rnorm(20)

## [1] -0.2880423 -0.7361104 0.4242357 1.8589778 -1.2536664 1.3137408  
## [7] 0.5529574 -0.7319361 1.2278631 -0.4532393 -0.7922563 0.4003026  
## [13] -0.9182291 -0.0175759 -0.2655825 -1.0816319 0.6723108 1.1295969  
## [19] 1.0471515 -0.9325637

1. Which R function would you use to get the current date from the system?

Sys.time()

## [1] "2022-11-12 23:55:39 -03"

1. What is a POSIX date?

Is type of data that stores Date and time along with timezone.

a <-Sys.time()  
 str(a)

## POSIXct[1:1], format: "2022-11-12 23:55:39"

as.POSIXlt(a, tz="UTC" ) #same but UTC instead

## [1] "2022-11-13 02:55:39 UTC"