

Hw 7

Ryan Dee

1/5/2024

Recall that in class we showed that for randomized response differential privacy based on a fair coin (that is a coin that lands heads up with probability 0.5), the estimated proportion of incriminating observations \hat{P} ¹ was given by $\hat{P} = 2\hat{\pi} - \frac{1}{2}$ where $\hat{\pi}$ is the proportion of people answering affirmative to the incriminating question.

I want you to generalize this result for a potentially biased coin. That is, for a differentially private mechanism that uses a coin landing heads up with probability $0 \leq \theta \leq 1$, find an estimate \hat{P} for the proportion of incriminating observations. This expression should be in terms of θ and $\hat{\pi}$.

Student Answer $\hat{P} = (\hat{\pi} - (1 - \theta) * \theta) / \theta$

Next, show that this expression reduces to our result from class in the special case where $\theta = \frac{1}{2}$.

Student Answer $\hat{p} = (\hat{\pi} - (1 - \frac{1}{2}) * \frac{1}{2}) / \frac{1}{2} \quad \pi = \frac{\hat{p}}{2} + \frac{1}{4} \quad \hat{p} = 2\pi - \frac{1}{2}$

Part of having an explainable model is being able to implement the algorithm from scratch. Let's try and do this with KNN. Write a function entitled `chebychev` that takes in two vectors and outputs the Chebychev or L^∞ distance between said vectors. I will test your function on two vectors below. Then, write a `nearest_neighbors` function that finds the user specified k nearest neighbors according to a user specified distance function (in this case L^∞) to a user specified data point observation.

```
#student input
#chebychev function
chebychev <- function(vector1, vector2){
  return(max(abs(vector1 - vector2)))
}
#nearest_neighbors function
nearest_neighbors = function(x, obs, k, dist_func){
  dist = apply(x, 1, dist_func, obs) #apply along the rows
  distances = sort(dist) [1: k]
  neighbor_list = which(dist %in% sort(dist)[1:k])
  return( list (neighbor_list, distances))
}
```

¹in class this was the estimated proportion of students having actually cheated

```
x<- c(3,4,5)
y<-c(7,10,1)
chebychev(x,y)
```

Finally create a `knn_classifier` function that takes the nearest neighbors specified from the above functions and assigns a class label based on the mode class label within these nearest neighbors. I will then test your functions by finding the five nearest neighbors to the very last observation in the `iris` dataset according to the `chebychev` distance and classifying this function accordingly.

```
library(class)
df <- data(iris)
#student input
knn_classifier = function(x,y){
  groups = table(x[,y])
  pred = groups[groups == max(groups)]
  return(pred)
}

#data less last observation
x = iris[1:(nrow(iris)-1),]
#observation to be classified
obs = iris[nrow(iris),]

#find nearest neighbors
ind = nearest_neighbors(x[,1:4], obs[,1:4],5, chebychev)[[1]]
as.matrix(x[ind,1:4])
obs[,1:4]
knn_classifier(x[ind,], 'Species')
obs[, 'Species']
```

Interpret this output. Did you get the correct classification? Also, if you specified $K = 5$, why do you have 7 observations included in the output dataframe?

Student Answer In this case the output is correct, however, tied chebychev distances are included in the output, that is why there are 7 instead of 5.

Earlier in this unit we learned about Google's DeepMind assisting in the management of acute kidney injury. Assistance in the health care sector is always welcome, particularly if it benefits the well-being of the patient. Even so, algorithmic assistance necessitates the acquisition and retention of sensitive health care data. With this in mind, who should be privy to this sensitive information? In particular, is data transfer allowed if the company managing the software is subsumed? Should the data be made available to insurance companies who could use this to better calibrate their actuarial risk but also deny care? Stake a position and defend it

using principles discussed from the class.

Student Answer I would say that unless explicit consent is given by the patient, there should be no sharing with insurance companies, and only in instances of immediate danger, should the information be shared with healthcare providers, and only when consent is given. To defend this I will appeal to virtue ethics and the idea of wading in the middle of two extremes. On the one hand there are those who never want their information shared with anyone, and on the other hand there are those who have no problems with it and feel open about their health. Google should look to work somewhere in the middle between these two extremes. In cases where it is an emergency, Google should absolutely share their information with healthcare providers, similar to the question of whether you should save someone from drowning on the way to class, Google should share information when someones life is in danger. However, they should also ask for consent in all other cases, and never share information with insurance providers. This is the “middle” between two extremes, at the same time one could appeal to deontology by saying that sharing data without consent is not something we would want to be a universal maxim, and thus Google should not do it. In addition, in case of absorption, data should only be transfered if consent agreements are reached with the new company. This is to protect the integrity of the agreement because the agreement was originally struck with Google. Imagine a situation where Google is bought by an insurance company, data on acute kidney injury should only be shared if the patients agree to having their data shared with the new insurance company.

#

I have described our responsibility to proper interpretation as an *obligation* or *duty*. How might a Kantian Deontologist defend such a claim?

Student Answer A Kantian Deontologist would defend the claim that statisticians have an obligation to proper interpretation because we would wish the same for everyone other professsion (otherwise we would wish it to be a universal maxim). One wouldn't want a doctor misinterpreting your injury. One wouldn't want an economist that misinterprets data and invest your money incorrectly. We would want this to be the highest importance, not a suggestion, but an obligation, and thus if we shoul dexude that obligation ourselves.