Hw 7

Sydney

12/2/2024

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 $\hat{\pi}$.

$\hat{P} = \frac{\hat{\pi}-(1-\theta)\theta}{\theta}$

Next, show that this expression reduces to our result from class in the special case where $\theta = \frac{1}{2}$. $\hat{P} = \frac{\hat{\pi}-(1-\theta)\theta}{\theta}$ $\hat{P} = \frac{\hat{\pi}-(1-\frac{1}{2})\frac{1}{2}}{\frac{1}{2}}$ $\hat{P} = 2(\hat{\pi} - (1-\frac{1}{2})\frac{1}{2})$ $\hat{P} = 2(\hat{\pi} - \frac{1}{4})$ $\hat{P} = 2\hat{\pi} - \frac{1}{2}$ # 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
#nearest_neighbors function
chebychev = function(x,y){
  max(abs(x-y))
}

nearest_neighbors = function(r, obs, k, cheby){
  distance= apply(r, 1, chebychev, obs)
  distances= sort(distance)[1:k]
  neighbors= which(distance %in% sort(distance)[1:k])
  return(list(neighbors, distances))
}

x<- c(3,4,5)
```

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

```
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?

The classification was correct. The reason that we had seven observations was there were some "ties" in terms of distance when KNN was run so the algorithm pulled more observations than was required due to those ties in distance.

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.

I think that the people who should have access to the data must be covered in original consent forms. My opinion isn't that every individual who might have access to the data be named, but instead that every entity who may have access should be covered in original consent and tacit consent should not be allowed. Therefore, data transfer should not be allowed without procuring consent from the original person who's data is being used. Along with that, I don't think that insurance companies should be provided health information. Although it would be helpful for assessing their risk, the issues of data privacy and a person's information being exposed are too great and it shouldn't be permissible unless a person gives explicit consent for their information to be shared.

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

The responsibility to proper interpretation under deontology can be defended as we have a moral duty to always seek to interpret data in a way that is fair and seeks not to cause harm. Deontology is based upon creating a set of rules or principles that we seek to adhere to in all circumstances. Often it is seen as too rigid, but when we are looking at maxims of a field, such as statistics, it is a good lens through which to look at things. We must always seek to properly interpret data- and the always is what makes it applicable under deontology. It is a constant thought whenever we are doing any analysis, and because it is a constant it can be considered under deontology.