Hw 7

Aditya Krishna

12/1/2024

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 $\hat{\pi}$.

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

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

$\hat{\pi} = \frac{1}{2}\hat{P} + (1-\frac{1}{2})\frac{1}{2}$ where $\theta = \frac{1}{2}$ $\hat{\pi} = \frac{1}{2}\hat{P} + \frac{1}{4}$ which was shown in class

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.

```r
#student input
#chebychev function
chebychev = function(x, y) {
  max(abs(x-y))
}
#nearest_neighbors function
nearest_neighbors = function(x, obs, k, df){
  dist = apply(x, 1, df, obs)
  distances = sort(dist)[1:k]
  neighbors_list = which(dist %in% sort(dist)[1:k])
  return(list(neighbors_list, distances))
}

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

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

```r
y<-c(7,10,1)
chebychev(x,y)
```

```
## [1] 6
```

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.

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

#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])
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width
## 71           5.9         3.2          4.8         1.8
## 84           6.0         2.7          5.1         1.6
## 102          5.8         2.7          5.1         1.9
## 127          6.2         2.8          4.8         1.8
## 128          6.1         3.0          4.9         1.8
## 139          6.0         3.0          4.8         1.8
## 143          5.8         2.7          5.1         1.9
```

```r
obs[,1:4]
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width
## 150          5.9           3          5.1         1.8
```

```r
knn_classifier(x[ind,], 'Species')
```

```
## virginica
##         5
```

```r
obs[,'Species']
```

```
## [1] virginica
## Levels: setosa versicolor virginica
```

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 KNN algorithm correctly classified the observation as virginica. The first part of the output shows the 7 nearest neighbors instead of 5 because the the Chebyshev's distance was equal for at least the 5th-7th observations. To ensure fairness, the algorithm should include these observations.

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.

The only people that should be privy to this sensitive information are the doctors involved with patient care and the patient. Data transfer should not be allowed under any condition, even to the insurance companies or if the company is subsumed. According to deontology, patients should have be able to have autonomy over their lives as moral agents. As a result, the patient should be full control over who gets access to their data. By going to a hospital, a patient understands that their sensitive information is given to doctors so they can diagnose and provide a cure to their ailment. However, a patient is not signing up for Google, insurance companies, or any other individuals/organizations gaining access to that information.

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

A Kantian Deontologist would defend the claim by using the Universalizability Principle. This principle essentially states that we should only act if everyone could also act as we did and the world would still be desirable. If everyone misinterpreted statistics, then we would not be able to trust any information given by statisticians around the globe. Also, by misinterpreting the statistics, we treat those who helped provide the information and those who utilize the information as a means to an end. Deontology is vehemently against treating people as a means to an end and also requires the Universalizability Principle to be upheld with every action. Therefore, a statistician's responsibility to proper interpretation is considered an obligation or duty.