Hw 7

Tara Hinton

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 $\hat{\pi}$.

Let's outline some proportions first. Coin lands on heads: θ; Coin lands on tails: 1 - θ; So, $\hat{\pi} = \theta\hat{P}$ + (1 - θ)θ; $\hat{P}$ = ($\hat{\pi}$ - (1 - θ)θ)/θ

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

Our result from class: $\hat{P} = 2\hat{\pi} - \frac{1}{2}$; $\hat{P}$ = ($\hat{\pi}$ - (1 - 0.5)0.5)/0.5;

$\hat{P}$ = ($\hat{\pi}$ - 0.25)0.5; Multiplying by 2, $\hat{P} = 2\hat{\pi}$ - $\frac{1}{2}$

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

cheby <- function(x,y){

  max(abs(x - y))
}
```

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

```r
#nearest_neighbors function

nearest_neighbors = function(r, obs, k, cheby){
  dist = apply(r, 1, cheby, obs) #apply along the rows
  distances = sort(dist)[1:k]
  neighbor_list = which(dist %in% sort(dist)[1:k])
  return(list(neighbor_list, distances))
}

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

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, cheby)[[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
```

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

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

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

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

```
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?

Using the Chebyshev distance and our respective KNN classifier function, we correctly classified this observation as the species "Virginica." We have 7 observations included in the output dataframe likely as a result of "ties" in the distance calculations. It is possible that the Chebyshev distance metric produced several nearest neighbors that were the same distance from our observation of interest and were therefore included in the output.

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.

Personal/sensitive data should remain solely accessible to companies under which consent was originally given (this consent being tacit, explicit, and informed). Included in this position is the sharing of information with insurance companies, which should be prohibited unless proper consent has been given. Philosopher John Rawls tells us that a just society should consider the most vulnerable among us to allocate resources. Applied to data privacy, Rawls position supports the idea that private data should remain private, particularly when harm to a vulnerable party is at stake. In the face of increased insurace rates or a complete denial of care, data privacy will ensure the utmost protection for vulnerable populations. This issue perhaps becomes murkier in the face of a company that might potentially use data for the good of vulnerable populations. Still, this paternalistic view may inflict harm on vulnerable populations if data security differs between companies, and limit the autonomy of individuals in these populations to receive good care. Tacit, informed, and explicit consent should therefore always be given the case of subsumed company management or data transfer.

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

Kantian Deontology understands morality as a set of duty-bound actions, consistent under a universal categorical imperative. We undertand the duty of the statistician to interpret results as a set of consistently applied actions taken to ensure honesty and communicate uncertainties, implications, and the scope of statistical results to outside audiences. These actions to fully interpret data apply to results both expansive and small, to every statistician across the world. This obligation ensures that decision-makers and other data-users understand the possibilities for good and harm inherent in statistical outcomes, and it provides a global imperative for thoughtful analysis. Thus, we can see comprehensive statistical interpretation as a moral obligation under deontology.