# Hw 7

Ava Klissouras

11/20/2024

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 $\theta$ and $\hat{\pi}$.

**Let us consider the possible ways individuals will answer affirmative to the incriminating question.**

**An individual who flips heads first must tell the truth. So, the probability of an individual answering affirmatively is the probability of flipping heads, $\theta$, times the estimated proportion of incriminating observations, $\hat{P}$.**

**An individual who flips tails first, and then heads, must also answer affirmatively. The probability of this occuring is $\theta(1-\theta)$, so this is the proportion of individuals who will answer affirmatively in this way.**

**These cover all possible cases in which an individual would answer affirmatively.**

**So, all in all, the proportion of individuals answering affirmatively, $\hat{\pi}$ is the sum of our above two proportions: $\theta(\hat{P})$ and $\theta(1-\theta)$. So we have $\hat{\pi} = \theta(\hat{P}) + \theta(1-\theta)$, and we want to solve for $\hat{P}$.**

**Subtracting $\theta(1-\theta)$ from both sides and dividing by $\theta$ yields $\hat{P} = \frac{\hat{\pi} - \theta(1-\theta)}{\theta}$.**

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

**Let $\theta = \frac{1}{2}$. Then $\hat{P} = \frac{\hat{\pi} - \frac{1}{2}(1-\frac{1}{2})}{\frac{1}{2}} = \frac{\hat{\pi} - \frac{1}{4}}{\frac{1}{2}} = 2\hat{\pi} - \frac{1}{2}$, the result we obtained from 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 $L^\infty$ distance between said vectors. I will test your function on two vectors below. Then, write

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

a `nearest_neighbors` function that finds the user specified $k$ nearest neighbors according to a user specified distance function (in this case $L^\infty$) to a user specified data point observation.

```
#student input
#chebychev function
chebychev = function (x,y){
    distmax = 0
    i = 1
    while (i <= length(x)){
    dist = abs(x[i] - as.numeric(y[i]))
    if (dist > distmax) {
      distmax = dist
    }
    i = i + 1
    }
    return (distmax)
}

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


#nearest_neighbors function

nearest_neighbors = function(x, obs, k, dist_func){
  x = as.matrix(x)
  dist = as.numeric(apply(x, 1, dist_func, obs)) #apply along the rows
  distances = sort(dist)
  neighbor_list = which(dist %in% distances[1:k])
  df = data.frame(Index = neighbor_list, Distance = dist[neighbor_list])
  return(df)
}
```

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?

**Yes, I got the correct classification. There are 7 observations included in the output data frame when specifying k = 5 because some of the distances were tied. When we found which were the 5 closest distances, multiple observations shared some of these distances, leading us to have 7 observations included in the output data frame.**

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.

**Deciding who should be privy to sensitive healthcare information is a difficult decision that relies on many factors. Above all, only individuals whom users consent to view/utilize their data should be able to access it, if consent can be feasibly given by users of these algorithms. Something as simple as a pop-up message informing users who will access their data upon setting up the app/accessing the algorithm, and requiring the user to accept these terms before proceeding, can resolve many of the ethical dilemmas involved. According to Kant, if consent is given, then even if an act harms an individual, it is still morally sound. Therefore, if users consent to certain individuals accessing their data, then this access is ethical.**

**That being said, there are some intricacies worth delving into. For instance, suppose another company acquires the original company. Should they have access to the same data? One option is notifying users of this acquisition and giving them the option to redact their individual data from the algorithm's database. However, this may result in a large loss of data if many users are uncomfortable with this data transfer, resulting in an algorithm that may exhibit increased bias or less accurate predictions. Some might argue that tacit consent may be sufficient in the case of the company managing the software being subsumed, because the users will still benefit from the algorithm's results, thereby participating in a social exchange of sorts. Moreover, if**

**users already consented to their data being used in a certain manner, if the new company uses their data in the same manner, and does use it for any additional purposes (such as selling it, or training additional algorithms for use outside of DeepMind) then perhaps tacit consent is sufficient. Otherwise, using the data without informing clients of this subsumption is not ethical.**

**Lastly, the data should not be made available to insurance companies to calibrate their actuarial risk but also deny care, unless informed consent is explicitly given on the users' part. Users whose data is given to insurance companies without their consent are treated as a mere means to an end when viewed through a Deontological framework. Their data can be utilized to deny them care, without them receiving any benefits in exchange. Since one of Kant's two formulations of the categorical imperative is thus violated, this application of users' data would be immoral.**

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 responsibility to proper interpretation by arguing that a lack of proper interpretation violates at least one of the two formulations of the categorical imperative. The first formulation of the categorical imperative is that an act is only moral if it can be universalized without logical contradiction. Suppose we generalize the act of disseminating non-interpretable models, or interpreting them incorrectly for the sake of convincing others of a certain result. Then, if all models are assumed to be uninterpretable or interpreted incorrectly, models that are properly interpreted will not convey any useful information. This is because nobody will believe the interpretation to be true, or they will not even try to interpet the model because they assume it is not interpretable. Therefore, we reach a logical contradiction, and lack of proper interpretation can be categorized as immoral. Conversely, we can therefore state that proper interpretation is a moral obligation.**

**For the sake of completeness, let us also examine Kant's second formulation of the categorical imperative, which is that an act is only moral if it treats others as ends, rather than mere means. We may consider a few cases in which proper interpretation is not satisfied.**

**The first is that a model is interpretable, but its results are interpreted incorrectly by the individuals who develop the model, in order to convey a certain result that may not necessarily directly follow from the model's output. This scenario treats the individuals who contribute their data to training/testing the model as means rather than ends. They are simply being used to provide data that is spun to serve the developers' goals, without receiving anything in exchange (as the model is not correctly interpreted and thus can pose no benefit to those whose data is used in its development).**

**Alternatively, suppose developers provide a model to stakeholders or individuals without a technical background, but that model is not readily interpretable in and of itself (perhaps something like an ensemble model) and fail to provide alternative methods of explanation, like feature importance or visualization. In this case, the individuals whose data is used to train/test the model algorithm are again used as a mere means to an end, albeit for a different reason. The algorithm can pose no benefit to stakeholders or society unless its results can be understood and appreciated by humans, so those who contribute data are not reaping any benefits of the algorithm by doing so.**

**Lastly, suppose developers provide an algorithm to stakeholders that lacks interpretability, and these stakeholders purposefully disseminate an improper interpretation of its results into society. This treats the developers as a mere means to an end, as their work is being used to**

**convey an idea or viewpoint that is incorrect and entirely a product of the stakeholders' motivations. Clearly, based upon deontology, proper intepretation is a moral obligation because improper intepretation violates both formulations of the categorical imperative.**