Hw 7

Quyen Dang

12/3/2024

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 $\hat{\pi}$.

Assuming that the coin lands with probability θ, $\hat{\pi} = \theta\hat{P} + (1-\theta)\hat{P}$. Solving this for $\hat{P}$,

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

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

Plugging in $\theta = \frac{1}{2}$:

$\hat{P} = \frac{\hat{\pi} - (1-1/2)1/2}{1/2}$

$\hat{P} = \frac{\hat{\pi} - 0.25}{0.5}$

$\hat{P} = \frac{\hat{\pi}}{0.5} - \frac{0.25}{0.5}$

$\hat{P} = 2\hat{\pi} - \frac{1}{2}$ This reduces to the result 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^∞ 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
chebychev <- function(x, y) {
  return(max(abs(x - y)))
}

nearest_neighbors <- function(data, point, k, distance_function) {
  distances <- apply(data, 1, function(row) distance_function(row, point))
```

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

```r
  cutoff_distance <- sort(distances, partial = k)[k]

  nearest_indices <- which(distances <= cutoff_distance)
  nearest_distances <- distances[nearest_indices]

  return(list(nearest_indices, nearest_distances))
}

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.

```r
library(class)
df <- data(iris)

knn_classifier <- function(nearest_neighbors_data, class_column) {
  class_labels <- nearest_neighbors_data[[class_column]]

  mode_label <- names(sort(table(class_labels), decreasing = TRUE))[1]

  return(mode_label)
}

#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, we got the correct observation, as both the actual species and the knn classified species outputted virginica. We received 7 observations despite specifying 5 because multiple observations in the dataset may have the same Chebychev distance to the point. When these ties exist, the knn function we implemented includes all observations with the same 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.

From a consequentialist perspective, access to sensitive healthcare data should be limited to individuals directly involved in patient care. The primary justification for gathering and using such data is to improve patient health, streamline medical processes, and enhance clinical decision-making. Granting access to individuals outside the immediate healthcare ecosystem risks potential misuse. This could lead to adverse outcomes, including losing trust in the healthcare system, reduced patient willingness to share critical information and poorer health outcomes. Thus, access must be restricted to professionals explicitly focused on improving patient care and well-being to maximize the benefits of sensitive healthcare data.

The transfer of sensitive healthcare data during a corporate acquisition must be carefully considered under consequentialism. Such a transfer could be ethically justifiable if it ensures the continuity of care, supports technological innovation or enhances the effectiveness of the tools used for patient treatment. However, if the transfer increases the risk of misuse or repurposing of data for profit motives unrelated to healthcare, it could lead to harm. To achieve the best outcomes, any data transfer should include strict safeguards, legal protections, and reassurances that the data will continue to be used only for original healthcare purposes. Requiring patient consent further reduces potential harm.

Granting insurance companies access to sensitive healthcare data is ethically problematic under consequentialism because the potential harms outweigh the benefits. While insurers could use the data to assess risks better and design more tailored policies, this could lead to discriminatory practices, such as denying coverage for high-risk individuals. Such actions would increase healthcare inequalities. These negative consequences would likely outweigh any benefits from improved risk assessment. Additionally, fear of data misuse by insurers could deter patients from seeking medical care or sharing accurate information with providers. Therefore, insurance companies should not have unrestricted access to sensitive healthcare data to maximize societal well-being and minimize harm.

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 as a moral duty based on the concept of the categorical imperative, which demands that actions must be guided by principles that can be universalized and respect individuals as ends in themselves. Proper interpretation of data ensures that we communicate truthfully and uphold the integrity of information. This aligns with the universal maxim formulation, which requires that our actions be generalizable without contradiction. Misinterpreting or deliberately skewing data would undermine trust in the broader community of interpretation, leading to a collapse in the reliability of data-driven decisions. Therefore, the duty of proper interpretation is grounded in maintaining a universally trustworthy system of knowledge.

Additionally, the ends-not-means formulation of Kant's categorical imperative reinforces this duty by prohibiting the instrumentalization of others through misleading or improper interpretations. Misinterpreting data could manipulate stakeholders, whether patients rely on accurate medical diagnoses or policymakers base decisions on flawed evidence. Such actions would treat these individuals as mere tools for achieving the interpreter's goals rather than respecting them as rational agents entitled to the truth. For a Kantian deontologist, adhering to proper interpretation is not just an act of professional ethics but a fundamental moral obligation to treat others with dignity and uphold the social contract of honest communication.