Hw

Lalida Kungval

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 $\hat{P} = 2\pi - \frac{1}{2}$ where π 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 π.

coin lands head and answer is yes: $\theta * \hat{P}$ coin lands tail and answer is yes: $(1-\theta)(\theta)$

proportion of "yes" answers: $\pi = \theta * \hat{P} + (1-\theta)(\theta)$

rearranged for $\hat{P}$: $\hat{P} = \frac{\pi - \theta + \theta^2}{\theta}$

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

substitute $\theta = \frac{1}{2}$: $\hat{P} = \frac{\pi - \frac{1}{2} + \frac{1}{2}^2}{\frac{1}{2}}$

simplify: $\hat{P} = \frac{\pi - \frac{1}{2} + \frac{1}{4}}{\frac{1}{2}}$ $\hat{P} = \frac{\pi - \frac{1}{4}}{\frac{1}{2}}$ $\hat{P} = 2\pi - \frac{1}{2}$

Consider the additive feature attribution model: $g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$ where we are aiming to explain prediction f with model g around input x with simplified input x'. Moreover, M is the number of input features.

Give an expression for the explanation model g in the case where all attributes are meaningless, and interpret this expression. Secondly, give an expression for the relative contribution of feature i to the explanation model.

If all attributes are meaningless, this implies that none of the features x'_i have any effect on the output of the model. Therefore, each feature contribution ϕ_i for $i = 1, \ldots, M$ should be zero, as no feature is contributing any information that affects the model's prediction. The expression for the explanation model $g(x')$ when all attributes are meaningless is then given by: $g(x') = \phi_0$

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

In this scenario, the output of the model is entirely determined by the intercept ϕ_0. The prediction does not change regardless of the input features.

The relative contribution of feature i to the explanation model is given by the term $\phi_i x'_i$.

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
cheby <- function(x, y) {max(abs(x - y))}
#nearest_neighbors function
nearest_neighbors <- function(data, observation, k, distance_func) {
  distances <- apply(data, 1, function(row) distance_func(row, observation))
  nearest_indices <- order(distances)[1:k]
  return(data[nearest_indices, ])
}


x<- c(3,4,5)
y<-c(7,10,1)
cheby(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)
#student input
knn_classifier <- function(neighbors, class_col_name) {
  most_common_label <- as.character(sort(table(neighbors[[class_col_name]]), decreasing = TRUE)[1])
  return(most_common_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])
```

```r
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 species is classified correctly as virginica. Having 7 observations although specifying K=5 might be due to multiple points having the same distance to the observation, exceeding the number of K value.

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.

** Consequentialism emphasizes the assessment of outcomes, suggesting that while the large-scale healthcare improvements from data sharing might justify the privacy risks involved, any potential for misuse, such as denial of care by insurance companies, presents a strong counterargument. In contrast, deontological ethics, particularly Kant's categorical imperatives, would argue for the upholding of patient autonomy and privacy, insisting that data not be used in any manner not explicitly consented to by patients, thereby opposing the unconsented transfer of data to third parties. Meanwhile, virtue ethics would expect organizations to exhibit benevolence, honesty, and fairness, treating patient data with the highest regard for respect and care. The principle of justifiability mandates that the benefits provided by the use of data must clearly outweigh the risks, thereby allowing for data transfer or sharing when it results in substantial public health advantages. Finally, drawing on philosophical measures of fairness and justice, Rawls's concept of the veil of ignorance would propose that data management policies should be such that everyone would consent to them without bias, while Nozick's perspective on historical justice would require that data is acquired and utilized fairly, with informed consent.

Given the above principles, the stance here is that sensitive health data should be managed with a priority on patient consent and privacy, used only in ways that directly benefit healthcare outcomes and research, and not for purposes that could lead to discrimination or harm, such as denying insurance coverage or care. Patients should be informed and give consent for how their data is used. This respects their autonomy and personal dignity. Data should not automatically be transferred if the managing company is subsumed. Patients should be informed and their consent obtained once more, respecting their right to privacy and acknowledging the change in data stewardship. Providing data to insurance companies poses significant risks, as it could lead to discrimination and exacerbate inequalities in healthcare. The principles of non-arbitrary discrimination, fairness, and the harm principle suggest that unless there are strong safeguards and transparency in place, this should not occur.

To ensure fairness and protecting privacy, federated Learning can be utilized to train algorithms on decentralized data, reducing the risk of data breaches and misuse. Implementing differential privacy techniques can help in using data for research and analysis while protecting individual identities. Binding agreements like NDA can ensure that data is not misused and is purged when no longer needed.**