HW 7

Grace Sun

4/19/24

1

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.

Student Answer

Assume $\hat{P}$ is the estimated proportion of students who actually cheated, and π is the proportion of students who responded yes.

$E[Y] = \pi = \theta\hat{P} + (1-\theta)\theta$

$\pi = \theta\hat{P} + \theta - \theta^2$

$\theta\hat{P} = \pi - \theta + \theta^2$

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

2

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

Student Answer

Substitute $\theta = \frac{1}{2}$ into $\hat{P} = \frac{\pi}{\theta} - 1 + \theta$.

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

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

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

3

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.

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

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.

Student Answer

In the case that all attributes are meaningless, $g(x') = \phi_0$. This is because the marginal impact of each attribute is 0, so $\phi_i = 0$ for i = 1,..., M, thus making each $\phi_i x_i$ equal to 0 and the entire summation also equal to zero. This expression means that the model g will always predict ϕ_0, which is the expected value or base value prediction that the model would make if we do not know any features to the current output.

The relative contribution of feature i to the explanation model is given by ϕ_i. Each feature i changes the prediction from the baseline of ϕ_0 by adding through the summation, hence the name "additive feature attribution model".

4

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
chebychev = function(x, y){
  max(abs(x-y))
}

#nearest_neighbors function
nearest_neighbors = function(x, obs, k, dist_func){
  dist = apply(x, 1, dist_func, obs)
  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)
chebychev(x,y)
```

5

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']
```

6

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?

Student Answer

Yes, the classification is correct – the chosen point was classified as virginica and is actually virginica. There are 7 observations included in the output because all seven points have a chebychev distance of 0 from our test point, meaning at least one of the features is exactly the same as the test point for each of these 7 observations. The logic of the nearest_neighbors function allows for these ties to be outputted even when there are larger than k matches because it finds the smallest k distances from the observation, then searches the the training data for which indices have those specified distances (as seen in this line from the function: neighbor_list = which(dist %in% sort(dist)[1:k])). Because the top 5 distances are all 0, and there are 7 indices with distances of 0 from the testing point, 7 observations are included instead of 5.

7

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.

Student Answer

The sensitive health care data that was collected by Google should only be accessible to the parties that are specifically outlined when informed consent is obtained by Google to access and store an individual's data. This data should not be allowed to be transferred if the company managing the software is aquired or available to insurance companies without explicit informed non-coercive consent from each individual. By allowing patients to maintain control over access to their own private data, patients can maintain personal autonomy.

The transfer of data to an acquiring company or insurance company can not be argued to be rightfully paternalistic and for the patients' "own good". The transfer of sensitive data to a new company if the current company is aqcuired is not paternalistic because in the case that the data is not used, the potential for harm is not to the extent of applying the harm principle to limit personal autonomy. While a highly accurate model that can help diagnose or suggest treatment for conditions can increase patient wellbeing, a model is not a patient's only option, as doctors are highly skilled and able to make these decisions and judgements as

well. Thus, it would be incorrect to assume that not using sensitive healthcare data to develop models is in some way causing harm to the extent of invoking the harm principle, so paternalism can not apply here. The sharing of data with insurance companies for use of calibrating actuarial risk is also not paternalistic, since the primary goal of actuarial calculations is for the purpose of making money for the insurance company with risk evaluation. Actuarial calculations are done to decide the insurance premium that individuals pay and determine if coverage will be offered or not. While some individuals may benefit from a lower insurance premium with more accurate actuarial models, the people that are turned away or have their premiums raised produce far more pain than the pleasure of saving some money. When seeing this from a utilitarian perspective, sharing sensitive healthcare data with insurance companies serves to hurt more than help.

There is not enough evidence that these models are significantly changing medical outcomes to the extent that paternalism can be argued to apply here. In other words, the transfer of sensitive healthcare information without informed consent can not be justified as for the patient's own good, especially with the harms that can arise from the sharing of this sensitive data, and the lack of tangible harm caused if this sensitive data is not shared. If a company is acquired and data needs to be transferred or the company wants to share data with insurance companies, the explicit informed and non-coercive consent of each and every individual needs to be obtained.