Homework 2

Yunjeon Lee(yl7143)

April 15, 2022

1 Problem 1

1.1 (a)

(1) Pre-existing bias

If the previous crime data including drug use has many datas that has attribute "race" labeled as black people, then the system would be trained to black people are high risk for the crime. In this problem, especially for black people in the Oakland area might be labeled as high risk for the drug use. The system can have race bias for the prediction. Later, even though we delete the attribute 'race' itself, the system can guess specific race with other attribute values such as income, location or education. And the system would make prediction for the black people as high risk because of pre-exisiting knowledge toward specific race.

(2) Technical bias

If the data has missing values, and if the system is designed to set default value as the most frequent value for the categorical data, then the race that constitutes the most in previous data can be biased and experience race disparity. This is because this race will be marked as higher risk than what they really is. The reason for the racial disparity here is because of processing the missing values, so it is technical bias.

(3) Emergent bias

The system is trained by previous arrest data from drug use, and this previous data affects the algorithm of the system continuously because that previous data is used for training the model in system. If the system predicts that the specific area(in the problem, for instance, West Oakland(1) and International Boulevard(2)) might have higher probability of occurrence of drug use, the police would target that area. This leads to more arrest to drug users in that specific area, and this new arrest data might be used for updating the algorithm itself. Within this updated data, the system would predict that the police should focus on the specific area again. Considering that this area is a place where many of non-white and low-income people live, this system leads to racial disparity toward non-white people.

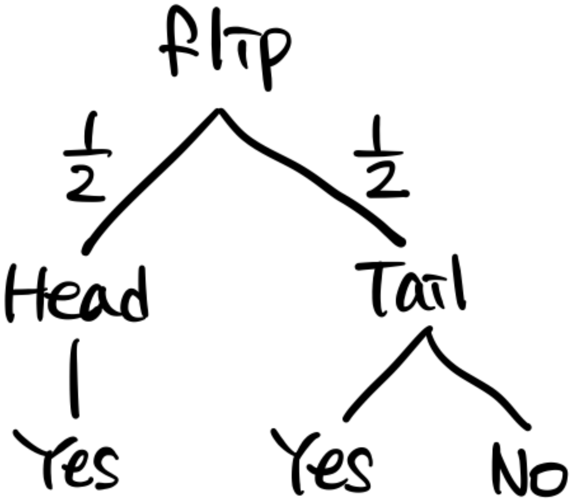
1.2 (b)

(1) I would use data cleaning to mitigate the disparity. The way to mitigate the racial disparity is to remove the column "race". The reason is that systems like in this problem does not need attribute race, because crime and race does not have direct relationship and this attribute causes disparity.

(2) Using the library mlinspect, we can conclude whether the data is biased or not. When we know that the data is biased, adding noise to the data can be one solution. This is because to add noise to the biased data can mitigate the bias, but there is a disadvantage that adding noise to the data can decrease utility. We should be aware of and careful about this disadvantage.

2 Problem 2. Randomized response

Below picture illustrates how randomized response works in this problem. If the coin comes up tail, the individual should answer truthfully which are "Yes" or "No". If the coin comes up head, the individual should always answer "Yes". In this question, the answer "No" is obtained when the individual should answer truthfully, which means randomness is not added for answer "No". Accordingly, this mechanism is not differentially private.



3 Problem 3. Classification association rules

3.1 (a)

item set	support	confidence
M → Yes	11	11/16
F → Yes	5	5/16
M / HS → Yes	3	3/6
M / BS → Yes	4	4/6
M / MS → Yes	4	4/4
F / HS → Yes	0	0/6
F / BS → Yes	3	3/6
F / MS → Yes	2	2/4
M → No	5	5/16
F → No	11	11/16
M / HS → No	3	3/6
M / BS → No	2	2/6
M / MS → No	0	0/4
F / HS → No	6	6/6
F / BS → No	3	3/6
F / MS → No	2	2/4

CARs

item set	support	confidence
M → Yes	11	11/16
F → No	11	11/16
M / BS → Yes	4	4/6
M / MS → Yes	4	4/4
F / HS → No	6	6/6

3.2 (b)

Let CARs from (a) as below:

$$\varepsilon_1 : \text{M} \rightarrow \text{Yes}$$
$$\varepsilon_2 : \text{F} \rightarrow \text{No}$$
$$\varepsilon_3 : \text{M / BS} \rightarrow \text{Yes}$$
$$\varepsilon_4 : \text{M / MS} \rightarrow \text{Yes}$$

$$\varepsilon_5 : \text{F / HS} \rightarrow \text{No}$$

For the perspective of M, ε_3 and ε_4 are included in ε_1. ε_3 and ε_4 are disjoint, because a person(especially male) can not have edu value as both BS and MS at the same time. Accordingly, Parallel composition can be applied to ε_3 and ε_4, and Sequential composition can be applied to ε_1, ε_3, and ε_4 as below.

$$\varepsilon_1 + \max(\varepsilon_3, \varepsilon_4)$$

For the perspective of F, Sequential composition can be applied to ε_2 and ε_5 as below.

$$\varepsilon_2 + \varepsilon_5$$

Since sex value can not be both M and F at the same time, Parallel composition can be applied with regard to sex as below. And set each equation as A and B respectively.

$$\text{A} = \varepsilon_1 + \max(\varepsilon_3, \varepsilon_4)$$
$$\text{B} = \varepsilon_2 + \varepsilon_5$$

A and B are disjoint because sex=F and sex=M are disjoint, which means Parallel composition should be applied.

$$\max(\text{A}, \text{B}) = 1 \text{ (overall privacy budget)}$$

High ε value means less noise and higher utility and the data is more faithful.

Considering that more noise to more general one makes more utility, we should set higher epsilon values to ε_1 and ε_2 (which are more general).

In order to maximize the utility, I set the maximum value for each element(epsilon in these equation) in max(). For example, if max(c, d) = 1, set c=1 and d=1.

$$\text{A} = 1,\ \varepsilon_1 > \varepsilon_3,\ \varepsilon_4$$
$$\text{B} = 1,\ \varepsilon_2 > \varepsilon_5$$

For example,

$$\text{if } \varepsilon_1 = 0.7,\ \max(\varepsilon_3, \varepsilon_4) = 0.3 \text{ which means } \varepsilon_3 = 0.3 \text{ and } \varepsilon_4 = 0.3$$
$$\text{if } \varepsilon_2 = 0.7,\ \varepsilon_5 = 0.3$$

4 Problem 4

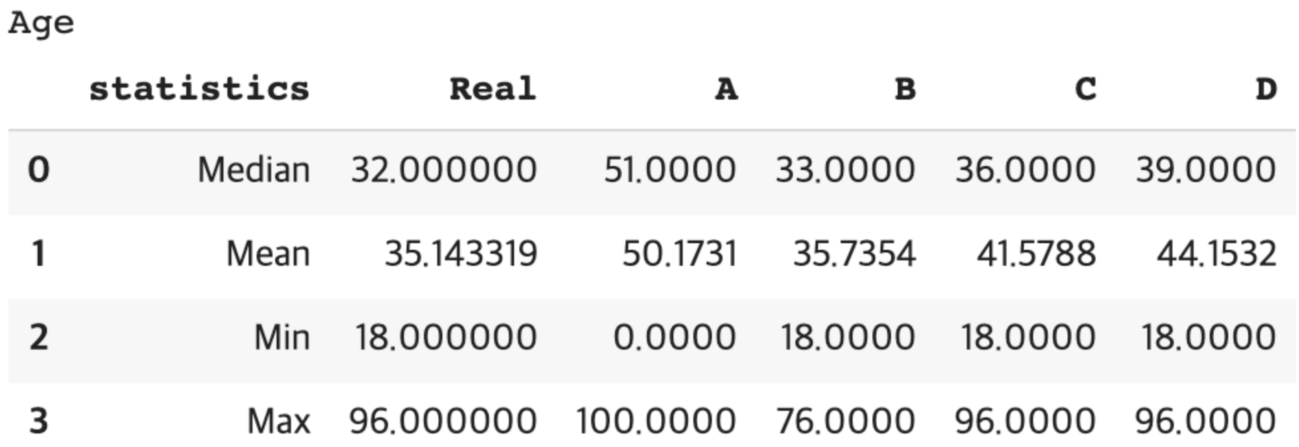
4.1 (a)

Q1.

For the attribute age, method B is the most accurate compared to other methods(A,C,D). The reason is that B shows similar values for Median, Mean, and Min when compared to real data, although there is some difference in Max values. However, considering that max is the extreme value, B shows the most accurate compare to other methods. Next, C is more accurate than other methods C and D, especially for median and mean (because min and max are the same for B,C,D). With the same reason, D and A is accurate in order.

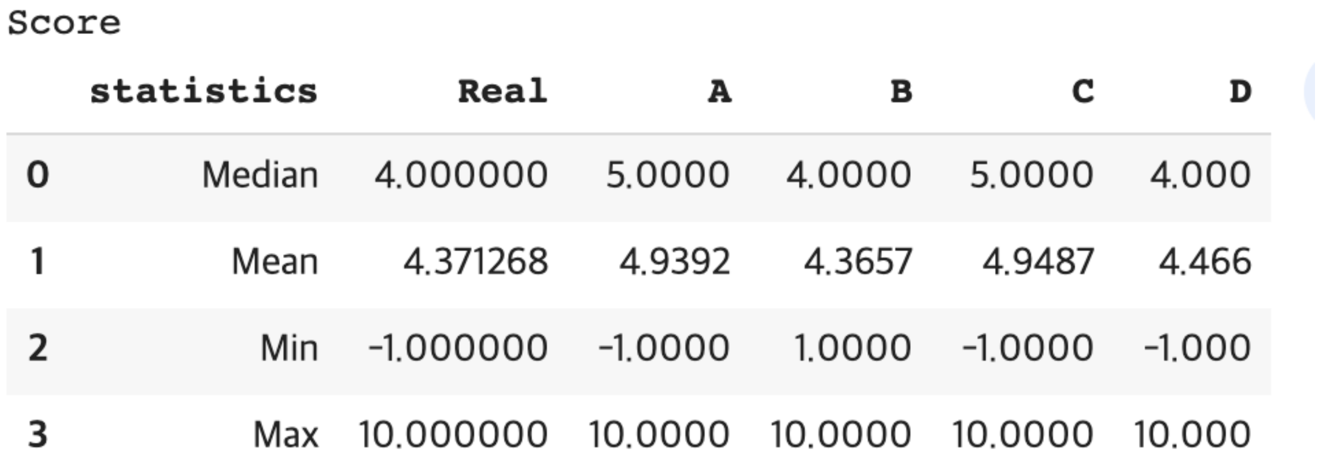
For the attribute score, the result is similar to the age. The method B is the most accurate, because Median, Mean, and Max shows similar value while Min has some difference. For the Min, considering that min is the extreme value, B shows the most accurate compared to other methods. Next C, D, and A, because C, D, A have same values for min and max values and the difference is median and mean. When comparing the mean and median values, C, D, A is accurate in order.

Accordingly, I can say that the accuracy of these methods are as follow.
B > C > D > A



Age

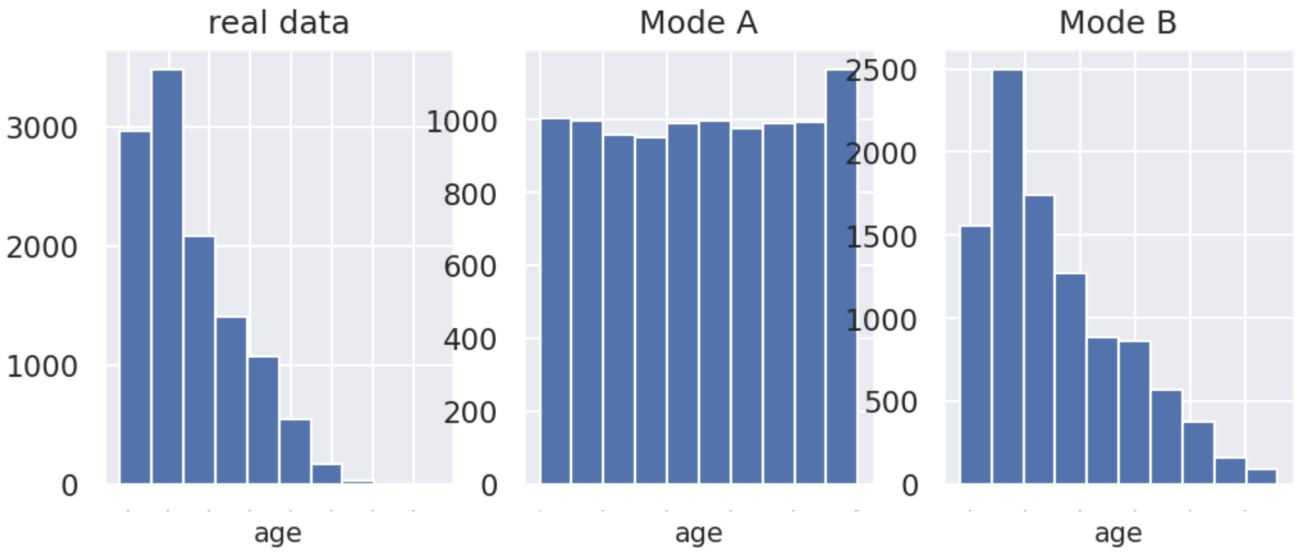
	statistics	Real	A	B	C	D
0	Median	32.000000	51.0000	33.0000	36.0000	39.0000
1	Mean	35.143319	50.1731	35.7354	41.5788	44.1532
2	Min	18.000000	0.0000	18.0000	18.0000	18.0000
3	Max	96.000000	100.0000	76.0000	96.0000	96.0000

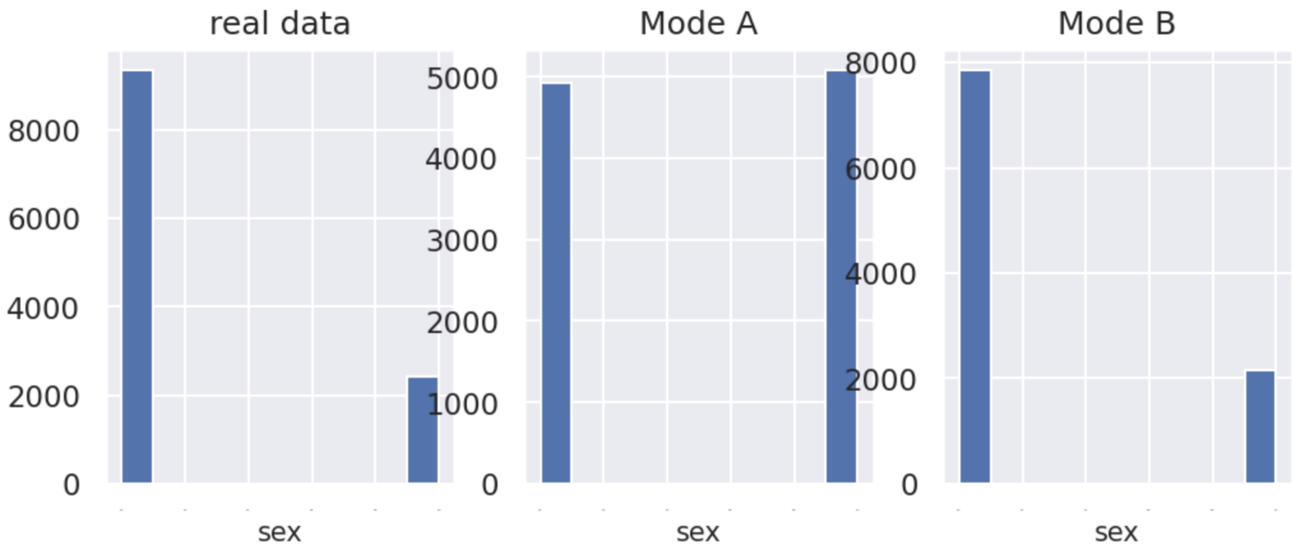


Score

	statistics	Real	A	B	C	D
0	Median	4.000000	5.0000	4.0000	5.0000	4.000
1	Mean	4.371268	4.9392	4.3657	4.9487	4.466
2	Min	-1.000000	-1.0000	1.0000	-1.0000	-1.000
3	Max	10.000000	10.0000	10.0000	10.0000	10.000

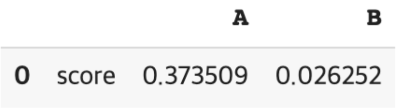
Q2.





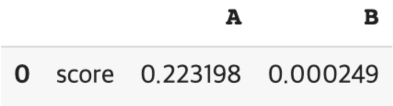
Comparing the histogram of Mode A and Mode B with real data, Mode B is more accurate than Mode A. For the attributes age and sex, the distribution of data is similar for the Mode B, while distribution of Mode A has a lot in difference when compared to the distribution of the real data. This is because Mode B uses independent attribute mode which preserves the distribution of the data, while Mode A uses Random mode.

ks test:



		A	B
0	score	0.373509	0.026252

kl test:

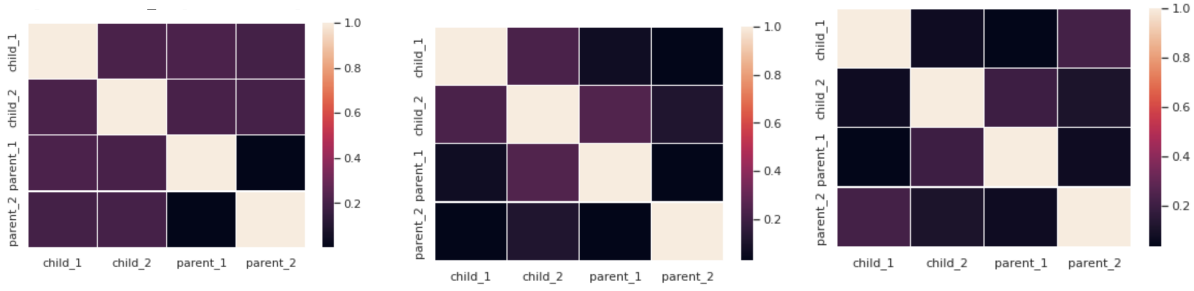


		A	B
0	score	0.223198	0.000249

Since the attribute 'age' is numerical attribute, I used KS test for this attribute. For the KS test, if the value is smaller than 0.05 we can not reject the null hypothesis and this means two data sets are similar. For the method B has value of 0.026252 which is smaller than 0.5, I can conclude that real data and data from Random Mode B is similar. However, the real data and data from Random Mode A are different because the KS test result is 0.373509

KL test shows the difference between two data, and since the attribute 'sex' is categorical data, I used KL test for the attribute 'sex'. For the KL test, the lower the value is, the similar two data are. In this table, method A has value of 0.223198 and method B has value of 0.000249 which is much smaller. This means Mode B is much similar than Random Mode A.

Q3. fake, C, D in order

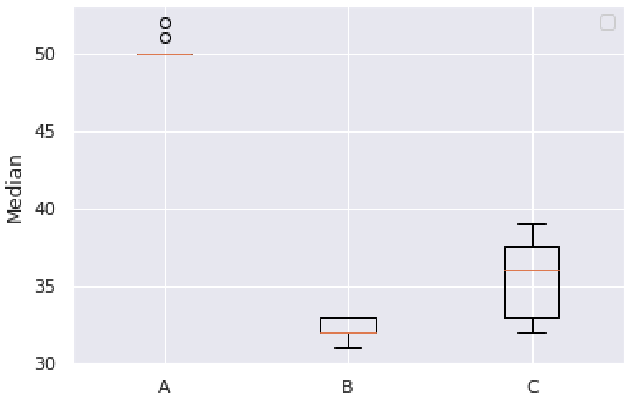


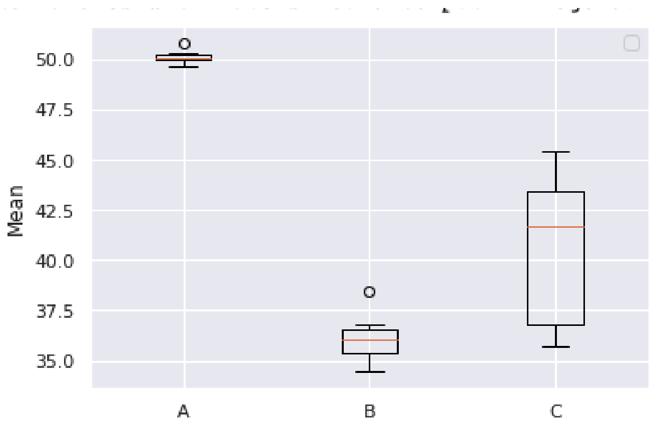
When I see the heatmaps of fake data, C, and D, it becomes darker(closer to 0) overall, which means the correlation becomes lower. Considering that C and D uses correlate attribute mode with only difference of parameter value k, k(:the number of parents a node can have) affects the result.

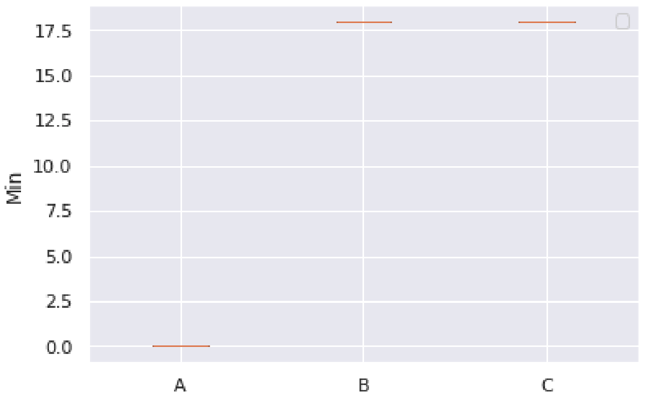
child1 and child2: becomes less correlation using correlate attribute mode with k=1 and k=2
child1 and parent1: becomes less correlation using correlate attribute mode with k=1 and k=2

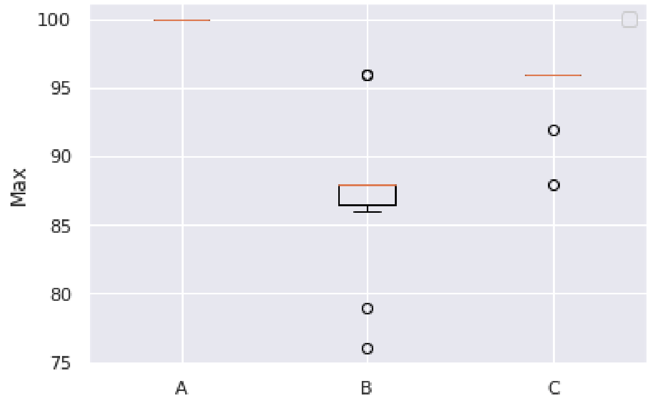
child1 and parent2: becomes less correlation using correlate attribute mode with k=1 and more correlation with k=2
child2 and parent1: correlation is similar
child2 and parent2: becomes less correlation using correlate attribute mode with k=1 and k=2
parent1 and parent2: always close to 0 which means there less correlation.

4.2 (b)









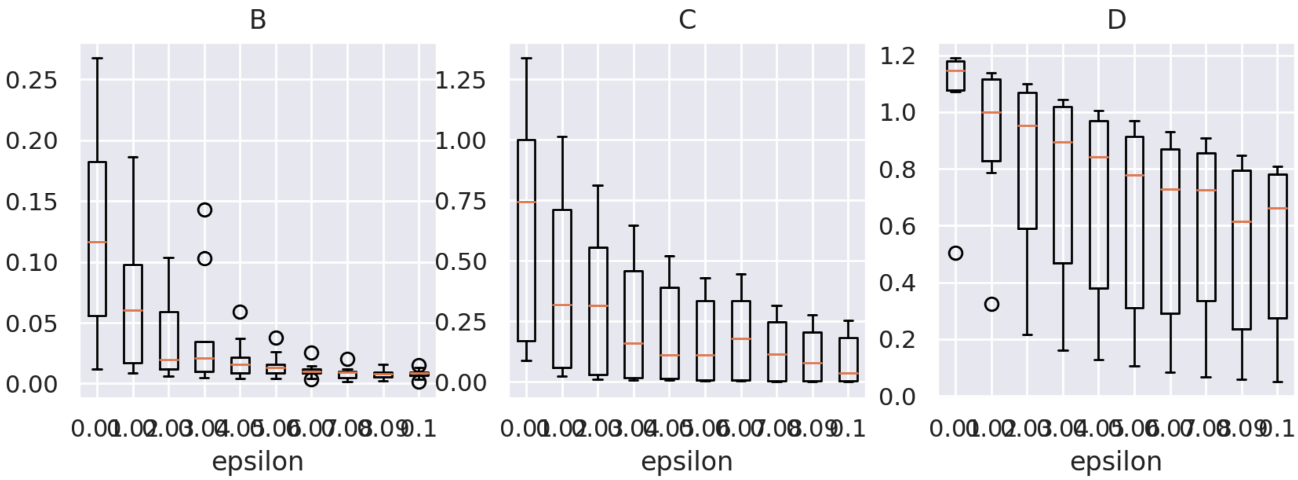
Accuracy: B > C > A
Variability: C > B > A

For the accuracy, considering that Median: 32 Mean: 35.143319 Min: 18 Max: 96 for the real data, data B is the most accurate. In the plot, Median values and Mean values are around 32 and 35 respectively for the B. Min values of B is the same which is 18, and max values are spread but still have some values around 96 for B. The next is C, which spread around Median, Mean, Min, and Max around those of real data. For the A, the values are more different when compare to B and C.

For the variability, C shows various values for Median, Mean, Min, and Max as in the box plot. The values from 10 data sets spread as in the box plot. B also spread a little, but A has barely spread of Median, Mean, Min, and Max, which means less variability.

4.3 (c)

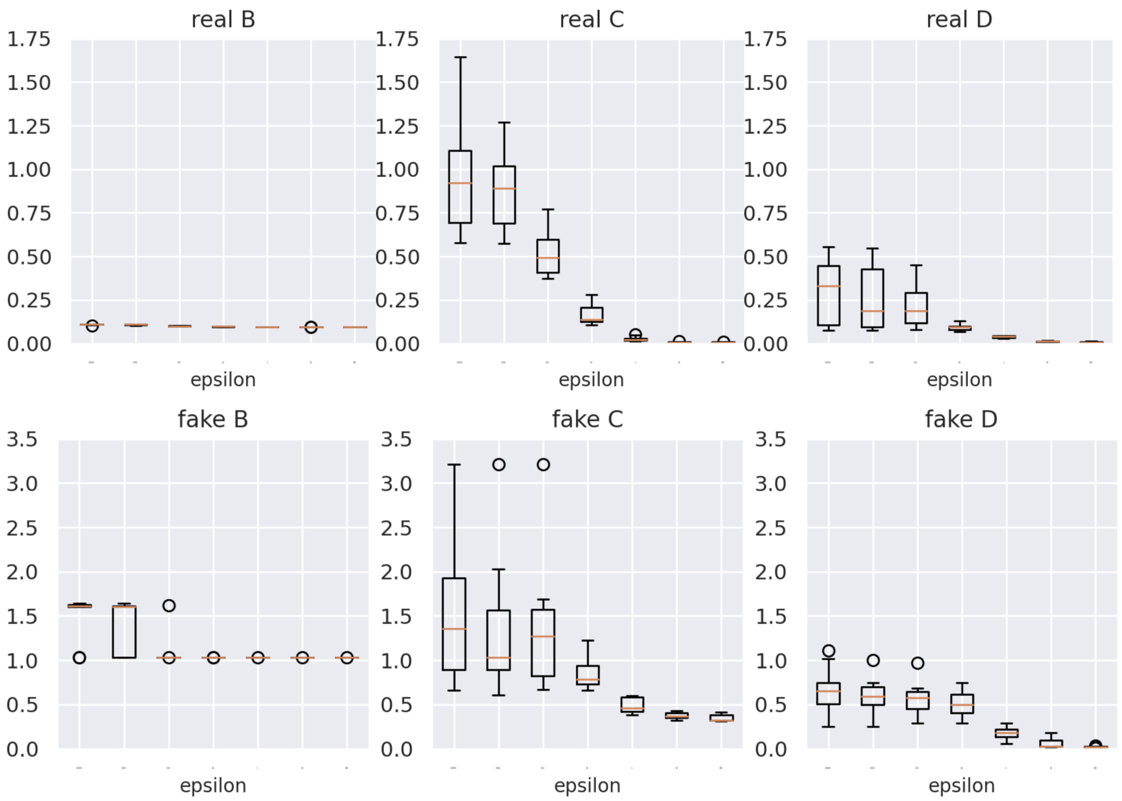
1. KL-divergence



From all 3 plots, as epsilon value increases, the values of KL-divergence decrease, which means two data(real data and generated by Synthesizer) are similar. This is because when epsilon value increases, it reduces noise which means two data sets are similar.

When compared to B, C, and D, B is more similar compared to C and D. This is because B used independent attribute mode which preserves the distribution, while C and D used correlate attribute mode which preserves correlations.

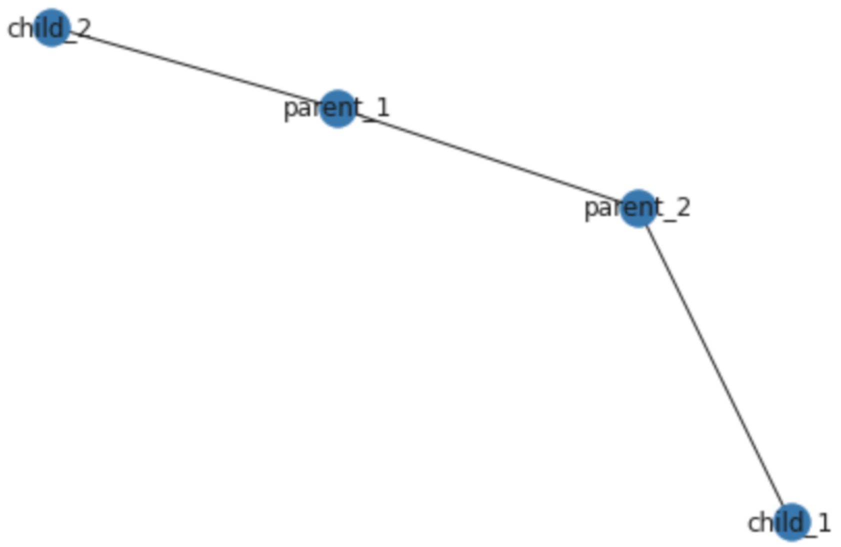
2. Difference in pairwise mutual information



For the real data, real B is the most similar because the difference of mutual information is almost 0. Next, real D which uses correlate attribute mode with Bayes network degree 2 is more similar than that of real C.

For the fake data, fake D is the most similar because the difference of mutual information is the lowest. Next, fake B which uses independent attribute mode is more similar than that of fake C. The reason why the fake D is the most similar is that fake D is generated by correlate attribute mode which preserves correlation.

4.4 (d)

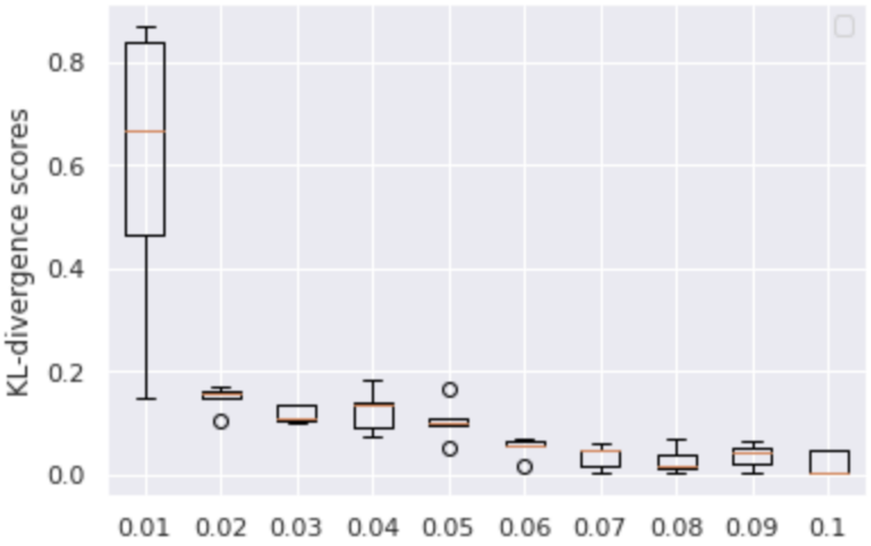


In the MST Synthesizer, the marginal can be selected manually by the domain expert or automatically, and in this problem the marginal is selected automatically.

For the MST Synthesizer, when we use the Bayesian Network, The shape of the graph should be directed and acyclic and there should be one marginal per attribute. Above graph is acyclic and can go in one direction. For example, if I start to traverse from child_2 and goes to parent_1, parent_2, child_1 in order, I would not go back to previously visited node. And in Bayesian Networks, the parameters are estimated by truncating noisy marginals to 0 and normalizing to obtain conditional probability tables.

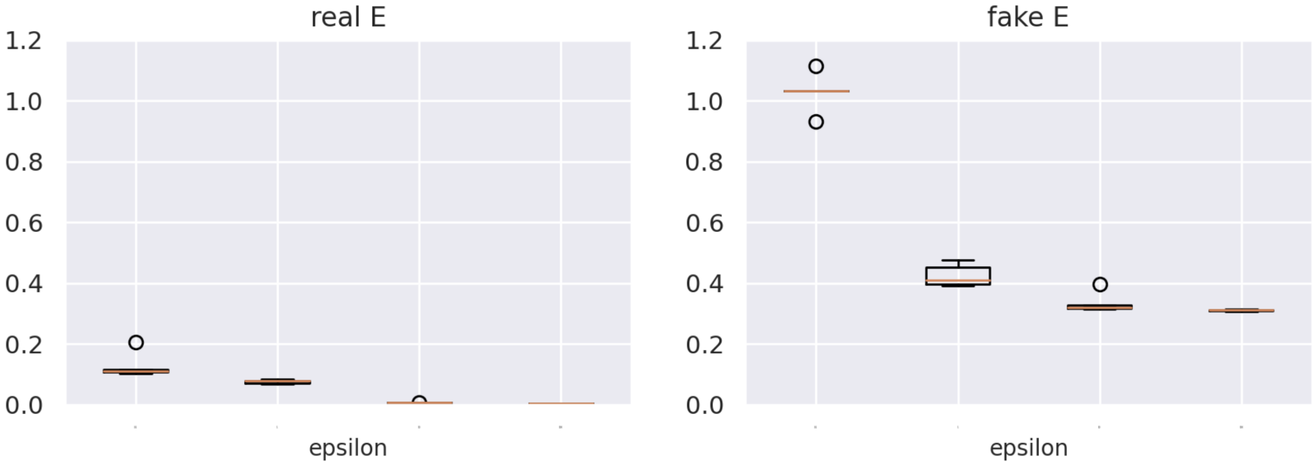
4.5 (e)

1. condition E



As epsilon value increases, the value of KL-divergence decrease, which means 'race' in two data(real data and generated by MST Synthesizer) are similar. This is because when epsilon value increases, it reduces noise which means two data sets are similar. And the KL-divergence scores are generally low in this plot because it uses MST synthesizer.

2.



* Run epsilons = [0.1, 1, 10, 100], because of error mentioned in Piazza.

Compared to plots in (c) for the method D, method E is better in that method E is more similar compared to original data. When comparing KL-divergence values, method E has smaller value (though not a big difference) which means data that is generated by MST Synthesizer is more similar to original data. Especially for the the plot real E shows that the difference is close to 0 and lower than 0.2. (There was no big difference for the fake D and fake E) This is because Data Synthesizer distributes equal amount of epsilon values to each, while MST Synthesizer distributes epsilon values considering the data. This difference leads to better performance for the MST Synthesizer.