1

2 Features correlations

2.1 10 features which are most correlated with Y

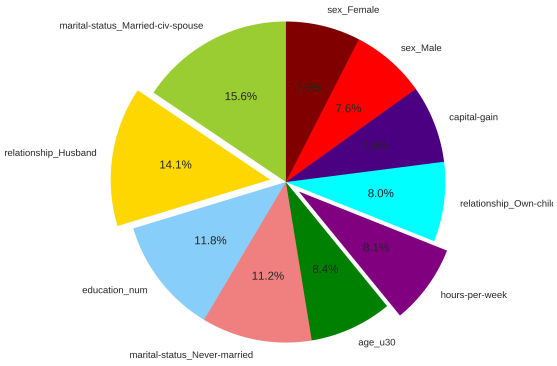


Figure 1: 10 features which are most correlated with Y

2.2 10 features which are most correlated with A

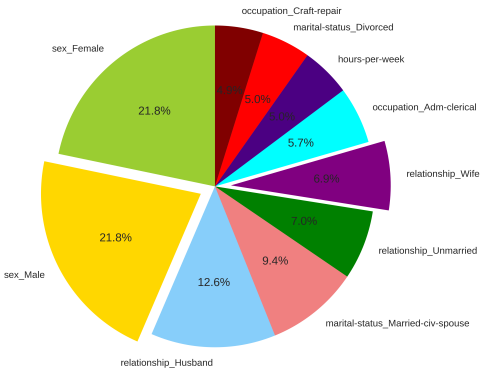


Figure 2: 10 features which are most correlated with A

3

3.1 Classification accuracy and Δ_DP on the test set for the trained classifier

Accuracy	0.77753
Δ_DP	0.11278

3.2 Classification accuracy and Δ_DP after removing the 10 attributes

Accuracy	0.80406
ΔDP	0.1183

Table 1: Accuracy and Δ_DP after removing 10 attribute

3.3 Which three features in the data are most correlated with $\hat{Y}$

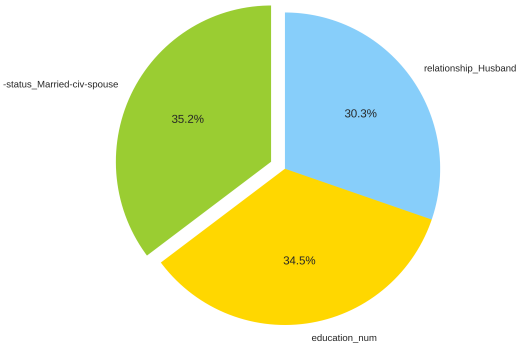
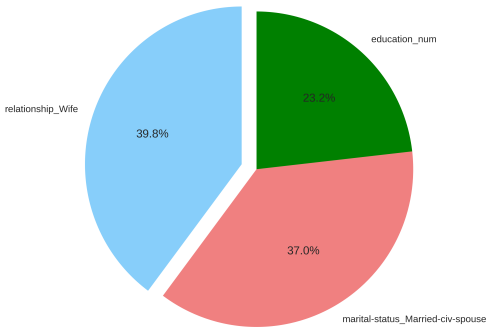
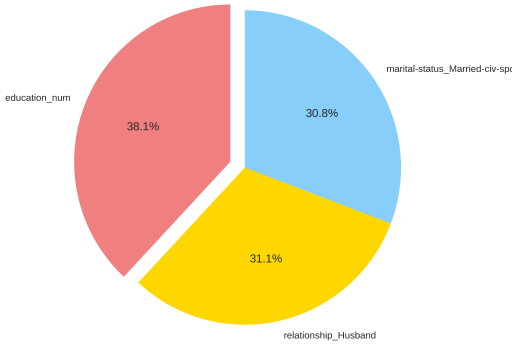


Figure 3: 3 features which are most correlated with $\hat{Y}$

3.4 Which three features are most correlated with $\hat{Y}$, only looking at examples where A = 0?



3.5 Which three features are most correlated with $\hat{Y}$, only looking at examples where A = 1?



3.6 Classification accuracy and re-weighted accuracy when predicting A with sex_Male and sex_female removed

Accuracy	0.762053
reweighted accuracy	0.75793

Table 2: Accuracy and reweighted accuracy after removing sex_Male and sex_female

3.7 Accuracy and reweighted accuracy after removing the 10 most correlated features with A

Accuracy	0.64996
reweighted accuracy	0.53570

Table 3: Accuracy and reweighted accuracy after removing 10 attribute

4 Representation Learning

4.1 Data Pre-processing using Gaussian distribution

4.1.1 Predicting Y

Accuracy	0.822308
Δ_DP	0.145205

Table 4: Predicting Y with the Gaussian normalized data

4.1.2 Predicting A

Accuracy	0.999631
re_weighted accuracy	0.99972

Table 5: Predicting A with the Gaussian normalized data

	Y	A
Accuracy without Gaussian normalization	0.77753	0.762053
Accuracy with Gaussian normalization	0.822308	0.99963

Table 6: Comparison between preprossing and without preprocessing

Normalizing the data by a Gaussian distribution make a big change on the accuracy for predicting $\hat{Y}$ and A

4.2 MMD Part: Which method was better

alpha	Classifier g (Y)	Classifier h (A)
0.0001	0.8261777532348893	0.9992629445365764
0.001	0.8324427246410412	0.9990786806707205
0.01	0.8300472943592874	0.9990172593821018
0.1	0.8267919661210756	0.9992629445365764
1	0.8345310484870236	0.999324365825195

Table 7: Training with MMD regularizer and different values for alpha

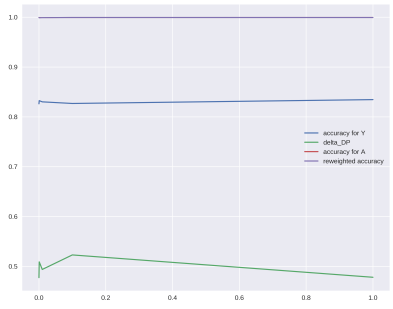


Figure 4: MMD classifiers

4.3 We compared two methods of removing sensitive information. What do you think is the best way to compare various methods for this task?

The objective of a classifier is to achieve a higher accuracy model while ensuring that the models are lesser discriminant in relation to sensitive/protected attributes. Therefore, the quality of the classifier is measured by its accuracy and the discrimination it makes on the basis of sensitive attributes; the more accurate, the better, and the less discriminant (based on sensitive attributes), the better.

4.4 Can you think of any other ways you might remove information about A from a representation?

The Gaussian normalization is a pre-processing and the MMD a in-processing. We can also use a post-processing. We have different approaches of post-processing:

- Equalized odds post-processing : The algorithm solves a linear program to find probabilities with which to change output labels to optimize equalized odds.
- Reject option classification: The idea is to give favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.
- Calibrated equalized odds post-processing: The algorithm optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.