Learning Fairness from an Adversary

June 18, 2015

The Adversarial Approach

- Many machine learning techniques can be reduced to finding parameters $\hat{\theta}$ for a model $M(\theta)$ solving the minimization problem

$$\min_{\theta} L(\theta),$$

where L is some loss function.

The Adversarial Approach

- Many machine learning techniques can be reduced to finding parameters $\hat{\theta}$ for a model $M(\theta)$ solving the minimization problem

$$\min_{\theta} L(\theta),$$

where L is some loss function.

- In an adversarial setting you also have a second model $A(\phi)$ and instead you solve the problem

$$\min_{\theta} \max_{\phi} L(\theta, \phi),$$

where the adversary opposes you (the model) by trying to maximize the loss function.

Generative Adversarial Networks

- Goodfellow *Generative Adversarial Networks* (2014).

Generative Adversarial Networks

- Goodfellow *Generative Adversarial Networks* (2014).
- The idea is that you learn a transformation f from a source distribution S so that it is indistinguishable from a target distribution T.

Generative Adversarial Networks

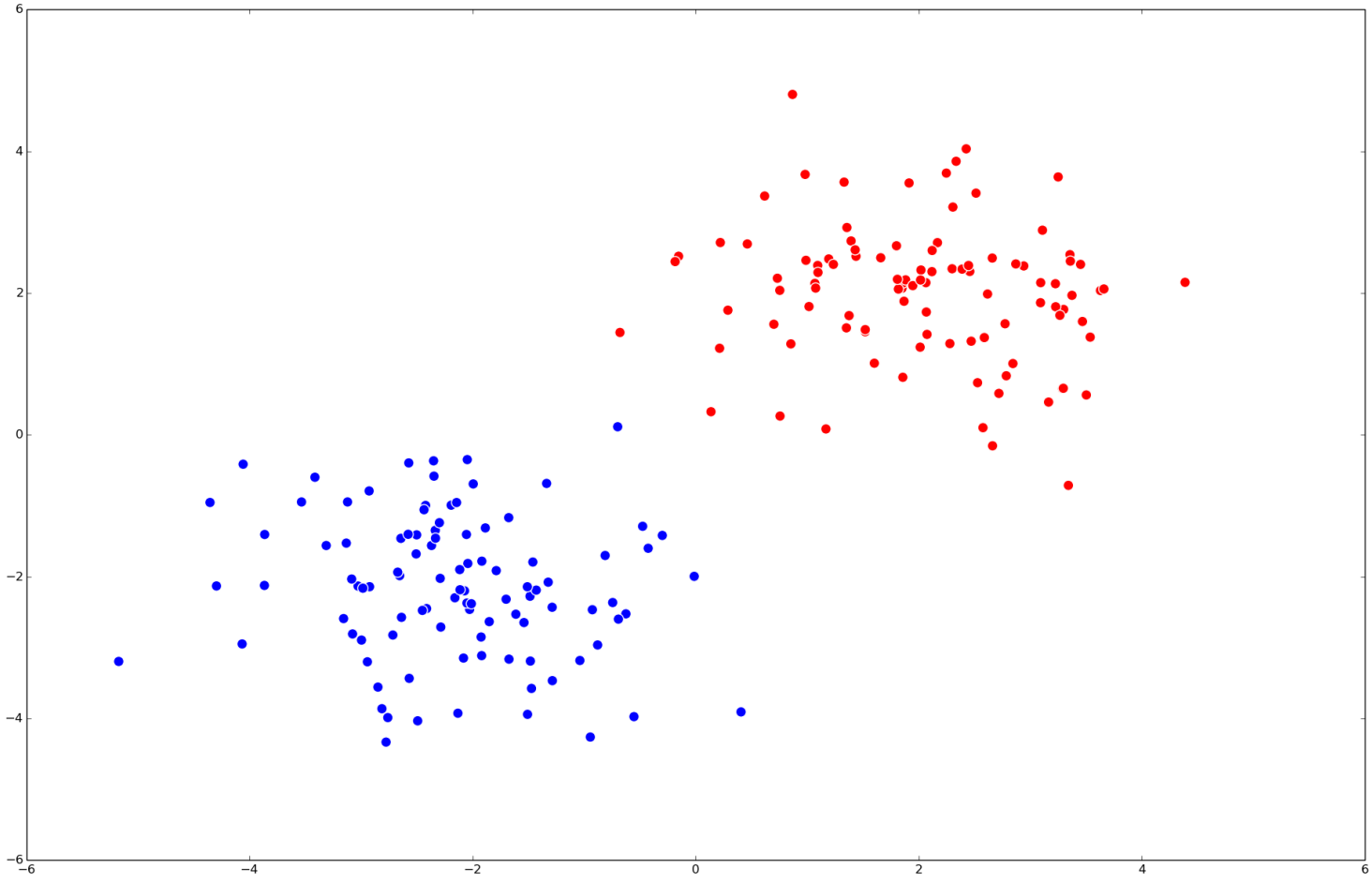
- Goodfellow *Generative Adversarial Networks* (2014).
- The idea is that you learn a transformation f from a source distribution S so that it is indistinguishable from a target distribution T.
- Once you have done this, if you can sample from S, you can sample from T.

Generative Adversarial Networks

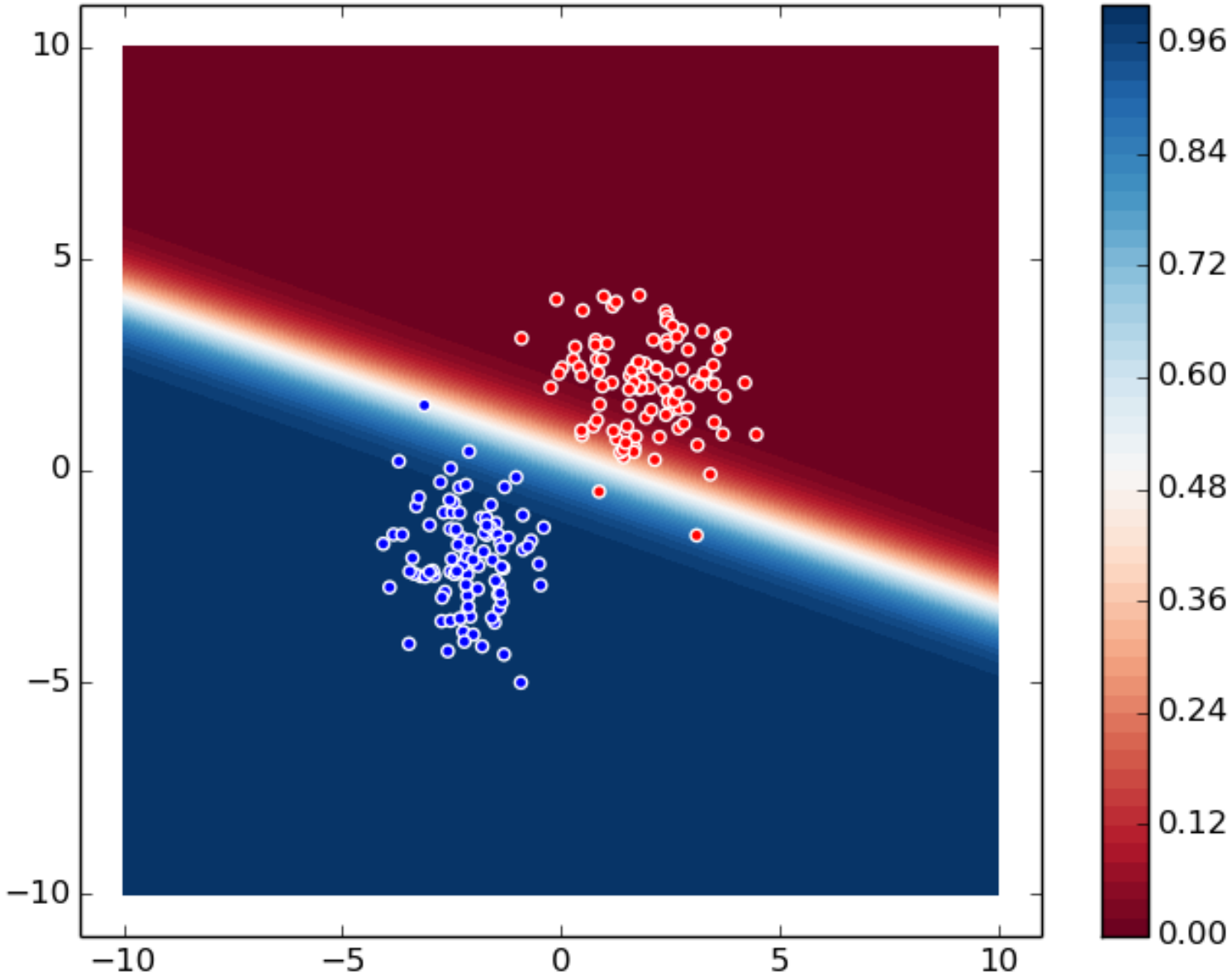
- Goodfellow *Generative Adversarial Networks* (2014).
- The idea is that you learn a transformation f from a source distribution S so that it is indistinguishable from a target distribution T.
- Once you have done this, if you can sample from S, you can sample from T.
- In this case the adversary is a classifier that tries to predict whether a sample is from T or $f(S)$

An Example

In our example we will try to make the red (source) data look like the blue (target) data, and the adversary will be a logistic regressor.



Example Cont'd



Fairness

- Back in 2006, the world looked like this.

Fairness

- Back in 2006, the world looked like this.
- But not everyone was happy: in 2011 the European Court of Justice ruled that "the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits".

Fairness

- Back in 2006, the world looked like this.
- But not everyone was happy: in 2011 the European Court of Justice ruled that "the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits".
- So we ask the question: how can we make machine learning algorithms whose decisions are *independent* of a given protected variable, like gender?

A Fair Classifier

Given features X, a binary protected variable D and binary targets Y, a *fair classifer* for Y with respect to D is a mapping $f : X \to [0, 1]$ where $f(x) = P_f(Y = 1|x)$ such that

$$P_f(Y = 1|D = 1) = P_f(Y = 1|D = 0).$$

We measure this using the *discrimination*:

$$|P_f(Y = 1|D = 1) - P_f(Y = 1|D = 0)|,$$

which is empirically estimated by

$$\left| \frac{1}{\#i : d_i = 1} \sum_{i:d_i=1} f(x_i) - \frac{1}{\#i : d_i = 0} \sum_{i:d_i=0} f(x_i) \right|.$$

Fair Representations Make a Fair Classifier

- In 2013 Richard Zemel *et al* had the idea in "Learning Fair Representations", that instead of penalizing a classifier with its discrimination, you could instead learn intermediate representations for your data which are fair, so that a classifier trained on this representation automatically be fair.

Fair Representations Make a Fair Classifier

- In 2013 Richard Zemel *et al* had the idea in "Learning Fair Representations", that instead of penalizing a classifier with its discrimination, you could instead learn intermediate representations for your data which are fair, so that a classifier trained on this representation automatically be fair.
- The intermediate representations are clusters $Z \in \{0 \dots m\}$ and each input x is mapped to a probability distribution over the clusters, so that cluster assignment is 'fair' that is

$$P(Z = k|D = 1) = P(Z = k|D = 0),$$

for each $k \in \{0 \dots m\}$.

Fair Representations Make a Fair Classifier

- In 2013 Richard Zemel *et al* had the idea in "Learning Fair Representations", that instead of penalizing a classifier with its discrimination, you could instead learn intermediate representations for your data which are fair, so that a classifier trained on this representation automatically be fair.
- The intermediate representations are clusters $Z \in \{0 \dots m\}$ and each input x is mapped to a probability distribution over the clusters, so that cluster assignment is 'fair' that is

$$P(Z = k|D = 1) = P(Z = k|D = 0),$$

  for each $k \in \{0 \dots m\}$.
- The classifier then makes decisions like:

$$P(Y = 1|x) = \sum_k P(Y = 1|Z = k)P(Z = k|x),$$

  and it follows that the classifier will be fair so long as the cluster assignment is fair.

Getting Adversarial

- Instead of using a fair clustering approach, one could instead learn a distributed intermediate representation $E : X \to H$ such that $E(X) \perp\!\!\!\perp D$.

Getting Adversarial

- Instead of using a fair clustering approach, one could instead learn a distributed intermediate representation $E : X \to H$ such that $E(X) \perp\!\!\!\perp D$.
- To do this we have an adversary try to predict D from H, and use this to learn H so that it is independent of D but still useful for predicting Y.

Getting Adversarial

- Instead of using a fair clustering approach, one could instead learn a distributed intermediate representation $E : X \to H$ such that $E(X) \perp\!\!\!\perp D$.
- To do this we have an adversary try to predict D from H, and use this to learn H so that it is independent of D but still useful for predicting Y.
- In both methods, there is a tradeoff between fairness and accuracy controlled by weighting parameters.

Experiments

- To compare the clustering approach with the adversarial one, I used the Adult data set from the *UCI Machine Learning* repository.

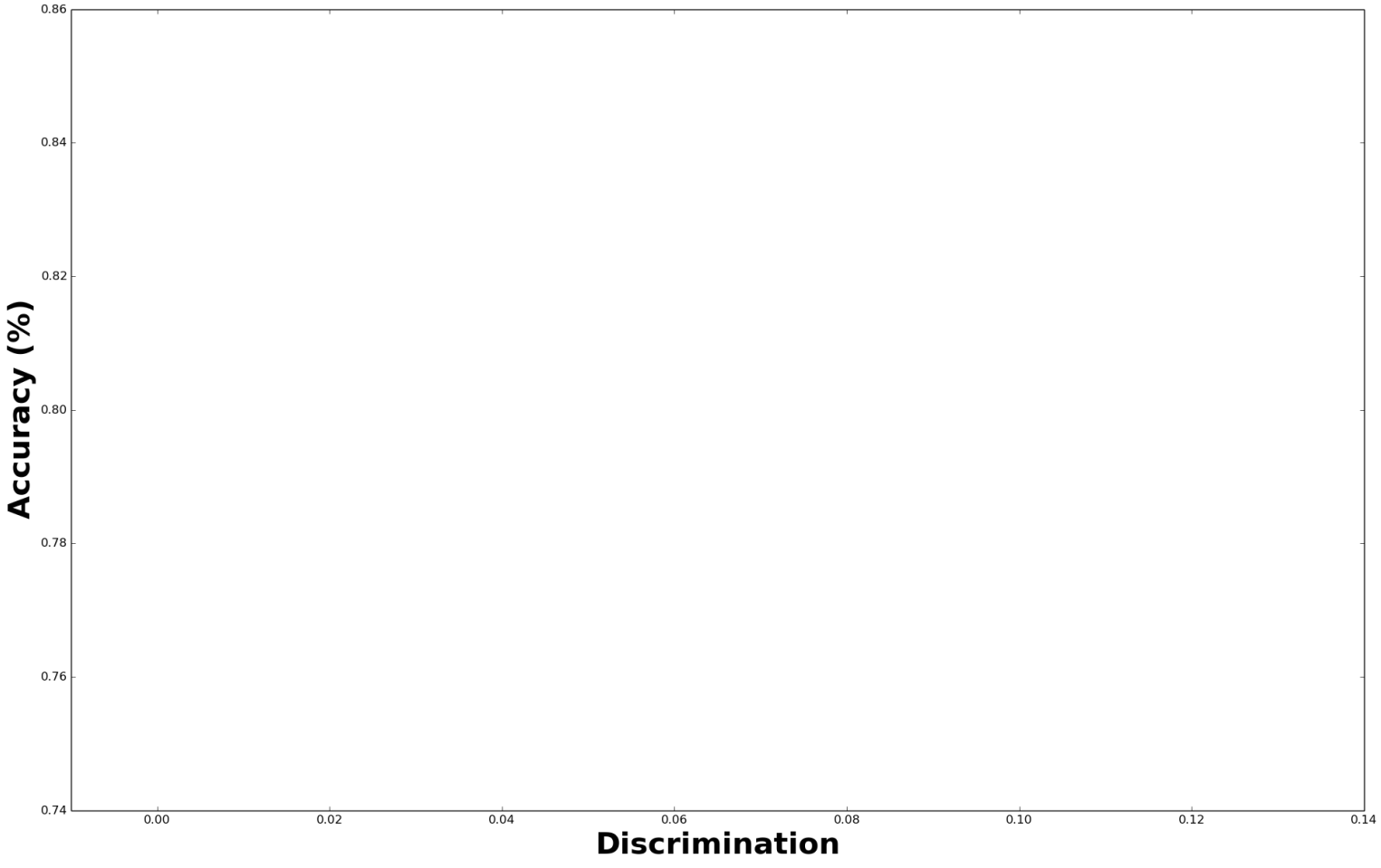
Experiments

- To compare the clustering approach with the adversarial one, I used the Adult data set from the *UCI Machine Learning* repository.
- The task is to predict whether a person's salary is above or below $\$50K$ and the protected variable is sex.

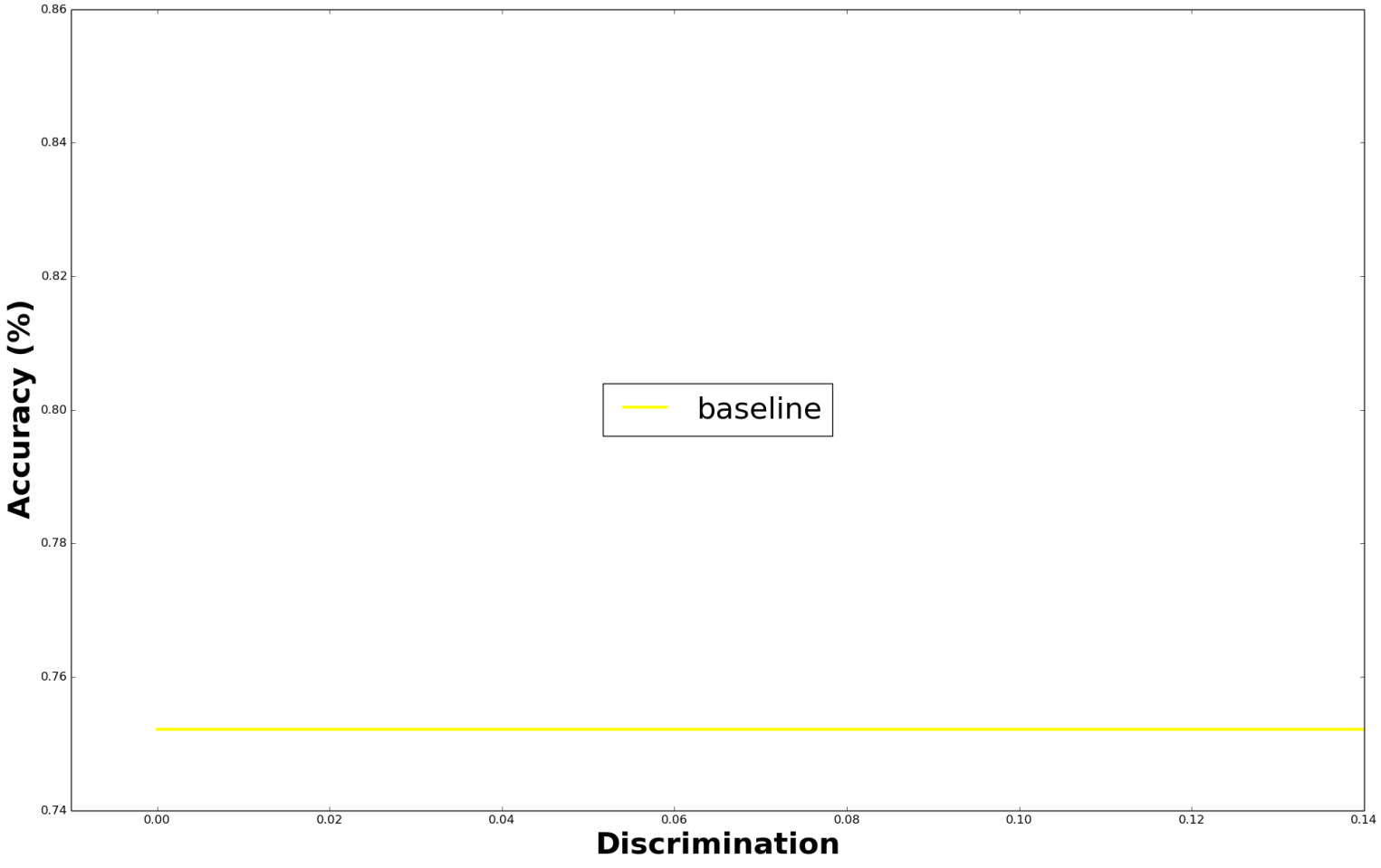
Experiments

- To compare the clustering approach with the adversarial one, I used the Adult data set from the *UCI Machine Learning* repository.
- The task is to predict whether a person's salary is above or below $\$50K$ and the protected variable is sex.
- The data set has around $50K$ entries and 100 features.

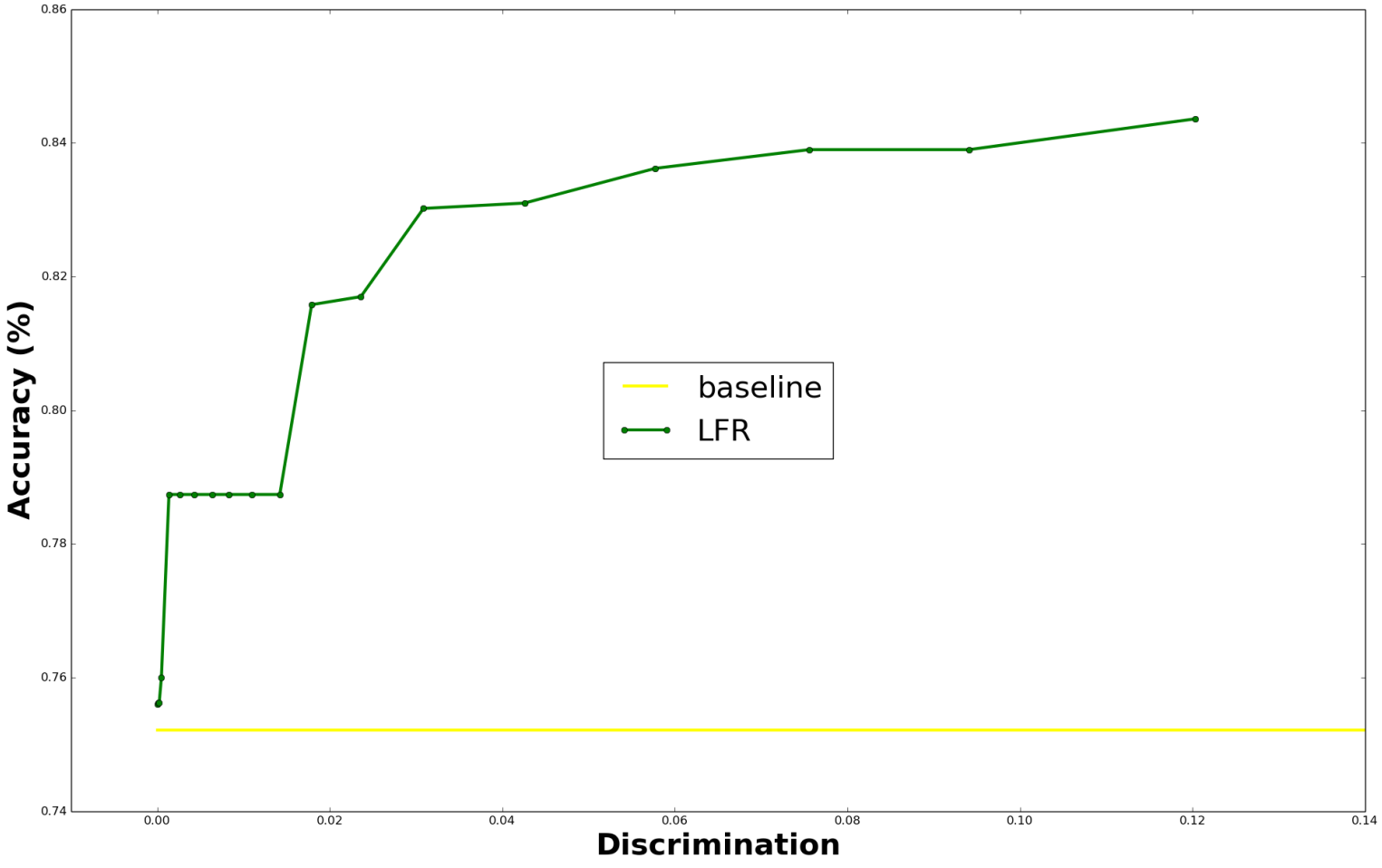
Results


0.86
0.84
0.82
0.80
0.78
0.76
0.74
Accuracy (%)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Discrimination

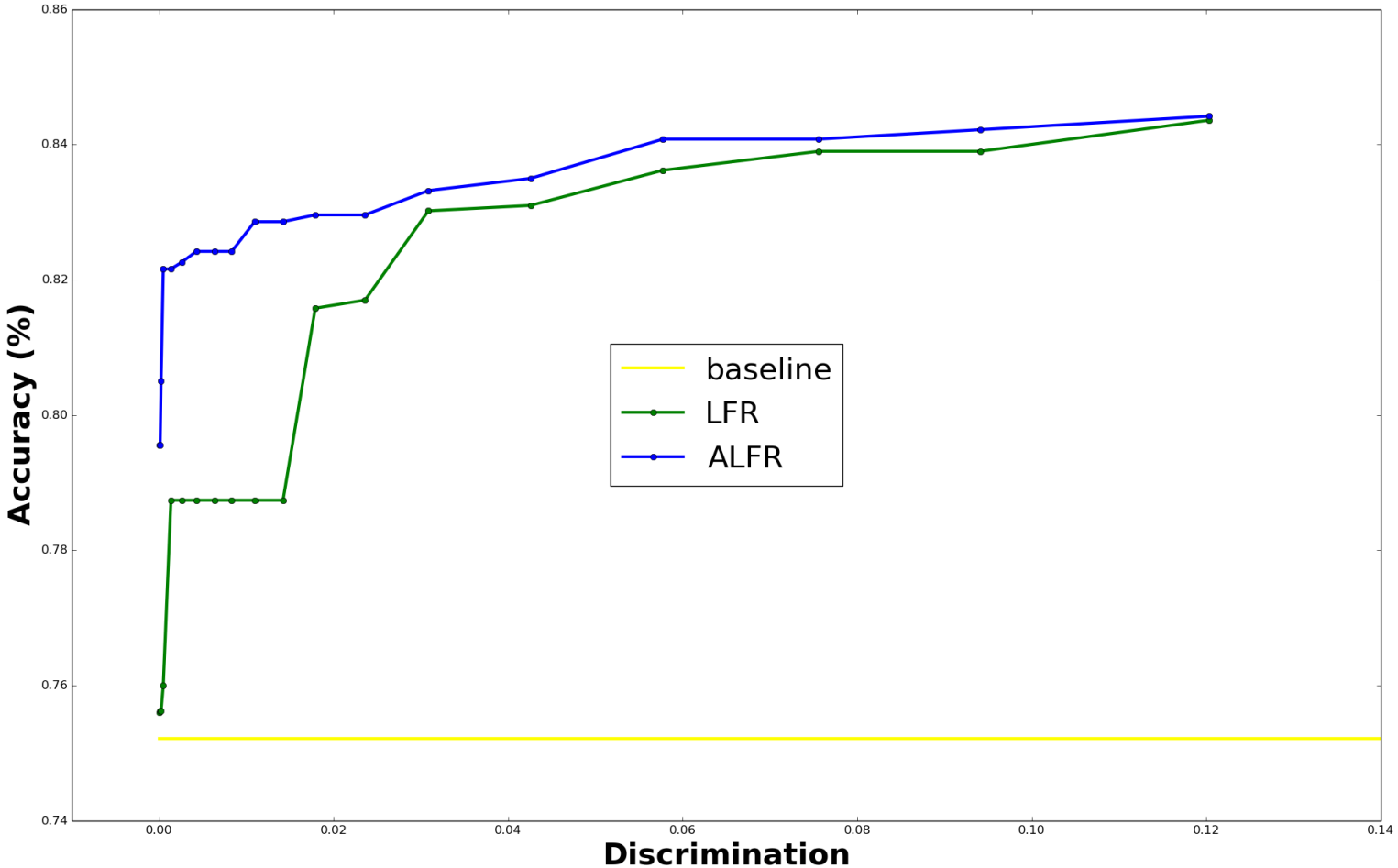
Results


0.86
0.84
0.82
0.80
0.78
0.76
0.74
Accuracy (%)
baseline
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Discrimination

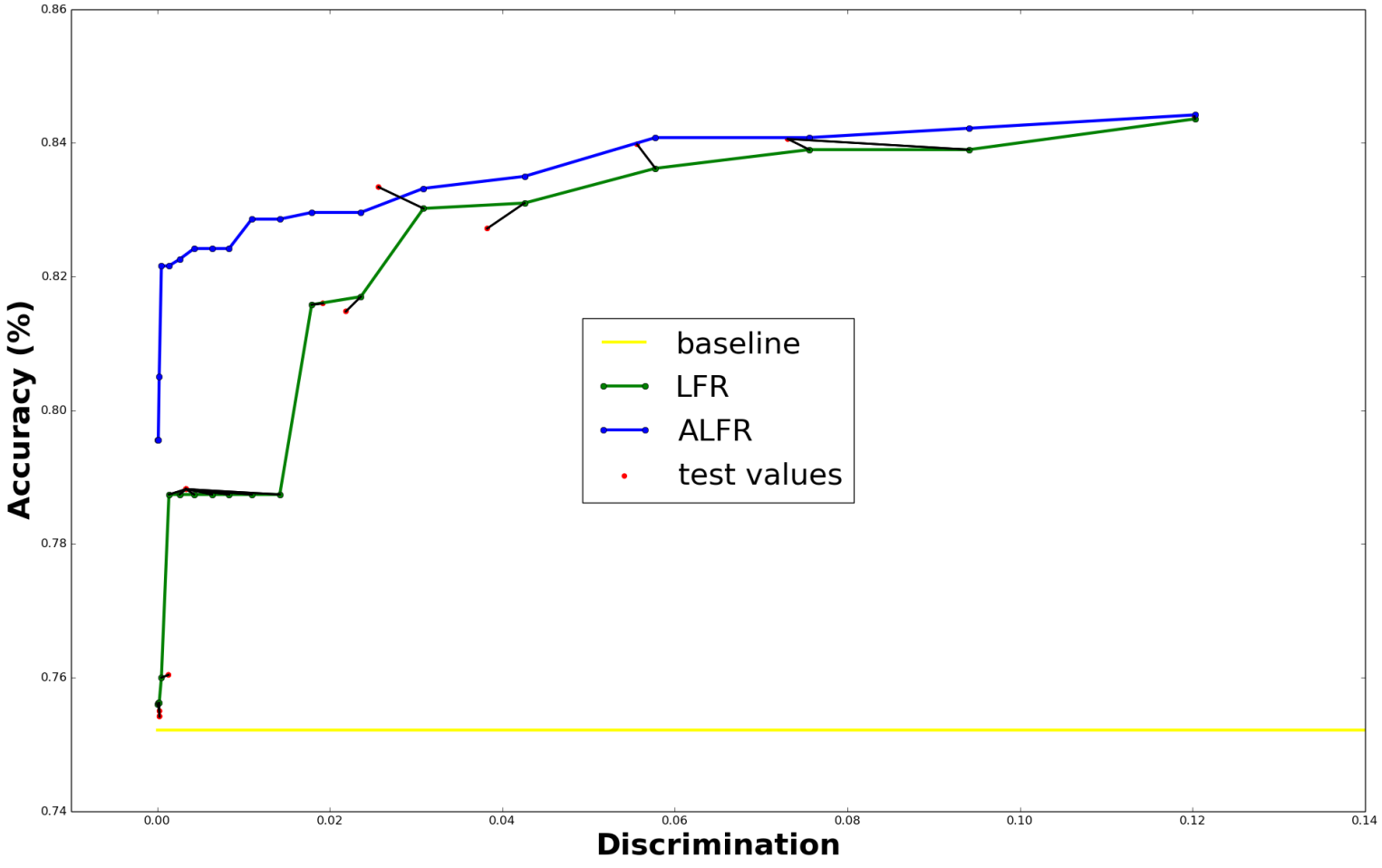
Results


0.86
0.84
0.82
0.80
0.78
0.76
0.74
Accuracy (%)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Discrimination
baseline
LFR

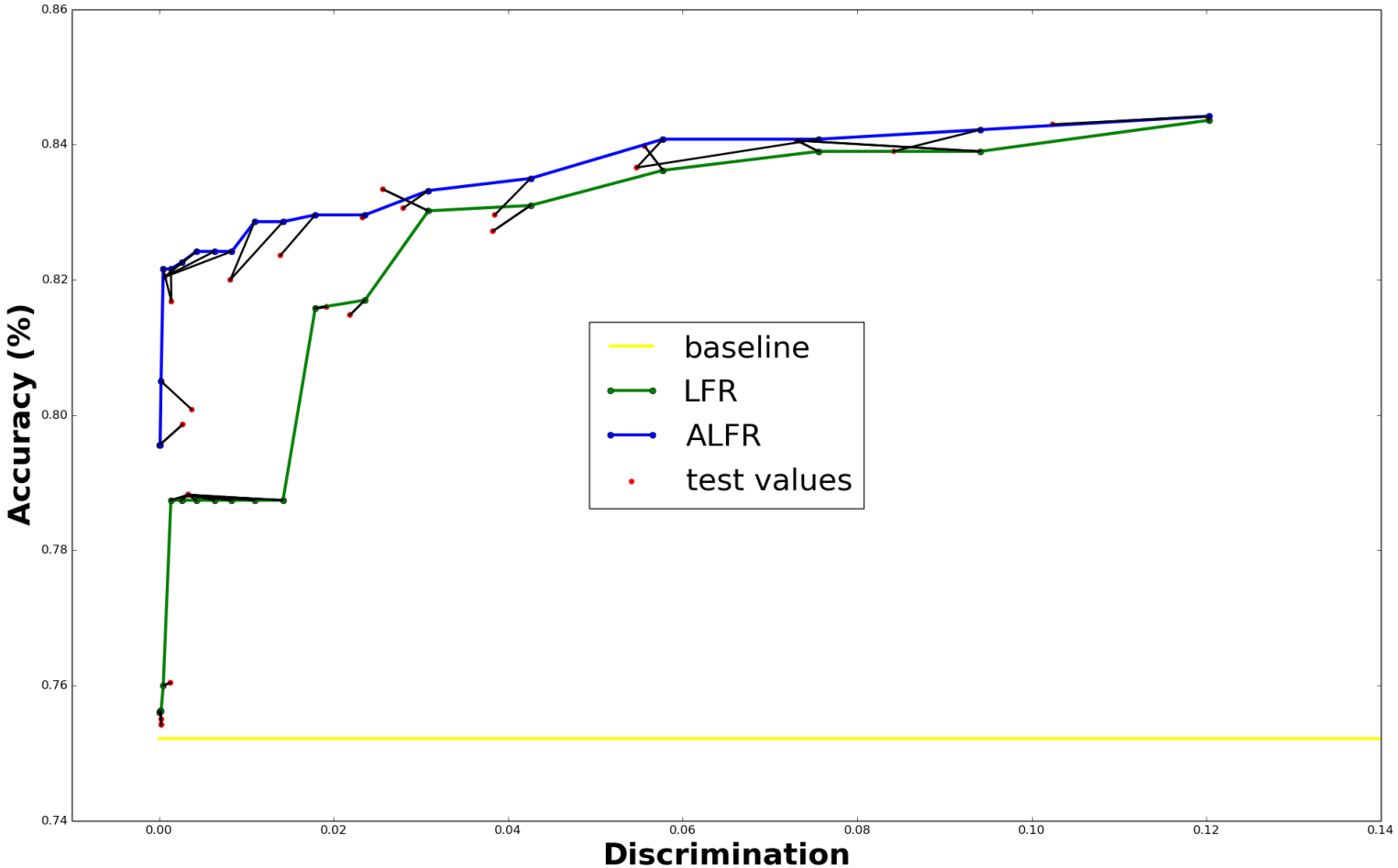
Results


0.86
0.84
0.82
0.80
0.78
0.76
0.74
Accuracy (%)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Discrimination
baseline
LFR
ALFR

Results


0.86
0.84
0.82
0.80
0.78
0.76
0.74
Accuracy (%)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Discrimination
baseline
LFR
ALFR
test values

Results


0.86
0.84
0.82
0.80
0.78
0.76
0.74
Accuracy (%)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Discrimination
baseline
LFR
ALFR
test values