Fairness

April 20, 2016; due May 4, 2016

1 (everyone)

Several applications of predictive methods are prohibited by law, regulatory means, or ethical considerations from using protected features in making classification decisions; examples of this include disallowing gender or ethnicity from influencing assessments of credit risk (under the Equal Credit Opportunities Act) or employment opportunities.

In this homework, your task is to design an algorithm that maximizes predictive accuracy while ensuring that the final classification decisions are not biased toward the protected feature—in our definition of "fair" here, the correlation between the protected features and your predictions must be less than 0.20. You have considerable flexibility in designing this algorithm, including how you choose to represent each data point from among the available seven features and what classification algorithm you decide to use.

protected feature	feat 1	feat 2	feat 3	feat 4	feat 5	feat 6	y
1	1	0	0	0	1	1	1
1	1	1	0	0	1	0	1
1	1	1	0	1	1	1	1
1	0	1	1	1	0	0	1
1	0	1	0	1	1	1	1
0	1	0	1	0	1	1	1
0	0	0	1	1	0	0	0
0	0	1	1	0	1	1	0
0	1	0	1	0	0	1	0
0	0	0	0	1	1	1	1
1	1	0	0	1	1	0	1
1	1	1	0	0	0	0	0
0	0	1	1	0	1	1	0
1	0	0	0	1	0	1	1
0	1	1	1	1	1	1	0

Table 1: Data containing one protected feature and six unprotected features

Deliverable: one-half page paper (single-spaced).

2 (choose either 2.1 or 2.2)

2.1 Implementation

Implement that algorithm using the data above (also provided on the GitHub repository as a .tsv file[1]). What is your cross-validated classification accuracy, and what is the correlation between the protected feature and your prediction?

Deliverable: one paragraph containing your written answers to the questions above, along with code to support it.

[1] `https://github.com/dbamman/dds/blob/master/homework/hw5/data/data.txt`

2.2 Critique

Ensuring that predictions are agnostic to protected features is only one aspect of fairness in data science. As we've been discussing all semester, bias of many forms can creep in at each stage of analysis, including the choice of data (and its representation), the questions that we set out to ask, and the models we choose to apply. Detail **four** specific ways in which the choices we make in data science design can lead to discriminatory or otherwise biased outcomes, and steps that we can take to counter them.

Deliverable: 1.5-page paper (single-spaced).