# *Assignment 3: Fairness, Maximum Likelihood Estimation, and Text Classification*

*Machine Learning*

*Fall 2019*

**Learning Objectives**

- TODO

## *1 Bayesian Networks and Algorithmic Fairness*

In assignment 1 of this module we discussed how Bayesian methods can be used to reason about algorithmic fairness. We've just had some lengthy discussions about fairness within the context of the Compas algorithm. We touched upon some of the limitations of statistically based notions of fairness. Nevertheless, they do have a potential role to play, and you should know what the most common definitions of fairness are and what assumptions they make.

As context for the reading and to help us have common notation, suppose we have the following random variables.

- $R$ is a prediction generated by our algorithm.
- $A$ is a sensitive attribute
- $Y$ is the thing we're trying to predict (we want $R = Y$ if we are predicting accurately)

**External Resource(s) (30 minutes)**

Read Fairness and Machine Learning Chapter 2. Start at the section *Formal non-discrimination criteria* and read up to (but not including) the section *Calibration and sufficiency*.

**Notice**

- Don't get hung up on the ROC curves. We can certainly discuss this on NB, but it is not required to understand what is going on here. The presentation earlier in the linked reading is also pretty clear.
- The notation they use in this reading for conditional independence is an upside down T with only one line (instead of our notation, $\perp\!\!\!\perp$).

**Exercise 1 (5 minutes)**

Thinking back to the COMPAS example, which definition of fairness given in the reading was Propublica using? Which definition of fairness was Northpointe using?

**☆ Solution**

- Northpointe is using sufficiency $Y \perp\!\!\!\perp A \mid R$. You'll notice that in the reading they say that sufficiency is the same thing as matching positive and negative predictive value for all values of the protected attribute.
- Propublica is using separation $R \perp\!\!\!\perp A \mid Y$. This fairness principle requires the false positive and true positive rates to be the same across for all values of the protected attribute.

TODO: possibly present the data from COMPAS in a notebook.

## *2 Text Classification with Bag of Words*

Next we'll be applying Naïve Bayes to the task of classifying text.

**External Resource(s) (45 minutes)**

This will be done in the Assignment 3 companion notebook.

## *3 The Intelligent Design of Jenny Chow*

This assignment is fully described on the Intelligent Design of Jenny Chow Canvas page. There is also an alternative described on the assignment page if you can't attend. Make sure to look at the assignment before going to the play since we are asking you to capture some of your reactions / thoughts so that you can bring them to class on Monday for discussion.