6.867 Project: Milestone 2

(Sapna Kumari, Archer Wang, Joe O'Connor)

As discussed in the last milestone, our goal is to evaluate the success of adversarial debiasing on the Compas Recidivism Risk Dataset. In order to do this, we will train two feed-forward neural network models, one of which will be trained in a non-adversarial way and one of which will be trained in an adversarial way. Both networks will use a sigmoid activation on the output layer so that we can interpret the result as the probability that a given convict will commit another crime. We plan to experiment with different network architectures (i.e., different numbers of layers, different numbers of neurons per layer, and different activation functions for the hidden layers). Our current plan is to use 90% of our data for training and 10% for testing; we will use cross-validation for hyperparameter selection.

Our input to the model will be a vector that is the result of concatenating several vectors corresponding to features from the Compas dataset. In particular, we will use word embeddings to represent the names, which we will obtain either explicitly from a suitable pretrained model or implicitly by projecting one-hot vectors in order to learn appropriate word embeddings as part of the overall network. Other features will include age (one-element vector), time spent in jail (one-element vector), crime arrested for (one-hot vector), and number of priors (one-element vector); we may add features as we further explore the dataset and begin implementing our model.

In terms of division of labor, we will all be involved in deciding on our final input representations and model architectures. Sapna will preprocess the data into CSV format, keeping only the data fields that we intend to use as input to our model. Once this is done, Joe will convert each row of the data into an appropriate input vector. Archer will divide the data into separate training and test sets.