CIS399: The Science of Data Ethics	Lecture 6: Feb 5, 2019
Instructor: Michael Kearns & Ani Nenkova	Scribes: Jacob Bass and Robert Zajac

1 Overview

This lecture was primarily focused on reviewing the methodology used in the *Wang and Kosinski study* about detecting sexual orientation based on a facial recognition model. The overarching question of the lecture was whether these methodologies were scientifically sound. We will go through each of the studies enacted, along with the processes used in these studies and their results.

2 The Data

The first important component of this study is how the researchers collected their data.

Data was collected from an unknown dating website and Facebook. It was composed of a feature space derived from a photo and a binary label (Gay or Straight). This data had to be cleaned significantly, both in "non-dubious" and "dubious" ways. In particular, it was not problematic to eliminate photos that were not of actual humans, contained more than one face, or contained obstructions. But more questionable changes included throwing out non-white subjects and balancing the data by age and by economic class.

An overall 60% of the data was thrown out, but there is no breakdown of that 60%. The conclusion is that findings based on this data would be questionable because the data does not significantly represent a snapshot of the population at large.

An aside by Professor Nenkova: When it comes to releasing data, institutions like Universities are significantly restricted in what they can publish. But for private companies, like a dating website, they can fully disclose their data if they wish. This might skew the types of data available to researchers, which is not completely applicable to this study but may affect similar studies.

2.1 Concerns About the Data

As has already been alluded to, we may have some concerns about the validity of this data. It was scraped from the Internet, and so there is an ethical concern that scraping this data, although legal, infringes upon individual privacy. The study also needed to be approved by a board for human study (in this case, Stanford's IRB). Perhaps the data collection was done by another party, which would make use of this data less concerning.

Questions we should keep in mind are: is using this data good science? Does the data accurately represent the world at large or just dating profiles?

3 Study 1

We now examine the procedures and results of Study 1.

The feature space these researchers worked with came from the VGG-Face System. The original purpose of the system system was to detect the same person across multiple photos. VGG itself is a neural network architecture trained on 3 million images, from which the researchers extracted 4,000 features for use. They then applied singular value decomposition (SVD) to simplify this 4,000 to 500 features.

As for the classifier, the researchers applied logistic regression, which gives the *probability* of each class. It was trained using cross-fold validation on 20 data folds. One caveat to note is that because of SVD, models trained in cross-fold validation may have used separate features than other folds.

The results reported by the researchers included an AUC of 0.81 on men with one photo and 0.91 on men with five photos. Similarly, they achieved AUC of 0.71 on women with one photo and 0.83 on women with five photos. Intuitively, the AUC tells us how sensitive this model is to the other class against all possible recalls. More photos of the subject clearly gives higher AUC because the model improves.

3.1 Study 1B

This study used the results of Study 1 to identify which features had most impact on the model's ability to predict sexual orientation. It took a sub-sample of 100 male images and 100 female images which were centered and standardized. Different parts of these images were iteratively blocked out and the model's subsequent predictions were measured against its original prediction. The study then measured changes in accuracy to conclude which parts of face were critical in predictions.

4 Study 2

In Study 2, the researchers trained a model on Facebook images that attempted to predict the probability of being a woman. The reported result was a 0.2 correlation between sexual orientation and model-given probability of being a woman. It was found that this result is significant statistically but not medically.

5 Study 4

(We elided discussion of Study 3 in class).

In study 4, 50,000 pairs of images were generated from the original data set, where we are guaranteed one is gay and one is straight. The study then measured human accuracy of predicting which person in a pair is gay and which is straight. The reported accuracy was 57% for men and 58% for women, though the methodology for measuring this accuracy was not given.

We do know that the participants were not trained in identifying gay or straight people. An interesting comparison to make would be to repeat the study with participants who are trained

this way. Still, the important conclusion made was that model accuracy was better than human accuracy in this classification task.

6 Study 5

In this study, the model was used to distinguish between straight men on dating sites and gay men on Facebook. Its reported accuracy in this task was 74%. When the model was used to distinguish gay men on dating sites and gay men on Facebook, its reported accuracy dropped to 53%.

7 Calculated Precision and Recall

The researchers created a subset of 930 straight images and 70 gay images and then reported model's precision on these values. Professor Nenkova then calculated the corresponding recall values, which are presented in Figure 1.

We can see from these precision and recall results that the model is not a reliable indicator of sexual orientation.

Table 1: Figure 1

	Reported Precision	Calculated Recall
Parameter Set 1	0.56	1.0
Parameter Set 3	0.75	0.3
Parameter Set 3	0.90	0.13

8 Conclusion

We should consider the question of how the model from this study was intended to be *deployed*. At first it seems only malicious use cases are possible, which might indicate that there was no purpose for the model to have been created at all. For example, hateful groups could use the model to profile their victims on Facebook. Still, one can imagine advertisers finding such a model helpful to tailor there advertisements based on a person's sexual orientation, which is a case that is decidedly grayer in terms of its ethical ramifications. In this case, Facebook users may still object to having their data gathered without explicit notification or consent.