CIS399: The Science of Data Ethics **Lecture 16: March 19, 2019**
Instructor: Michael Kearns & Ani Nenkova **Scribes: Kellie Dinh & Emily Lobel**

1 Methods of data collection

In order to have data to perform analysis on, we can use:

- historic data
- scrape existing data
- collect online data, often with elements of deceit
- collection annotations, often crowd-sourced
- perform own collection by asking people

Things we have thought about:

- What happens when you get a dataset and learn from it?
- How can you discover that there are problems with the dataset or model?

Overview of today's lecture:

- What if you would like to collect a dataset that doesn't have problems?
- What if we collect annotations? How do we ask people to annotate data for us?
- How do you collect online traces? (See assigned reading)
- What information can we get from online searches? (using medical searches for symptoms)
- Directly asking people - will they respond truthfully? Will their answers be protected?

Example to guide the lecture: what predictive task would we like to automate?

- college admissions? accept or deny applicant
- future potential of a company (as an investor)? should we fund them or not?
- level of emergency of calls to child services? should we allocate resources/make this call a higher priority?

2 Predictive task: Should we fund a start-up company?

2.1 Data collection

Historic data: We could look at historic data but there's risk of bias: women founders are overall funded less than male founders but women founded start-ups have overall higher return on investment.

What should we do first to collect this dataset?

Features Attain information about the company. Figure out what to ask them

Feature Selection Come up with criteria and what decisions were made based on company information

Issues:

- Decision could be from a small detail that isn't helpful in prediction i.e. founder mentioning a shared interest like golf
- Annotating the data (which features factored into the decision) is tedious
- Humans have different notions of what features should be within the decision-making criteria

We don't want to look at old data to inform the decision because of implicit bias.

2.2 Making predictions

How to make the decision:

Ask multiple people/committee

Computer inter-annotator agreement

- If you have a good inter-annotator agreement, you can say your guidelines for the automation are very good
- If people disagree, it means that the decision will be almost random.
    - With low inter-annotator agreement, you have to go back and change the instructions to be more explicit so to limit subjectivity.
- Example: standardized testing has essay component graded with strict time limit of 3 minutes for essay writing

2.2.1 Relevant Concepts and Equations

To calculate measurement bias, consider how much each feature is weighted in the final decision. Percent agreement (for two raters) is denoted as the number of times they agree on the class (a) over the number of sample points (s).

$P_0 = \frac{a}{s}$

This ties back into using precision and recall rather than accuracy to assess a model.

Kappa Stat:

Where P_e is expected agreement and P_0 is observed agreement,

$k = P_0 - \frac{P_e}{\wedge - P_e}$.

		Person1	
		Yes	No
Person2	Yes	a	b
	No	c	d

The propensity of a person to mark yes (where $N = a + b + c + d$) is:

$P_{\text{yes}} = (a + c/N)(a + b/N)$

$P_{\text{no}} = (c + d/N)(b + d/N)$

And,

$P_{\text{expected}} = P_{\text{yes}} + P_{\text{no}}$

2.3 Continuing to make predictions

In the past, data with low kappa values has not been worth automating.
Now, there is a larger push for automating decisions overall, despite the value of the kappa stat.

Because data has been collected via crowd sourcing, one should be considerate to adjust the kappa statistic if needed. For example, one should adjust the kappa statistic if one no longer has everyone responding to the same questions. Consider, how much are people agreeing with the annotations of others? When one group disagrees with another group's annotations, it is unclear who is correct.

Consider the reliability of your data. Are we able to predict the correct outcome given historic data? Is there bias in previous decision making? (Example of auditions for an orchestra being done behind a curtain leading to a more diverse group of instrumentalists.)

Can we predict the confidence/certainty of the class assignment? Which decisions were unanimous? Which were split?

3 Future class plans

A reading will cover information about our health searches on Google (and other websites) being leaked to third parties and storing information about searches so to help predict what to advertise. How often are people not aware that theyre being tracked?

The author looked at around 1000 websites and their privacy policies and analyzed how long it takes to read the policy. The author discovered that it takes 6 minutes on average to read a website such as Facebook's privacy policy, however it takes 15 minutes on average to read third party collectors policies.

The author examines if it is clear that the website will or will not track your history. As for

Facebook, it's not clear and it is easy to consent to send over your data without fully reading the details.

Keep in mind that, rather than being tricky, websites can just ask people for their data and promise them to be private!