Prevalence Estimation in Social Media Using Black Box Classifiers

Siqi Wu, Paul Resnick

University of Michigan
{siqiwu, presnick}@umich.edu

Abstract

Many problems in computational social science require estimating the proportion of items with a particular property. This counting task is called prevalence estimation or quantification. Frequently, researchers have a pre-trained classifier available to them. However, it is usually not safe to simply apply the classifier to all items and count the predictions of each class, because the test dataset may differ in important ways from the dataset on which the classifier was trained, a phenomenon called distribution shift. In addition, a second type of distribution shift may occur when one wishes to compare the prevalence between multiple datasets, such as tracking changes over time. To cope with that, some assumptions need to be made about the nature of possible distribution shifts across datasets, a process that we call extrapolation.

This tutorial will introduce an end-to-end framework for prevalence estimation using black box (pre-trained) classifiers, with a focus on social media datasets. The framework consists of a calibration phase and an extrapolation phase, aiming to address the two types of distribution shifts described above. We will provide hands-on exercises that walk the participants through solving a real world problem of quantifying positive tweets in datasets from two separate time periods. All datasets, pre-trained models, and example codes will be provided in a Jupyter notebook. After attending this tutorial, participants will be able to understand the basics of the prevalence estimation problem in social media, and construct a data analysis pipeline to conduct prevalence estimation for their projects.

Organizers

The organizers are Siqi Wu (main contact) and Paul Resnick. Their short biographies are as follows.

Siqi Wu[1] is a postdoctoral research fellow at the University of Michigan School of Information. Prior to that, he was a research fellow at the Australian National University, where he also earned his Ph.D. in Computer Science from. His research interests include computational social science and social computing. He has published papers at ICWSM, CSCW, CIKM, and WWW. He is the recipient of a best paper honorable mention award at CSCW, a spotlight paper award at ICWSM, and a Google PhD fellowship.

Paul Resnick[2] is the Michael D. Cohen Collegiate Professor of Information and Associate Dean for Research and Innovation at the University of Michigan School of Information. He was a pioneer in the field of recommender systems (sometimes called collaborative filtering). The GroupLens system he helped develop was awarded the 2010 ACM Software Systems Award. His articles have appeared in Scientific American, Wired, Communications of the ACM, The American Economic Review, Management Science, and many other venues. His 2012 MIT Press book (co-authored with Robert Kraut), was titled "Building Successful Online Communities: Evidence-based Social Design". He is an ACM Fellow, past chair of the RecSys Conference and past program chair and co-editor-in-chief for ICWSM.

Tutorial Type, Duration, Schedule, and Activities

Instead of presenting a lecture-style course first and hands-on exercises at the end, this tutorial will be split into a few segments, each consisting of lead-in slides that introduce new concepts and techniques, a hands-on programming exercise in provided Jupyter notebook, and detailed reflection slides that summarize what the participants have implemented. The scheduled duration is four hours. The outline is as follows.

1. **Introduction** (60 mins). This opening segment contains the participant introduction that lists the definition, motivation, and applications of the prevalence estimation problem in social media. We also review the prerequisite knowledge in machine learning and probability theory. It ends with the first hands-on component, where the participants will get familiar with the end-to-end prevalence estimation framework, simulated data generation process, and see what would go wrong with the naive classify and count method in the face of distribution shift.
2. **Distribution shift** (30 mins). This segment details the challenge of distribution shift and describes what kinds of distribution shifts are plausible. Using simulated datasets, we will enumerate all possible distribution shift scenarios and see how off the estimations would be if making the wrong stability assumptions.

[1] Personal website: https://avalanchesiqi.github.io/

[2] Personal website: http://presnick.people.si.umich.edu/

3. **Coffee break** (30 mins).
4. **An end-to-end example** (15 mins). We pick a real world task of estimating the proportion of positive tweets in datasets from one time period, or from multiple separate periods. Using our framework, we will walk the participants through a complete example. This segment introduces the tweet datasets, and a variety of classifiers with different discriminative power.
   (a) **Calibration in binary case** (25 mins). This segment implements the calibration phase when the classifiers output binary outcome, covering the steps of purposefully selecting data sample, collecting human labels for it, and extending those labels to an inferred joint distribution between classifier outputs and human labels. The calibration phase copes with the distribution shift from unknown training dataset to the base test dataset.
   (b) **Extrapolation in binary case** (25 mins). This segment implements the extrapolation phase when the classifiers output binary outcome, involving a choice of stability assumptions, each of which yields a different extrapolation to an inferred joint distribution for a new dataset, from which a prevalence estimate can be read. The extrapolation phase copes with the distribution shift from the base dataset to the target dataset.
   (c) **Calibration and extrapolation in continuous case** (30 mins). This segment implements both the calibration and extrapolation phase when the classifiers output continuous scores between 0 and 1.
   (d) **The choice of stability assumptions** (15 mins). This segment details the challenge of doing prevalence estimation for multiple time periods, using a real dataset where the "true" answer is not known.
5. **Conclusion and takeaways** (10 mins). This segment concludes the tutorial, and gives a step-by-step guide for participants to build the end-to-end prevalence estimation pipeline for their own projects.

Audience, Prerequisites, and Objectives

The target audience are students, researchers, and practitioners who want to estimate the prevalence of phenomena in social media, using black box machine learning models that have been pre-trained by others.

Prerequisites. The audience are expected to have some basic knowledge in the following fields:

- Machine learning. Understand the basics of logistic regression or some other text classification methods, in particular that they take as input a set of labeled items and produce as output a function from items to scores.
- Probability theory. Be familiar with the representation of a probability distribution as a graph of the probability density function. Understand the notion of a joint distribution between two random variables; in our case, the joint distribution between classifier outputs and ground truth labels.
- Python. Sufficient python knowledge to understand and make changes to a Jupyter notebook that invokes a pre-trained classifier on text inputs.

Learning objectives. After attending this tutorial, participants will be able to:

- describe the prevalence estimation (i.e., counting or quantification) problem.
- understand the fundamental challenge of using any model to make inferences on datasets that the model was not trained on, and extrapolating the knowledge learned from one dataset to a future dataset.
- describe scenarios in which alternative distribution shift assumptions (i.e., stability assumptions) hold.
- build a data analysis pipeline to conduct prevalence estimation for their projects:
  - Calibration phase. Select a data sample, collect human labels for it, and use those labels to calibrate a pre-trained black box classifier.
  - Extrapolation phase. Calculate prevalence based on extrapolating a variety of alternative distribution shift assumptions to the target dataset.

Materials

We will set up a web page to host the learning materials for this tutorial, including

- tutorial abstract and schedule;
- presentation slides with links to referenced papers;
- a Jupyter notebook with example codes. Required python packages, external datasets, and pre-trained models will be downloaded directly by commands in the notebook.