Uncertainty-aware generative models for inferring document class prevalence

Katherine A. Keith and Brendan O'Connor
College of Information and Computer Sciences, University of Massachusetts Amherst
http://slanglab.cs.umass.edu/doc_prevalence/

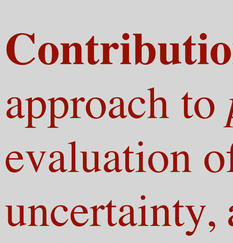
Summary

- **Contributions:** (1) generative probabilistic modeling approach to *prevalence estimation*, (2) construction and evaluation of prevalence credible intervals to measure uncertainty, and (3) a large-scale and replicable empirical evaluation.
- **Evaluation:** Empirically, our LR-implicit method (1) provides better confidence interval coverage and (2) is more robust to shifts in class distributions between training and testing than other models.

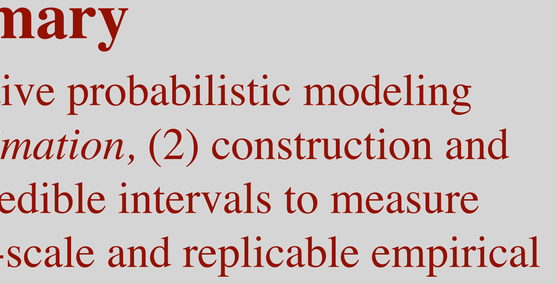
Prevalence estimation overview

Predict class distribution over unlabeled examples in a group

Training



Inference / Testing



Prevalence estimate → $\theta^* = \frac{1}{n}\sum_{i=1}^{n} 1\{y_i = 1\} = 0.66$

Differences from classification:

- Class priors assumed to be different in training vs. testing
- Possibly multiple test groups

Example applications:

- How do topics discussed on Twitter change per day?
- What is the positive sentiment per business on Yelp? *(our empirical experiments)*

Discriminative baselines

- **Obvious method:** Train a discriminative classifier and aggregate individual classifications at test-time
- If a classifier is *perfectly accurate*, it will give perfect prevalence estimates. However, classifiers often exhibit errors due to (1) shifts in the class distribution between training and testing, and (2) difficult tasks (e.g. predicting sentiment or sarcasm)

Prob. of positive class from discriminative classifier

Classify and count (CC) (Forman, 2005)

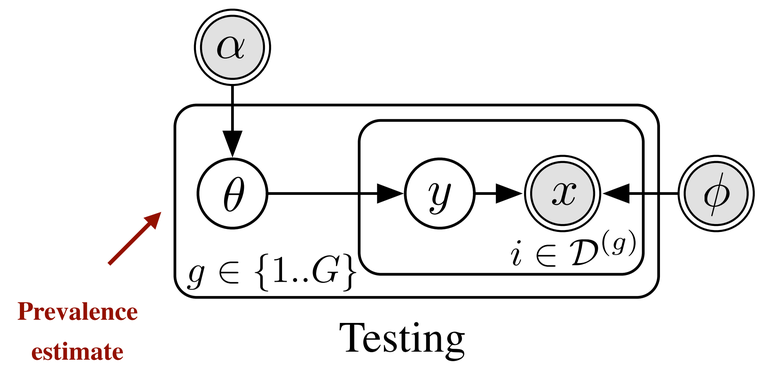
$$\hat{\theta}^{CC} = \frac{1}{n}\sum_i 1\{p_i > 0.5\}$$

Poisson-Binomial probabilistic classify and count (PB-PCC)

$$y_i \sim p_i$$
$$\hat{\theta}^{PB\text{-}PCC} = \frac{\sum_i y_i}{n}$$

Additional methods (in the paper): ACC, Readme (Hopkins and King, 2010)

Generative model for prevalence estimation



Class-conditional language models ϕ_y

- Multinomial Naive Bayes (MNB)
- SAGE (Eisenstein, 2011), i.e. log-linear

Implicit likelihoods from discriminative classifiers (LR-implicit)

$$p_{\text{discriminative}}(y|x) = \frac{p_{\text{implicit}}(x|y)p_{\text{train}}(y)}{p(x)}$$
$$p_{\text{implicit}}(x|y) \propto \frac{p_{\text{discriminative}}(y|x)}{\hat{\theta}_{\text{train}}}$$

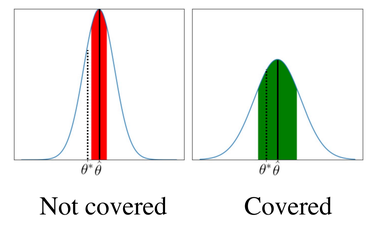
Inference: Obtain a posterior distribution over θ via marginal log likelihood (MLL)

$$\text{MLL}_g(\theta) \equiv \log p(\mathcal{D}^{(g)}|\theta) = \sum_{i \in \mathcal{D}^{(g)}} \log\Big(\theta p(x_i|y_i = 1) + (1-\theta)p(x_i|y_i = 0)\Big)$$

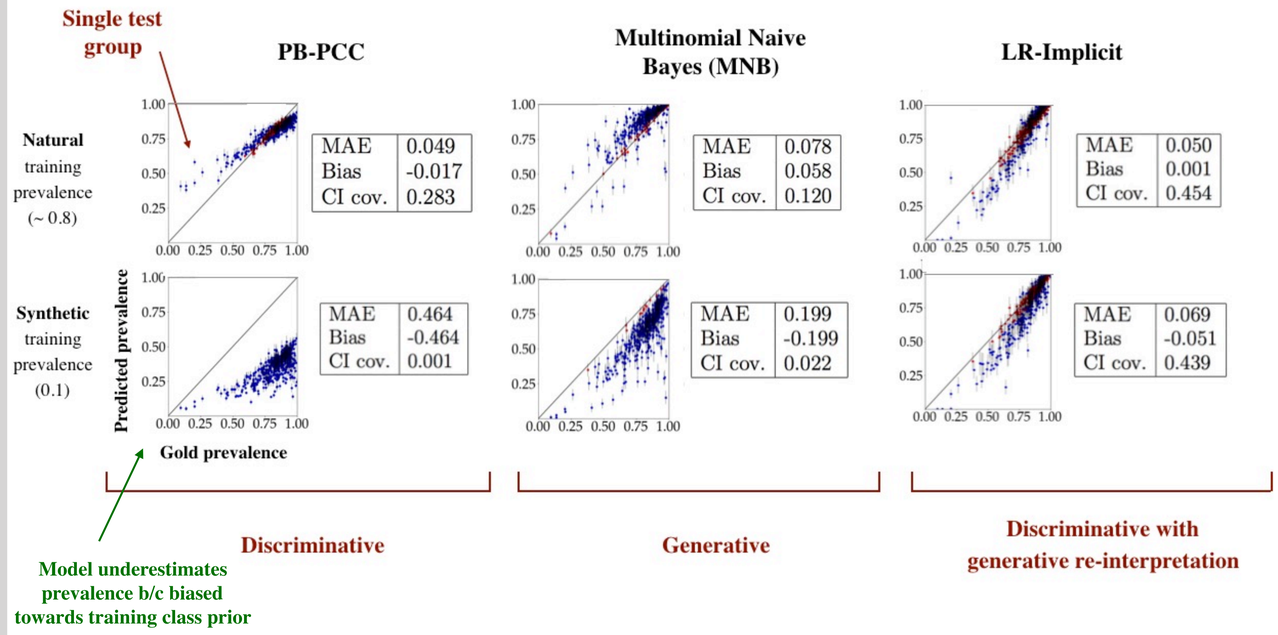
Modeling goals

(1) Better test-time **point estimates** of prevalence across multiple groups

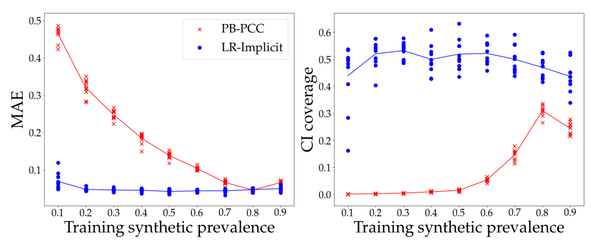
(2) Represent uncertainty via **credible interval coverage** (CI cov.)



Results highlights



Varying training prevalence



Varying training size