Conditioning on "and nothing else": Simple Models of Missing Data between Naive Bayes and Logistic Regression

David Poole, Ali Mohammad Mehr, Wan Shing Martin Wang
Department of Computer Science
University of British Columbia

ICML Artemiss Workshop, July 2020

Motivation

- Example: people volunteer information about themselves:
  - Half the people have siblings.
  - Siblings mentioned in 90% of the cases with siblings.
  - Siblings never mentioned when there were no siblings

  $P(has\_sibling) = 0.5$ but
  $P(has\_sibling \mid sibling\_was\_not\_mentioned) = 5/55 \approx 0.09.$
- Simple models are important (e.g., relational representations equivalent to logistic regression in some cases).
- We want to ignore observations not mentioned, but take them into account.
  Non-observations should have zero computation cost.

Idea (LR±)

For simple models of missing data:

- Model phenomenon of interest assuming all data is missing.
- For each possible observation, model how that observation would change the prediction.
- Logisitic regression for Boolean variables, two parameters per variable: w_i^+ when X_i is true and w_i^- when X_i is false

$$P(y \mid X_1 \ldots X_n \text{ and nothing else})$$
$$= sigmoid(w_0 + \sum_{i=1}^{n} w_i^+ X_i^+ + w_i^- X_i^-)$$

$X_i^+ = 1$ when X_i is observed to be true
$X_i^- = 1$ when X_i is observed to be false

- $sigmoid(w_0) = P(y \mid nothing\_was\_mentioned)$

Results

- for simple models LR$\pm$ can fit data better than explicitly modelling the missing data
- motivated by relational models where most variables are unobserved
- easy to extend to other discrete variables (indicator variable for each value)