Data Reduction for Efficient Probit Regression

Christian Peters

September 21, 2021

Objectives

- Develop a data reduction algorithm, $n \to \text{poly}(d \log(n))$, for μ-complex datasets
  - Obtain $(1 \pm \epsilon)$-coreset
- Extend the algorithm to row-by-row streams (one pass)
- Evaluate loss function approximation quality on 3 real world datasets
- Evaluate applicability to Bayesian data analysis on 3 real world datasets using the efficient Gibbs Sampler
- Outperform uniform sampling and stochastic gradient descent

Theory section

- Introduce probit model + Gibbs sampler
  - Start with latent variable model
  - Derive optimization procedure for the ML method
- Introduce μ-complexity for the probit model
- Introduce coresets
- Use the sensitivity framework to construct a coreset
  - Bound the sensitivities + Bound the VC-dimension
  - Introduce leverage score sampling
- Show how leverage scores can be approximated online
  - Show that we still have a coreset

Experiments section

- Implement ML optimizer
  - Deal with numerical issues
- Evaluate approximation quality of the loss function

$$\text{Quality measure:} \quad \frac{f(\tilde{\beta}_{opt})}{f(\beta_{opt})}$$

- Bayes Experiments
  - Implement Gibbs Sampler (with sample weights!)
  - Select measures to compare distributions (MMD, distance between means)
  - Show how much time can be saved