Evaluation Framework for Probabilistic Programming Languages

- Widespread adoption of statistical modeling in many domains
- Increasing number of PPLs, inference methods, models applied
- Need to quantify model performance across PPLs, inference methods

Conceptual Framework of PPL Bench:

- Model Instantiation and Data Generation

\[ P_0(X, Z) = P_0(Z) P_0(X | Z) \]
\[ Z_1 \sim P_0(Z) \]
\[ X_{\text{train}} \overset{iid}{\sim} P_0(X | Z = Z_1) \]
\[ X_{\text{test}} \overset{iid}{\sim} P_0(X | Z = Z_1) \]

- PPL Implementation and Posterior Sampling

\[ Z_1^{n}, \ldots, n \sim P_0(Z | X = X_{\text{train}}) \]

- Evaluation of Posterior Samples

\[ \text{Predictive Log Likelihood}(n) = \log \left( \frac{1}{n} \sum_{i=1}^{n} P(X_{\text{test}} | Z_i) \right) \]

PPL Bench supports adding new models
- Specify data generation
- Predictive Log Likelihood computation
- Other custom evaluation metrics

Bayesian Logistic Regression:
- Simple model; baseline
- Log-concave posterior, easy convergence

Noisy-Or Topic Model:
- Inferring topics from words in document
- Bayesian Network structure with topics and words as nodes
- Supports hierarchical topics

Crowdsourced Annotation:
- Inferring true label of an object given multiple labeler’s label assignments
- Maintain confusion matrix of each labeler
- Includes inferring the unknown prevalence of labels

PPL Bench Summary:
- Modular Customizable framework for evaluating PPLs, statistical models, and inference methods
- Open source, with lot of potential for extension

Using PPL Bench:
- Comparing model performance across PPLs
- Comparing effectiveness of inference algorithms across models
- Evaluating new inference algorithms

Contributing to PPL Bench:
- Add new models
- Add model-specific eval metrics
- Add new PPLs
- Add PPL implementations of existing Models

Next Steps: PPL Bench Open Source

Website: pplbench.org
GitHub Repository: github.com/facebookresearch/pplbench