9-2019

Development of Probabilistic Cardinal Models

Oliver Serang
University of Montana, Missoula

University Grant Program 2018-2019 Final Report

Name: Oliver Serang
Department: Computer Science
Project: Development of Probabilistic Cardinal Models

Objective
Cardinal models solve problems of the form Y=X_1+X_2+...+X_n, where we have discrete distributions on each random variable. The objective was to improve the usability and performance of Evergreen, an engine for solving cardinal models and to create cardinal models using them.

Summary of Results
Funding was used to sponsor undergraduate students in computer science.

A modeling language for probabilistic cardinal models.
The most significant result is the development of a basic modeling language for fast prototyping and development of cardinal models. This was developed as an interpreter by using the UNIX tools yacc and flex. The result is a modeling language: instead of directly using the C++ library, the modeling language allows a directly interpreted version of the model.

For example, the modeling language can encode a problem using the following script:

```
PMF (Y) (0) [0.1, 0.2, 0.05, 0.15, 0.5]
PMF (X_1) (0) (3) UNIFORM
PMF (X_2) (-1) (3) UNIFORM
PMF (X_3) (0) [0.9, 0.1]
PMF (X_4) (1) [0.2, 0.8]
Y=X_1+X_2+X_3+X_4
Pr(X_1)
```

The final line outputs the posterior distribution on the random variable X_1. The output is

```
X_1 PMF:{[0] to [3]} t:[0.325867, 0.264451, 0.233743, 0.175939]
```

Importantly, the underlying computations for solving the posterior are computed via the high-performance C++ library.

Posteriors can be computed via loopy belief propagation (default), which is a fast numeric approximation, or via brute force by prepending the line

```
@engine=brute_force()
```

Demonstration of Evergreen for solving subset-sum and knapsack problems
Evergreen can be used to solve NP-complete problems, such as subset-sum and knapsack problems.

The following script solves a subset-sum problem wherein four people order food at a restaurant. The first person spends either $6 or $9. The second person orders either $0, $5, or $7. The third person orders either $5 or $9. The fourth person orders either $4 or $8. The output reveals that the total bill could be $15 but cannot be $16 or $17, and that it could be $18, $19, *etc*.

```
PMF (PERSON_1) (6) [1, 0, 0, 1]
PMF (PERSON_2) (0) [1, 0, 0, 0, 0, 1, 0, 1]
PMF (PERSON_3) (5) [1, 0, 0, 0, 1]
```

PMF (PERSON_4) (4) [1, 0, 0, 0, 1]
TOTAL=PERSON_1+PERSON_2+PERSON_3+PERSON_4
Pr(TOTAL)

This subset-sum problem can be used to solve the corresponding knapsack problem by prepending the script with @p=inf and modifying the script so that each value indicates a person's preferences. For example, the line

PMF (PERSON_1) (6) [1, 0, 0, 1]

could be replaced with

PMF (PERSON_1) (6) [10, 0, 0, 1]

to indicate that the first person is 10 times happier about spending $6 than spending $9.

@p=inf
PMF (PERSON_1) (6) [10, 0, 0, 1]
PMF (PERSON_2) (0) [1, 0, 0, 0, 0, 5, 0, 2]
PMF (PERSON_3) (5) [1, 0, 0, 0, 7]
PMF (PERSON_4) (4) [1, 0, 0, 0, 4]
PMF (TOTAL) (10) [1, 0, 2, 0, 3, 1, 0, 4, 2, 0, 1]
TOTAL=PERSON_1+PERSON_2+PERSON_3+PERSON_4
Pr(TOTAL; PERSON_1)

The output reveals that the optimal total bill is $20 and that the optimal order for PERSON_1 is $6.

Application of Evergreen to proteomic / RNASeq data

A standard protein inference model (*i.e.*, noisy-or from Serang *et al.* 2010) was built using the modeling language. This protein inference model was able to compute marginals for proteins (*i.e.*, infer which proteins are in the data set) using loopy belief propagation. For large data sets, this was frequently faster than the best available methods.

Availability of results

The engine, including both C++ code underpinnings and the language, are available from https://bitbucket.org/orserang/evergreenforest .

Development of external funds

This small grant was used to generate preliminary results for an NSF CAREER award application. This application was submitted on July 18, 2018 and was awarded May 30, 2019. The NSF CAREER award is valued at roughly $700k (direct) and roughly $1M (total) [award number 1845465].