PPAML Hackathon Problem Statement

Tom Dietterich, Chad Scherrer, Eric Woldridge, Max Orhai Version 6, January 22, 2017

Introduction to the Data Set

Gapminder is a data set created by the Swedish Gapminder Foundation, whose goal is to help people become more informed about the countries of the world and especially about progress in development. It provides a curated (and in some cases, smoothed) database with 519 variables by country and by year. Different variables have different coverage across time and countries and there is a huge percentage of missing values.

TA1 has designed a custom version of the Gapminder database for this Hackathon. This has involved a major effort at variable selection, row selection (e.g., to eliminate tiny countries and city states), and data cleaning. It has also involved making modifications and seeding "ground truth" into some parts of the database. Therefore, the teams are directed to not compare the Hackathon database against the Gapminder original data, because this would allow you in some cases to detect the changes we have made.

The database is made available as a CSV file for easy ingest. Each variable has a short name, e.g., "sugar", and a longer description "Sugar per person (g per day)". The mapping between these is given in the file "variable-meaning.csv". Where possible, we have transformed the variables to be per capita ("per1") or per 1000 people ("per1k"). In several cases, we provide the log-transformed version of the variable as well. Missing values are encoded as "NA".

The data folders are laid out as follows:

```
_data_
        raw-data
                country-year.csv
        meta-data
                variable-meaning.csv
        task1
                imputation-variables.csv
        task4
                task4-model.pdf
        task5
                ta1-invest-gdp-per1k.csv
                ta1-invest-mortality-kid.csv
                ta1-invest-mortality-maternal.csv
                ta1-invest-life-expectancy.csv
```

Input, Output, and Contextual Variables

The following tables list the variables that we will use in the hackathon. We have divided the variables into three general sets: (a) variables that could be viewed as interventions (e.g., foreign aid), and (b) output variables that can be viewed as measures of success of those interventions (e.g., GDP per person), and (c) all other variables.

Intervention Variables	Description
aid_received	Total foreign aid received ($)
invest_foreign_per1k	Foreign investment per 1000 people
invest_domestic_per1k	Domestic investment per 1000 people
spending_health_per1	health spending per person (US $)

Outcome Variables	Description
gdp_per1k	GDP per 1000 people
mortality_kid	Child deaths before age 5 per 1,000 live births
life_expectancy	Life Expectancy at Birth
mortality_maternal	Maternal deaths per 100,000 live births

Other Variables	Description
population	population
population_density	population per square km
surface	land area in square km
gini	gini index
continent	continent where country is located
completion	Primary School Completion Rate
agriculture	area of agricultural land (sq km)
energy_per1*	Energy Consumption Per Capita
disaster*	People Affected by Natural Disasters
physicians	Doctors Per 1000 People
hiv	People living with HIV
malaria	Malaria Reported Cases per 100000
professional_birth	Births attended by Trained Staff
sugar	Sugar consumption per person
food	Food supply per person (kcals)
infection	New TB Cases per 100000
labor	% Labor Force Participation (Total, 15-64)
broadband	Broadband subscribers per 100 people
cell_phone	Cell phones per 100 people
internet	Internet users per 100 people
computers	PCs per 100 people

it_tel	Total phone subscribers per 100 people
sanitation	Proportion of population using improved sanitation
water_source	Proportion of population using improved drinking water
child_immune*	Child immunization percentage

Three variables in the above table have an asterisk. In the case of "child_immune" and "disaster", these variables do not exist. However, we provide several other variables that could be processed to create values for these variables.

Variables for inferring People Affected by Natural Disasters:	Filename
drought	Drought affected
earthquake	Earthquake affected
epidemic	Epidemic affected
extreme_temp	Extreme temperature affected
flood	Flood affected
storm	Storm affected
tsunami	Tsunami affected
air_accident	Air accident affected

Variables for Inferring Amount of Childhood Immunization	Filename
immune_diphtheria	DTP3 immunized (% of one-year-olds)
immune_hepatitis	HepB3 immunized (% of one-year-olds)
immune_hib	Hib3 immunized (% of one-year-olds)
mcv	MCV immunized (% of one-year-olds)
immune_tetanus	PAB immunized (% of newborns)

The third variable, "energy_per1" (Energy Consumption per Capita) exists, but it is frequently missing. We suggest that the teams impute its value using the variables listed in the following table:

Variables for imputing Energy Consumption Per Capita:	Description
coal_per1	Coal Energy Use Per Capita
electricity_per1	Electrial Energy Use Per Capita
oil_per1	Oil Consumption Per Capita
natural_gas_per1	Natural Gas Production Per Capita
co2_per1	CO2 Emissions Per Capita

Analysis Tasks

During the hackathon, you will work on each of the following tasks. We have structured the Hackathon so that each of these tasks can be performed independently and concurrently. However, it would obviously be more elegant and a greater demonstration of the power of probabilistic programming if Tasks 1-4 could be attacked jointly. So we encourage the teams to pursue this if there is time.

We provide an overview of the task here. Details on deliverables are given on subsequent pages.

Task 1: Imputation

The goal of this task is to create and fit a model of the joint distribution over the variables in the database and then draw samples from it to impute the missing values. Specifically, you should develop an imputation model for the variables listed in the file "imputation-variables.csv" in the "task1" folder.

The model can, of course, be used to improve the solutions to the other tasks.

Task 2: Modeling the Effect of Intervention Variables on Outcome Variables

For each intervention variable and each outcome variable, is there any evidence that the intervention affects the outcome, possibly after some delay? If possible, formulate and test causal hypotheses. For example, does a change in the intervention variable lead to a change in the outcome variable? Consider segmenting the countries to uncover groups of countries where the intervention variables have some effect. It is reasonable to expect that in highly-developed countries, the interventions have no effect because the outcomes are already so good. Conversely, very poor countries may lack institutions that allow the intervention variables to work.

Consider additional variables (e.g., natural disasters) that might interfere with the effect of the interventions. Consider whether some of the other variables might be useful for modeling intermediate variables. Consider introducing latent variables that generate both the outcome variables and some of the other variables (e.g., investment might lead to improvements in medical infrastructure that then lead to increased number of doctors, doctor-assisted births, immunizations, etc. that in turn lead to improved child mortality, maternal mortality, and life expectancy).

Consider creating submodels for Energy Consumption Per Capita, Natural Disasters, and Childhood Immunization and incorporating them into the intervention model.

We have provided variables for the first differences of the (log-transformed) intervention variables and the outcome variables. These have names like "delta_aid_received_log", which is equal to $\log aid\_received(t) - \log aid\_received(t-1)$ for year t.

To gain some understanding of the effect of interventions, each team will do a "zero-intervention simulation". Specifically, you will compute the values of the outcome variables that your model predicts would have been observed if all intervention variables had been set to zero for the entire time period (see below for deliverable format).

Notes on the Outcome Variables: Different Outcome variables may respond on different time scales. Child mortality, maternal mortality, and GDP are likely to respond relatively quickly to interventions, whereas life expectancy may respond more slowly.

Task 3: Modeling Missingness Processes

Standard imputation methods assume that the missing values are missing at random. However, there are likely to be values missing not at random (MNAR). The goal of this task is to create a model of the processes that determine whether a variable will be missing. For example, does the value of the variable influence whether it is missing? Are there other variables or combinations of variables that can predict whether a variable will be missing? For each variable, determine whether it can safely be treated as missing at random or whether you need to include a conditional observation model. If a conditional observation model is needed, then fit that model.

Task 4: Model Criticism

In the folder "task4-model.pdf", there is a model for mortality_maternal. This is *not* a causal model but instead just models the behavior of the mortality_maternal variable as a function over other variables. Criticize this model (e.g., by analysis of residuals, posterior predictive checks, sensitivity analysis) to identify its shortcomings. Then design a new model to address these shortcomings. Examples of shortcomings might include missing or redundant predictor variables, inappropriate data transformations, overly-simplistic models (e. g., that ignore change points or mixtures of processes), and so forth.

Task 5: Decision Analysis

Suppose you had a budget of $10 billion ($2 billion per year for five years). How would you invest it most effectively? Specifically, for each country c and intervention variable m you should choose a constant amount of money $x_{c,m}$ to spend in each of the years 2006, 2007, 2008, 2009, and 2010 (for a total expenditure of $\$5 \times \sum_m x_{c,m}$ on country c). Let X consist of all of the values of $x_{c,m}$ for all c and m.

We will explore this question separately for each of the outcome variables, and we will measure the impact of the investments over the years 2006, 2007, 2008, 2009, 2010. Note that for the year 2010 investment we will only measure the immediate benefit *in that same year* (2010).

For mortality_kid, the goal is to maximize the number of child deaths prevented over the five-year period. Hence, the objective should be to minimize the sum over countries (and over the five years) of mortality_kid × population × live_births ÷ 1000. However, we don't know how many live births there are, so we will just assume that it is a fixed fraction of the population. Hence, the objective is

$$J_{mortality.kid}(X) = \sum_{y=2006}^{2010} \sum_{c} mortality\_kid(c,y) \times population(c,y)$$

For mortality_maternal, the goal is analogous and so is the problem that we don't know the number of live births. Hence, we will adopt the following objective:

$$J_{mortality.maternal}(X) = \sum_{y=2006}^{2010} \sum_{c} mortality\_maternal(c, y) \times population(c, y)$$

For GDP_per1k, the goal is to maximize the number of people living in countries whose GDP_per1k exceeds $10 Hence, the focus will be on bringing countries where GDP_per1k is below $10 up to this level. Hence, the objective will be

$$J_{GDP.per1k}(X) = \sum_{y=2006}^{2010} \sum_{c} \mathbb{I}[GDP\_per1k(c, y) \geq 10] \times population(c, y),$$

where $\mathbb{I}[b]$ is the indicator function that is 1 if b is true and 0 otherwise.

Finally, for life_expectancy, the goal is to maximize the total number of person_years of increased life. Let life_expectancy_0 be the life expectancy with no additional investment in the intervention variables, and life_expectancy be the (predicted) life expectancy with the additional investment. Then the objective is

$$J_{life.expectancy}(X) = \sum_{y=2006}^{2010} \sum_{c} [life\_expectancy(c, y) - life\_expectancy.0(c, y)] \times population(c, y)$$

To solve these optimization problems, one approach could be to use a general purpose package for constrained optimization. An alternative is to try to formulate the objective as a likelihood function and employ maximum likelihood inference to find the optimum..

Deliverables

Here is a description of the information to be delivered in response to each task. All files and documents should be pushed to the git repository area for your group.

In addition, we would like to capture the various models that you explore while doing these tasks. We would like to document the claim that probabilistic programming languages support rapid exploration of a wide range of possible models. To support this, we encourage you to push each model that you develop to a git branch. If a model undergoes a sequence of refinements, these could be committed as model changes, and we can use the git change history to reconstruct the various model versions.

Task 1

Please submit a document that describes the approach you took to imputing the missing values. Along with this, submit the probabilistic program(s) and scripts that you used to perform the imputation. If you fitted a joint model of the data, please submit the fitted model in the agreed-upon digital form (typically as a probabilistic program with either point MAP estimates or posterior samples for the relevant parameters).

To submit samples from the joint posterior, produce 100 files having the same format as "country-year.csv" but with all NA's for the variables listed in "imputation-variables.csv" replaced by an imputed value. Name the files "country-year-imputation-nn.csv", where nn ranges from 00 to 99. Place these files in the "task1" folder of your team's area on gitlab. The imputed values in each file should be a coherent draw from the joint posterior.

Task 2:

Please submit a document that describes the approach you took to analyzing the effects of interventions on outcomes and your findings. For each outcome variable, describe your findings concerning which intervention variables had an effect and quantify that effect (e.g., by providing estimates of model coefficients, estimates of mutual information, and so on). If a variable had an effect, describe the conditions under which the effect is obtained. Were there other variables (intermediate or other outcome variables) that appeared to have a strong causal effect on the outcome variable? Does the intervention variable act indirectly through such other variables?

Submit the probabilistic program(s) and scripts that you developed. Please submit the fitted model(s) in the agreed-upon digital form.

Finally, for the zero-intervention simulation, please submit 20 samples from the posterior distribution over the time series for each outcome variable. Name the files "zero-intervention-nn.csv", where nn = 00 through 19. Each csv file should including the following information:

country_id	year	gdp_per1k	mortality_kid	life_expectancy	mortality_maternal

Place these in the "task2" folder in your team's area of the repository.

Task 3:

Please submit a document that describes the approach you took to detecting and modeling data missing not at random and your findings. If you fit explicit observation models, please describe them. Did you find a way to use observation models to improve the imputation of missing values? Submit the probabilistic program(s) and scripts that you developed. Please submit any fitted model(s) in the agreed-upon digital form. If your observation models were included in your solution to Task 1, then you do not need to submit them a second time for Task 3. In this case, please commit a file named "task-3-results-included-in-task-1.txt" to the "task3" folder on the repository, so that we know where to find your solution.

Task 4:

Please submit a document that describes the failures that you identified in the provided model for mortality_maternal. The document should also describe the refined model that you developed to address those failures. Submit the probabilistic program(s) and scripts that you developed. Please submit the refined model in the agreed-upon digital form. Please submit your fitted (predicted) values for mortality_maternal as a csv file in the following format:

country_id	year	mortality_maternal

The file should be named "task-4-mortality-maternal.csv" and be placed in your team's "task4" folder on the repository. If the refined model was included in your solution to Task 2, you do not need to submit the refined model and fitted values for mortality_maternal a second time for Task 4. Instead, please push an empty file named "task-4-results-included-in-task-2.txt" so that we know where to look for your Task 4 results.

Task 5:

Please submit four files, one per outcome variable, describing your proposed allocation of the $10B budget. Each file should be a CSV of the following form, with one row per country c:

country_id	aid_received	invest_foreign_per1k	invest_domestic_per1k	spending_health_per1k
c	$x_{c,aid}$	$x_{c,invest.f}$	$x_{c,invest.d}$	$x_{c,spend.h}$

The sum over all countries and x values should be $2B, because the amounts will be repeated in all five years.

The files should be named "invest-gdp-per1k.csv", "invest-mortality-kid.csv", "invest-mortality-maternal.csv", and "invest-life-expectancy.csv" and be placed in the "task5" folder in your team's area on gitlab.

In addition, please submit the expected value that your models predict will be attained for each objective function. Submit this as a CSV file named "objective-function-expected-values.csv" in the following format:

gdp_per1k	mortality_kid	life_expectancy	mortality_maternal
$J_{GDP.per1k}(X)$	$J_{mortality.kid}(X)$	$J_{life.expectancy}(X)$	$J_{mortality.maternal}(X)$

We will provide a baseline investment strategy in the same format as the "invest" files. These will be placed in the "task5" folder within the "_data_" tree. Please apply your fitted models to predict the value of the objective function for the TA1 baseline investment strategy. This will require you to develop code that can read in an investment strategy and evaluate each objective function using your models. Submit the resulting objective function values as "ta1-objective-function-expected-values.csv".

We hope that the results will show the benefit of improved modeling over our baseline modeling approach.

This may be too ambitious for the 2-day hackathon, so we will decide whether to attempt Task 5 after the first day based on progress on Tasks 1-4.