DEPARTMENT OF COMPUTER SCIENCE - CSC591 - FALL 2018 - Edu Tech

TAM N. NGUYEN - GITHUB REPO

BAYESIAN KNOWLEDGE TRACING

DATA IMPORT and PRE-PROCESSING

Data was imported using Python Pandas’ read_csv procedure. Please refer to the GitHub repository or attached zip file for README and Jupyter notebook with Python source code.

Imported data frame was turned into a multi-index one...

```
mod1_data = raw_data.set_index(['Student','StepID'])
```

... and was shorted again to make sure everything is in order

```
mod1_data.sortlevel(inplace=True)
```

Columns for the probabilities were inserted.

PROBABILITY FUNCTIONS

u : student
k : skill
function d: skill mastery
function e: correct application of skill in the future

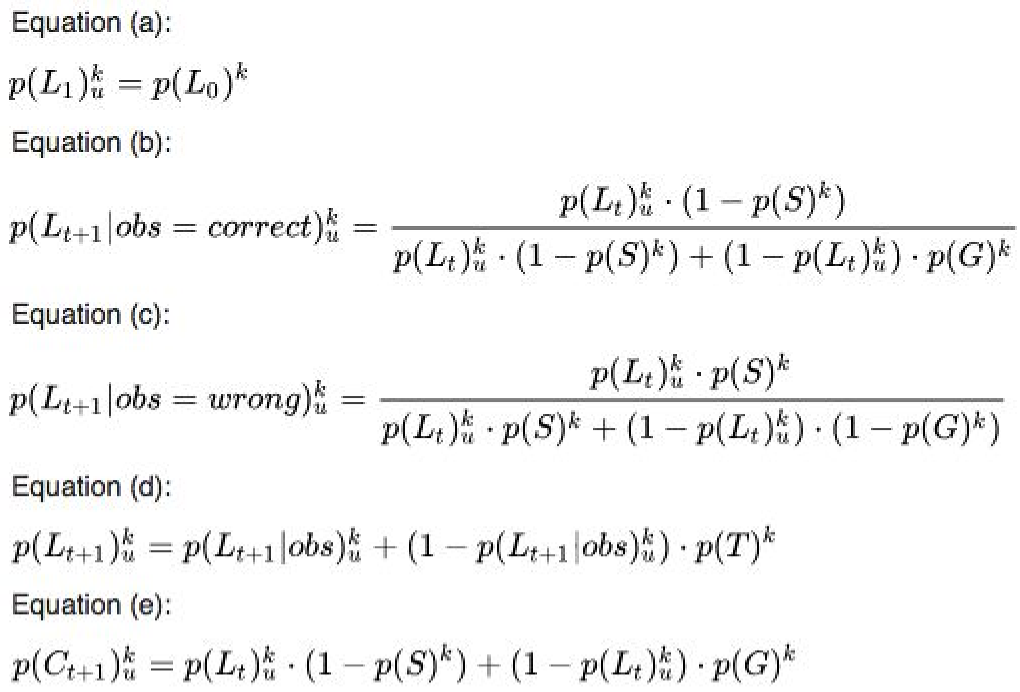
We assume these following parameters are fixed:

P(Lo) = 0.5

P(T) = 0.1

P(S) = 0.1

P(G) = 0.1



Equation (a):

$$p(L_1)_u^k = p(L_0)^k$$

Equation (b):

$$p(L_{t+1}|obs = correct)_u^k = \frac{p(L_t)_u^k \cdot (1 - p(S)^k)}{p(L_t)_u^k \cdot (1 - p(S)^k) + (1 - p(L_t)_u^k) \cdot p(G)^k}$$

Equation (c):

$$p(L_{t+1}|obs = wrong)_u^k = \frac{p(L_t)_u^k \cdot p(S)^k}{p(L_t)_u^k \cdot p(S)^k + (1 - p(L_t)_u^k) \cdot (1 - p(G)^k)}$$

Equation (d):

$$p(L_{t+1})_u^k = p(L_{t+1}|obs)_u^k + (1 - p(L_{t+1}|obs)_u^k) \cdot p(T)^k$$

Equation (e):

$$p(C_{t+1})_u^k = p(L_t)_u^k \cdot (1 - p(S)^k) + (1 - p(L_t)_u^k) \cdot p(G)^k$$

Initial values were set per:

Hawkins W.J., Heffernan N.T., Baker R.S.J.D. (2014) **Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities.** In: Trausan-Matu S., Boyer K.E., Crosby M., Panourgia K. (eds) Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science, vol 8474. Springer, Cham

Previous efforts regarding the science behind the setting of these parameters can be found in papers such as:

Yudelson M.V., Koedinger K.R., Gordon G.J. (2013) I**ndividualized Bayesian Knowledge Tracing Models**. In: Lane H.C., Yacef K., Mostow J., Pavlik P. (eds) Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science, vol 7926. Springer, Berlin, Heidelberg

We can also use machine learning approaches to change these parameters along the way but such approach is beyond the scope of this assignment.

Functions are defined as followed:

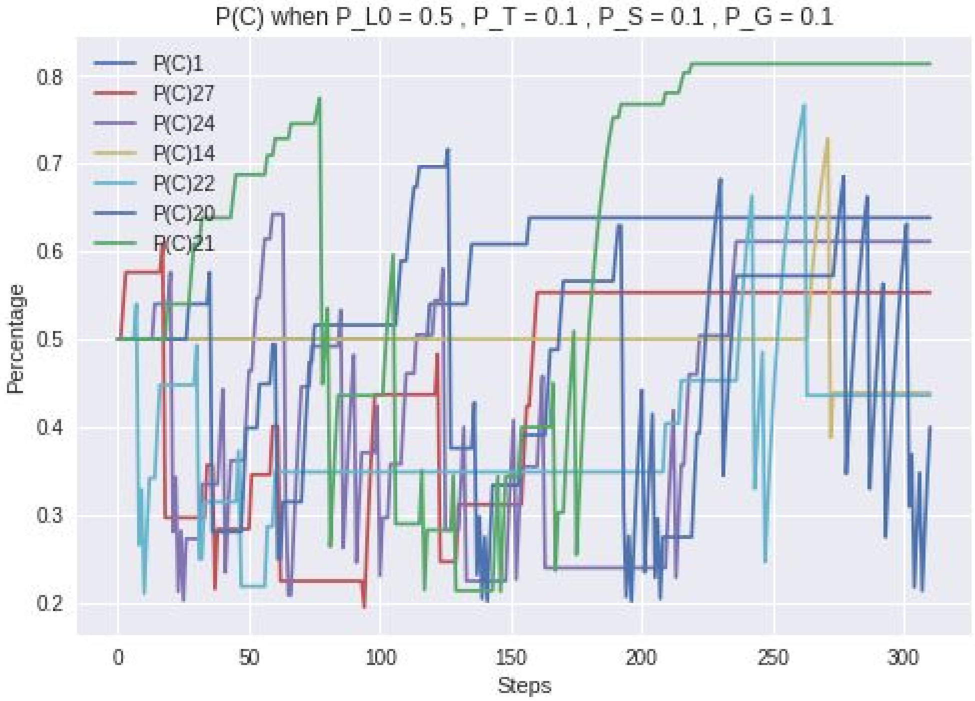
```
def P_L_func ( correct, P_L_previous):
  if correct==1:
   P_L_obs = (P_L_previous*(1-P_S))/(P_L_previous*(1-P_S) +
(1-P_L_previous)*(1-P_G))
 else:
   P_L_obs = (P_L_previous*P_S)/(P_L_previous*P_S + (1-P_L_previous)*(1-P_G))
  P_L_current = P_L_obs + (1-P_L_obs)*P_T
  return P_L_current
def P_C_func (P_L_previous):
 P_C_current = P_L_previous*(1-P_S) + (1-P_L_previous)*P_G
  return P_C_current
```

FILLING IN CALCULATED RESULTS

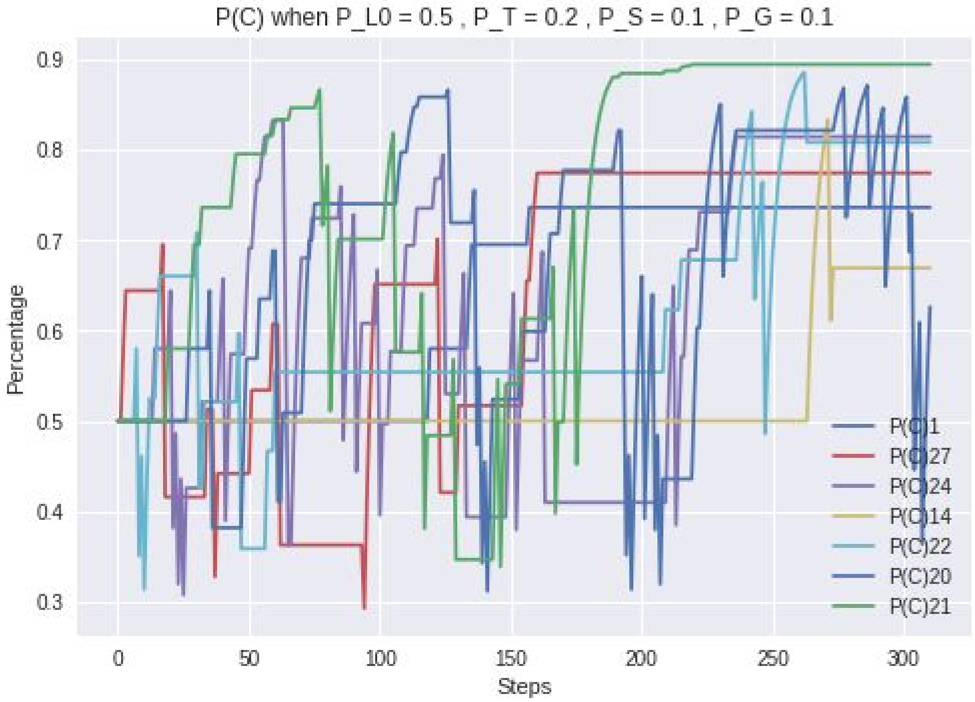
Results will be exported to separate .csv files with the names are the student IDs
Linux “cat” command can then be used to merge all files if necessary.
Sample output files were included in the zip file.
Executing the Jupyter notebook will export results for ALL students. There is a “break” option which was marked with exclaimation marks to limit execution to the first student only (for quick examination)

ANALYSIS

The below graphs show different scenarios with different initial values of P(G), P(S), P(T)



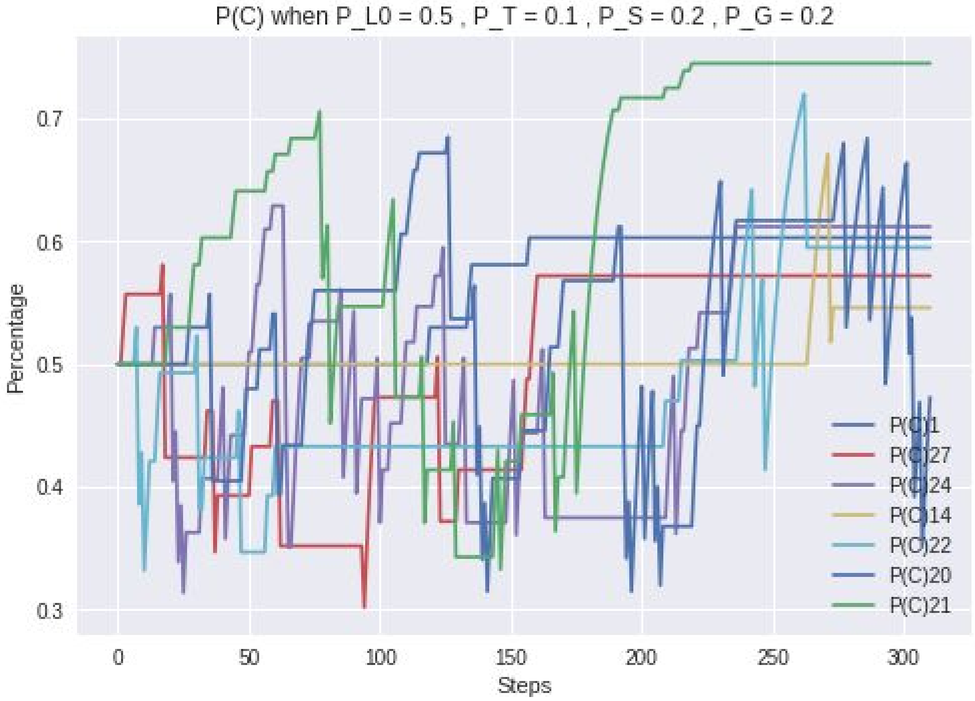
When P(T) got increased by 10%



In this case, P(T) = 0.2 instead of 0.1 by default. We note that the whole landscape remains almost the same but was moved up by 0.1 (10%).

This means P(T) affects P(C) directly and should be considered carefully

When P(S) and P(G) got increased by 10%



We note that the landscape shrinked towards the inside - meaning the bottom line got moved up (10% to 0.3) and the ceiling got lowered down (for approximately the same percentage)

The limits on how far P(T), P(S), P(G) can go

We note that the innitial value of P(L) = 0.5 makes sense. Therefore, other innitial values should not go over 0.5 as well. We can increase P(T) in order to raise the bottom line, increase P(S) and P(G) to reduce the margins between the top and the bottom lines. How much of an increase is dependent on specific domain knowledge as well as familiarity with students' innitial skill levels.