BASIC ASSIGNMENT 3
CORE METHODS IN EDUCATIONAL DATA MINING
PROFESSOR BAKER
FEATURE ENGINEERING
DUE NOON, WEDNESDAY OCTOBER 15

In this assignment, you need to build a Bayesian Knowledge Tracing model for data file AsgnBA3-dataset.csv. This data set is a subset of the data set used in

Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008) Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18, 3, 287-314.

This paper can be found at http://www.columbia.edu/~rsb2162/USER475.pdf

This data set's variables are:

- ID – a unique ID for every student action in the Cognitive Tutor used
- Lesson – the tutor lesson the action comes from
- Student – a deidentified ID for the student
- KC – the knowledge component (skill) involved
- Item – the problem step in the learning system
- Right – is the student action right (1) or not right (0)
- Firstattempt – is this the student's first attempt at the problem step (1)?
- Time – how long did the student attempt take?

You should complete questions 1-9 of this assignment in Microsoft Excel, or a similar spreadsheet program. Question 10, the bonus question, can be completed using Java and BKT-BF, available at http://www.columbia.edu/~rsb2162/BKT-BruteForce.zip

Question 1:

Filter out all actions from (a copy of) the data set, until you only have actions for KC "VALUING-CAT-FEATURES". How many rows of data remain?

Question 2: We need to delete some rows, based on the assumptions of Bayesian Knowledge tracing. With reference to the firstattempt column, which rows do we need to delete?

A) Firstattempt = 1
B) Firstattempt = 0
C) No rows
D) All rows

Question 3: Go ahead and delete the rows which match your answer on the previous question. How many rows of data remain?

Question 4:

We're going to create a Bayesian Knowledge Tracing model for VALUING-CAT-FEATURES. Create variable columns P(Ln-1), P(Ln-1|RESULT), and P(Ln), and leave them empty for now. (If you're not sure what these represent, re-watch the lecture). To the right of this, type into four cells, (cell M2) L0, (M3) T, (M4) S, and (M5) G. Now type 0.3, 0.1, 0.2, and 0.25 to the right of (respectively) L0, T, S, and G (e.g. cells N2, N3, N4, N5). What is your slip parameter?

Question 5: Just temporarily, set K3 to have = I2+0.1, and propagate that formula all the way down (using copy-and-paste, for example), so that K4 has = I3+0.1, and so on (this pretends that the student always gets 10% better each time, even going over 100%, which is clearly wrong… we'll fix it later). What should the formula be for Column I, P(Ln-1)? If you're not sure which of these is right, try them each in Excel. Now, what should the formula for cell I2 be?

A) =IF(C2<>C1, N2, K1)
B) =IF(C2=C1, N2, K1)
C) =IF(C2<>C1, N2, K1)
D) =IF(C2<>C1, N2, K1)
E) =IF(C2<>C1, N2, K1)
F) =IF(C2=C1, N2, K1)
G) =IF(C2=C1, N2, K1)
H) =IF(C2=C1, N2, K1)

Question 6: Propagate the correct formula for column I all the way down (using copy-and-paste). Just temporarily, set J2 to have =I2, and propagate that formula all the way down (this eliminates Bayesian updating, which is not correct within BKT… we'll fix it later). Now, what should the formula for cell K2 be? (Propagate it down)

A) ((1-J2)*N3)
B) ((1-J2)*N3)
C) (J2*N3)
D) (J2*N3)
E) J2 + ((1-J2)*N3)
F) J2 + ((1-J2)*N3)
G) J2 + (J2*N3)
H) J2 + (J2*N3)
I) J2 - ((1-J2)*N3)
J) J2 - ((1-J2)*N3)
K) J2 - (J2*N3)
L) J2 - (J2*N3)

Question 7: What should the formula for cell J2 be? (Propagate it down)

A) =IF(F2=1,(I2*(1-N3))/((I2*(1-N3))+((1-I2)*N3)),(I2*N3)/((I2*N3)+((1-I2)*(1-N3))))
B) =IF(F2=1,(I2*N4))/((I2*N4)+((1-I2)*N5)),(I2*N4)/((I2*N4)+((1-I2)*N5)))
C) =IF(F2=1,(I2*(1-N5))/((I2*(1-N5))+((1-I2)*N4)),(I2*N5)/((I2*N5)+((1-I2)*(1-N4))))
D) =IF(F2=1,(I2*(1-N4))/((I2*(1-N4))+((1-I2)*N5)),(I2*N4)/((I2*N4)+((1-I2)*(1-N5))))

Question 8: If a student starts the tutor and then gets 3 problems right in a row for the skill, what is his/her final P(Ln)?

A) 0.856
B) 0.950
C) 0.955
D) 1.000

Question 9: If a student starts the tutor and then gets 3 problems wrong in a row for the skill, what is his/her final P(Ln)?

A) 0.046
B) 0.142
C) 0.154
D) 1.000

Question 10 (BONUS): Run this data through BKT-BF, available at http://www.columbia.edu/~rsb2162/BKT-BruteForce.zip . What is the value for P(G) for VALUING-CAT-FEATURES? (Note: this will take a few minutes to run even on a good computer.)

A) 0.100
B) 0.299
C) 0.541
D) 0.635