Seoyeon’s journal

<run bruteforce bkt code>

-download ‘[Geometry Area Study Spring 2014](https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=986)’ from datashop ([**Geometry Cognitive Model Discovery Closing-the-Loop)**](https://pslcdatashop.web.cmu.edu/Project?id=68)

-making input file for ‘computeKTparamsAll.java’ (Bruteforce bkt java code)

: it should be ‘num lesson student skill cell right eol’ format.

:student- id/ skill-kctracedskill(convert space to ‘-’)/ right - first attempt(1; correct, 0;incorrect&hint)

-copy and paste the input text file to java package.

-run as argumentation-set arguments: input text file.

-change the last part as the name of the text file.

-get 4 parameters for bkt:)

-안돌아가면 스킬이름 ‘-’ 살려야함.

<run bkt code>

-run bkt\_python code with python. (wd: run - configuration)

-bktoutput file(geometry data-> geometry0913)….-> read bkt article again to find out why the updated L value are different and can I have the value 1?

-이거 돌릴때도 rb-> r/ wb-> w

------------

back to sanjay’s data.

I added ‘decision’ from ‘compromised data non-mastery’ data to ‘value output sanjat’ data. -> ‘sanjay&decision’ data. (exclude cells with ‘~~’)

---------------

download sanjay’s raw data. (wheelspinning seoyeon -> geometrydata-> sanjayraw)

1. exclude blank KC
2. exclude ~~ KC and opportunity
3. created ‘ makingindevar’ data for making independent variable from the raw data.
4. making variables on ‘excoppour’ sheet.
5. creating ‘IDKC’ column -> ID&”,”&KC

|  |  |  |
| --- | --- | --- |
| name | meaning | function |
| guess | * How often the student was rapidly guessing, computed across all skills, defined as submitting responses less than 2 seconds apart on successive items. We took the geometric mean in the same manner as for response time.   (Beck&Gong, 2013) | =IF(J3<=2,1,0) |
| sum of guess | sum of guess per IDKC | =VLOOKUP(R3,Sheet11!$D:$E,2,FALSE) |
| sdaverage | average of step duration per IDKC | =VLOOKUP(R4,Sheet4!$D:$E,2,FALSE) |
| cdaverage | average of correct step duration per IDKC | =VLOOKUP(R5,Sheet4!$D:$I,6,FALSE) |
| sumofincorrect | sum of incorrect per IDKC | =VLOOKUP(R4,Sheet4!$D:$F,3,FALSE) |
| sumofhints | sum of hints per IDKC | =VLOOKUP(R3,Sheet4!$D:$H,5,FALSE) |
| maxhints | max hints per IDKC | =VLOOKUP(R2,Sheet4!$D:$J,7,FALSE) |
| sumofcorrects | sum of corrects per IDKC | =VLOOKUP(R2,Sheet4!$D:$G,4,FALSE) |
| KCcircle~KCsubtracted | if KC has a ‘circle’(~’subtracted’) word, it shows 1. | =IF(ISERROR(FIND("circle",AA2)),0,1) |
| KCsum | sum of KCcircle~KCsubtracted per IDKC | =SUM(AB2:AV2) |
| MaxOppor | Max opportunity per IDKC | =VLOOKUP(R2,Sheet11!$D:$F,3,FALSE) |
| averagepEr | average of predicted error rate per IDKC | =VLOOKUP(R2,Sheet4!$D:$K,8,FALSE) |
| bottomouthint | if student asked hints >=3, the value is 1. | =IF(U2>=3,1,0) |
| averageL | average L value per IDKC |  |
| averageDiffL | average the difference of L value per IDKC | if two cells are the same, calculate the difference. |

Matsuda said me to make a dataset as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| student | KC | w/m | #observation | average independent variables | …. |
| s1 | kc1 | c(mastery) | 9 | 3.7 | ... |

\*making pivot table ( visualize first)

|  |  |  |
| --- | --- | --- |
| student | kc | average of sth |
| row | row | value |

>>try logistic regression ( rapidminor& r)

rapidminor error -> dependent variable should be binominal! (not just categorical)

r -> ‘sdaverage’ made an error so deleted it.

I ran all the independent variable respectively, and found ‘observation/sumofincorrect/bottomouthint/sumofhints/sumofcorrects/kcsum/MaxOpp/averagepEr/averageL/averageDiff.of.L’ variable works (when just simple LR)

# Logistic Regression

dataset = read.csv('valueoffinal.csv')

View(dataset)

dataset = dataset[,5:18]

dataset$numdeision = as.factor(dataset$numdeision)

str(dataset)

library(caTools)

set.seed(123)

split = sample.split(dataset$numdeision, SplitRatio = 0.75)

training\_set = subset(dataset, split == TRUE)

test\_set = subset(dataset, split == FALSE)

result = glm(numdeision ~ .,

family = binomial,

data = training\_set)

result

summary(result)

anova(result)

>the result with all the variables

Call:

glm(formula = numdeision ~ ., family = binomial, data = training\_set)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.3274 0.1312 0.1824 0.2728 1.2957

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.538657 2.530489 0.608 0.54316

observation 0.109167 0.109539 0.997 0.31896

sum.of.guess 1.058899 0.607470 1.743 0.08131 .

cdaverage 0.004289 0.022367 0.192 0.84794

sumofincorrect 0.058885 0.067167 0.877 0.38065

bottomouthint -1.248812 0.620373 -2.013 0.04411 \*

sumofhints 0.066293 0.036561 1.813 0.06980 .

maxhints 0.079983 0.194069 0.412 0.68024

sumofcorrects -0.245872 0.141667 -1.736 0.08264 .

Kcsum -0.436606 0.671734 -0.650 0.51571

MaxOpp 0.002373 0.089848 0.026 0.97893

averagepEr -2.608471 2.031637 -1.284 0.19917

averageL 5.290605 1.886558 2.804 0.00504 \*\*

averageDiff.of.L -13.869195 6.166710 -2.249 0.02451 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 220.67 on 543 degrees of freedom

Residual deviance: 160.58 on 530 degrees of freedom

AIC: 188.58

Number of Fisher Scoring iterations: 7

I changed the code for the stepwise regression( I also did backward but the result was the same)

> summary(result)

Call:

glm(formula = numdeision ~ sum.of.guess + sumofincorrect + bottomouthint +

sumofhints + sumofcorrects + averageL + averageDiff.of.L,

family = binomial, data = training\_set)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.1396 0.1397 0.1853 0.2694 1.3509

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.63770 0.98828 -0.645 0.5188

sum.of.guess 0.72706 0.47529 1.530 0.1261

sumofincorrect 0.07969 0.05177 1.540 0.1237

bottomouthint -1.16156 0.55267 -2.102 0.0356 \*

sumofhints 0.05834 0.02711 2.152 0.0314 \*

sumofcorrects -0.16224 0.07846 -2.068 0.0387 \*

averageL 5.74982 1.12301 5.120 3.06e-07 \*\*\*

averageDiff.of.L -12.64733 5.76039 -2.196 0.0281 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 220.67 on 543 degrees of freedom

Residual deviance: 163.39 on 536 degrees of freedom

AIC: 179.39

Number of Fisher Scoring iterations: 7

>>>this data got skewness problem……..

>>>>find out data scum….(Sanjay’s data handling was wrong…)

|  |  |  |  |
| --- | --- | --- | --- |
| Stu\_6cc7f80e4c8102717d28a1c652920c01 | C | 1 0 0 1 1 1 | Find added area |
| Stu\_6cc7f80e4c8102717d28a1c652920c01 | W | 0 1 1 0 1 1 1 0 | Find added area~~Find individual area~~Successfully pose component areas |

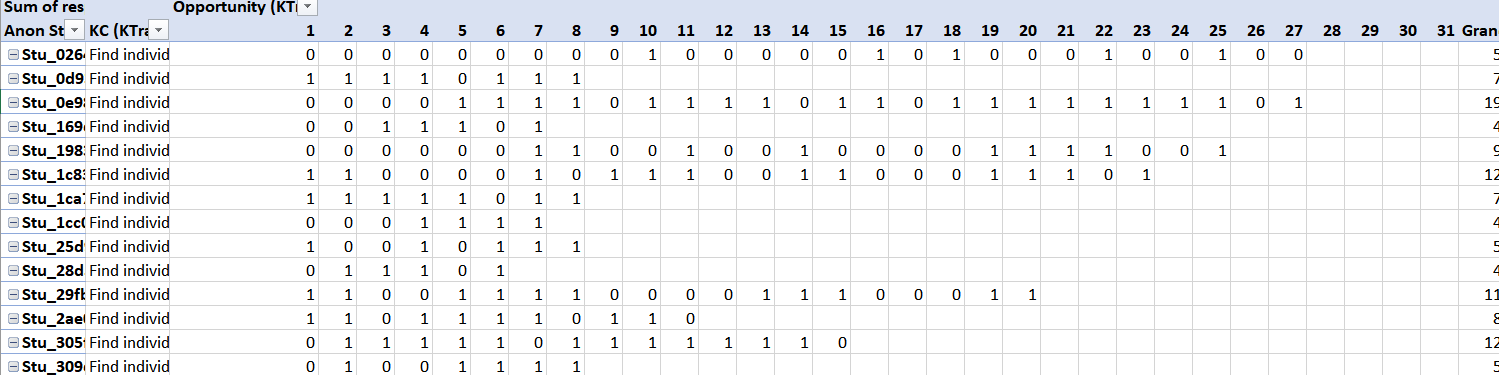
>>>>this means that the student actually tried more than 6 times on the ‘find added area’ skill. so we need to split up and combined/ or deleted both columns

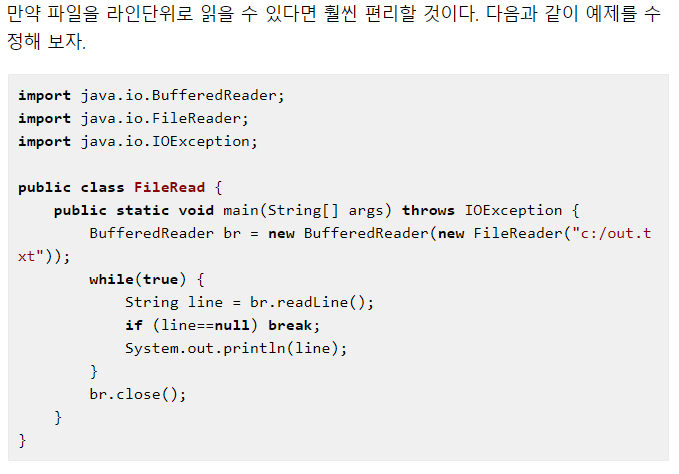
delete those students’ columns -> valueoffinal .

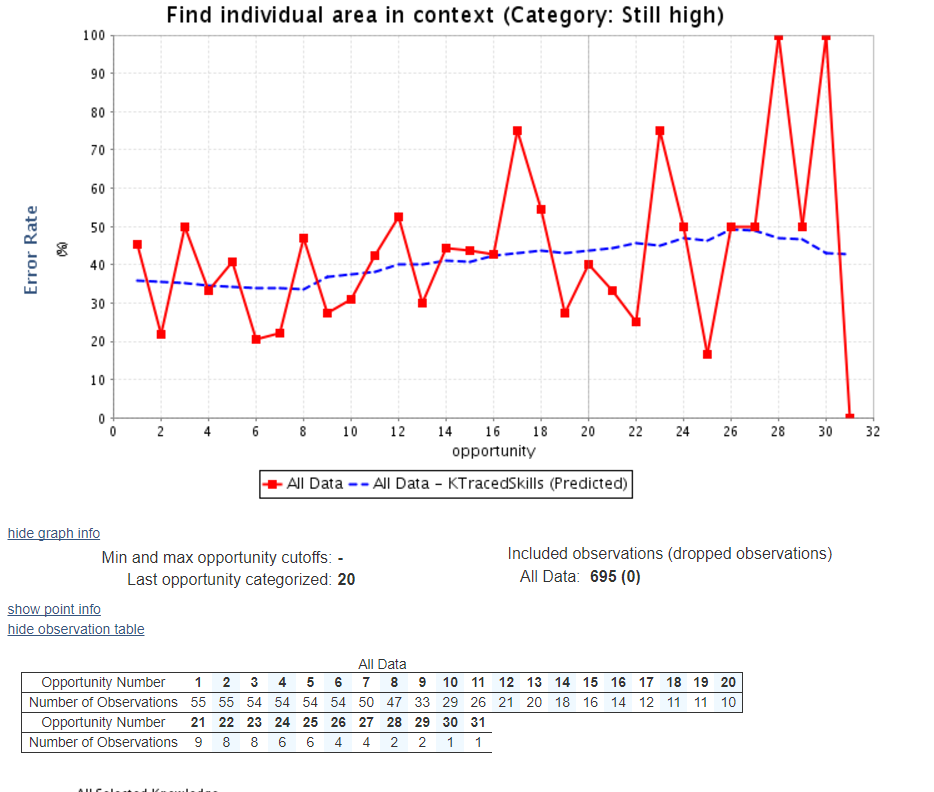
>>>>noboru said I could start to focus on the skill that shows wheel-spinning

so extracted data on ‘find the individual area’ skill. -> made ‘pilot data’folder ‘ in onedrive.

>>to make response sequence, make pivot table-> cancatenate function.

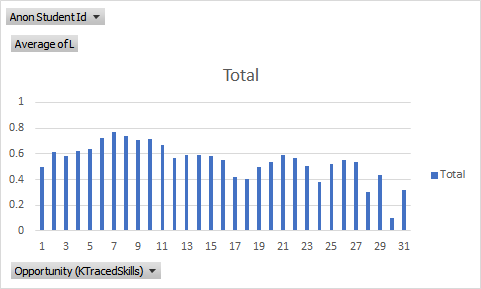






-interpreting above Opportunity 1 has 55 observations. red line is actual error rate(the number of incorrect /55 maybe). blue line is predicted error rate.

>>>Average of L per opportunity



>>>Matsuda want me to make predicted learning curve using logistic regression…

log(responce...correct/incorrect firstattempt)=b0+b1\*Opportunity

P(response)=f(Opportunity)

dataset <-findtheindividualarea\_main

dataset = dataset[, c(12,18)]

str(dataset)

result= glm(formula =response ~ Opportunity,

family = binomial,

data = dataset)

result

summary(result)

coef(result)

exp(coef(result))

prob\_pred = predict(result, type = 'response', newdata = dataset)

dataset\_pred<-data.frame(dataset,prob\_pred)

dataset\_pred

write.csv(dataset\_pred,"dataset\_pred.csv")

> summary(result)

Call:

glm(formula = response ~ Opportunity, family = binomial, data = dataset)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4824 -1.3792 0.9176 0.9614 1.1766

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.71635 0.13170 5.439 5.35e-08 \*\*\*

Opportunity -0.02305 0.01194 -1.930 0.0536 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

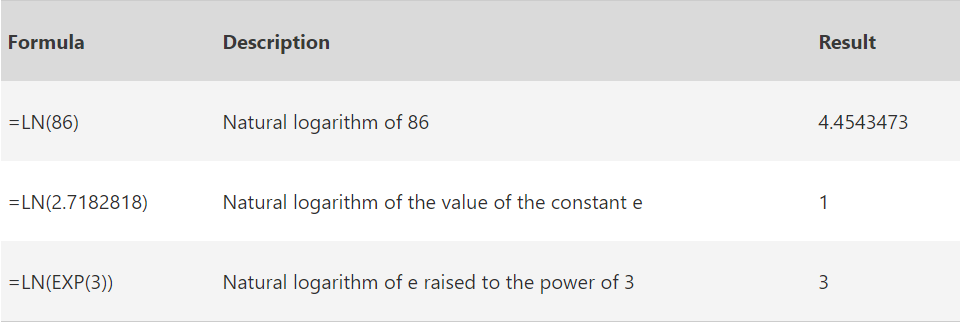
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 918.93 on 694 degrees of freedom

Residual deviance: 915.22 on 693 degrees of freedom

AIC: 919.22

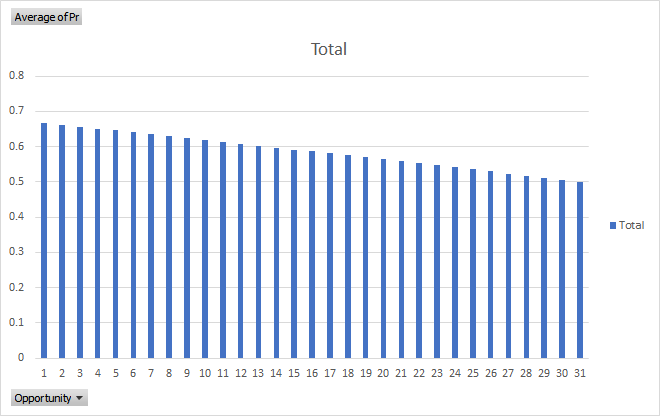
Number of Fisher Scoring iterations: 4



>>so the log(Pr/1-Pr)=0.71635-0.02305\*Opportunity

>>calculate P(response)

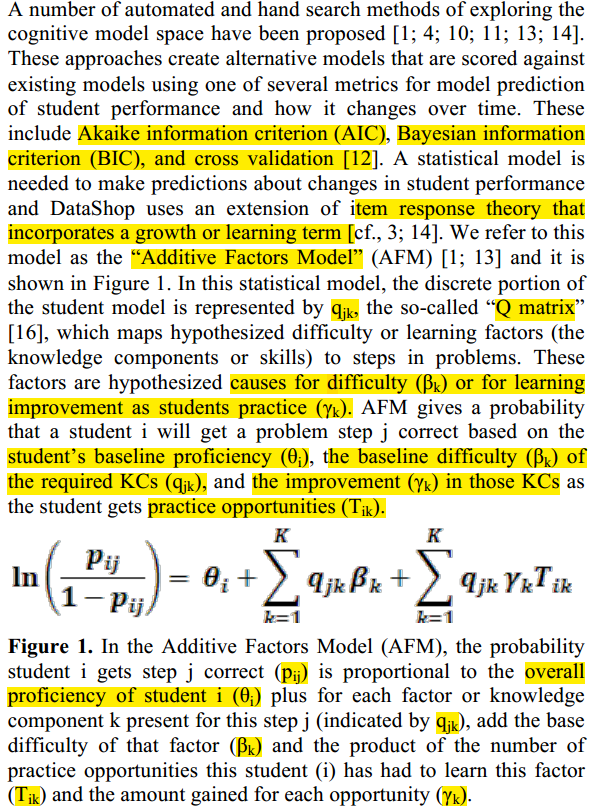
|  |  |  |
| --- | --- | --- |
| log(pr/1-pr) |  | =0.71635-0.02305\*R2 |
| e^log(pr/1-pr) |  | =EXP(U2) |
| 1+e^log(pr/1-pr) |  | =1+V2 |
| Pr |  | =V2/W2 |



>>this graph doesn’t match with the learning curve from the datashop.

>>pick another KC(find the trezoid area) and try again to make same graph. (check the student number!) - # is 903.

>>the skill was not the matter…..the matter is the datashop use different function to calculate the predicted error rate. (Additive Factors Model)



<https://pslcdatashop.web.cmu.edu/help?page=modelValues>

Using R notation, the AFM model (applied to a modified student-step export file called "ds") can be approximately\* represented as:

R> L = length(ds$Anon.Student.Id)

R> success = vector(mode="numeric", length=L)

R> success[ds$First.Attempt=="correct"]=1

R> model1.lmer <- lmer(success~knowledge\_component+  
 knowledge\_component:opportunity+(1|anon\_student\_id),data=ds,family=binomial())

f ratio of interation in regression.

하...여기서부터 내 일주일이 꼬이기 시작….

-try to run that r code to get the same predicted error rate of datashop.

-to this, I need to run the student-step roll up data with whole skills..-> it didn’t work because it need a size over 24 giga.

-so I downloaded another dataset (digital game dataset) from datashop which is much smaller and pick ‘benchunit’ KC.

-create ‘smaller.csv’data.

-run that r code to smaller data.

library(lme4)

ds<-read.csv("smaller.csv")

str(ds)

L=length(ds$Anon.Student.Id)

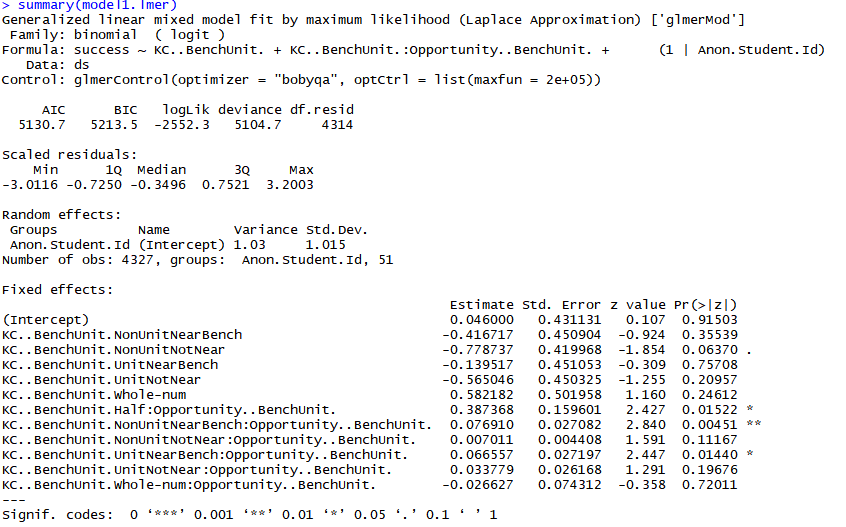
success=vector(mode="numeric",length=L)

success[ds$First.Attempt=="correct"]=1

model1.lmer<-glmer(success~KC..BenchUnit.+KC..BenchUnit.:Opportunity..BenchUnit.+(1|Anon.Student.Id),data=ds,family=binomial,control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=2e5)))

model1.lmer

-뒤에 control function은 자꾸 에러나서 그거 같다 붙엿더니 에러 없어짐.



-using this ourcome, compute predicted error rate.

>>>z: kcopp coefficient \* Opportunity + kc coeffecient+ Intercept

>>my error rate(1-p) and predicted error rate doesn’t match ……….I don’t know...Noboru doesn’t know……

---so look through how datashop get the error rate again and email to them….!

>>matsuda doesn’t remember why he asked me to do this logistic regression thing……..

matsuda suggests to calculate 1-predicted error rate as dependent variable.

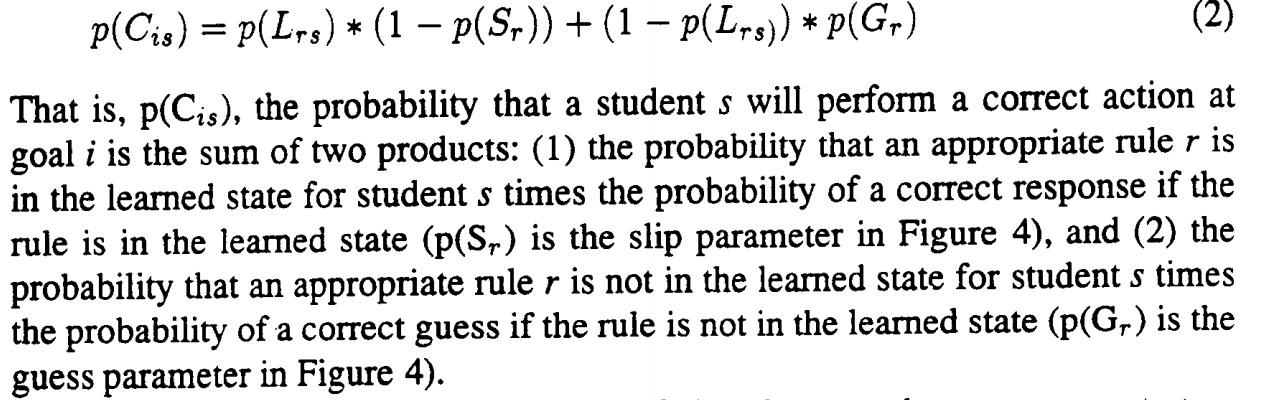
>>brainstorm how we define wheel-spin/master and its criteria

* master/ interminate/ whel-spin :
* Defining mastery may vary between systems. One measure of mastery includes next problem correctness, another is performance on a transfer question, and yet another is performance on a delayed retention test.

>> told matsuda that L value cannot be a dependent variable on wheel spinning. talked about ‘identifiability issue’. reading Joe beck’s and Yan wang’s paper about that.

>> matsuda said using performance curve then. he told me to compare the performance curve and 3 correct. calculate the performance curve using bkt.

>> performance curve calculation. (corbett -knowledge tracing article)



>> calculate performance data using (L). in criteria 0.95, there was no mastery. all “W”.

so Matsuda said why we need to use 3-in-row. justification. we cannot use this because some said so. we need to set up our reason.

>>이제 transaction data 가지고 놀아보장.

|  |  |  |
| --- | --- | --- |
| Category | Question | Reference |
| reducing detection speed | Can we detect WS on 7th OPP?  : After PO 7, Wheel Spinning is the majority class. | Kelly, Wang, Thompson, & Heffernan(2015) : 14 OPP  Beck&Gong(2013) : 10 OPP |

how? - maybe time series analysis. because we might be able to analyze what feature made a signal WS before it happened.