User Activity Measures in a MOOC

Ilia Rushkin

Problem: create a measure of user activity level in a HarvardX MOOC that can be applied across courses.

Solution overview

Among the readily available variables from HarvardX data, the following four are identified as likely proxies for activity by a user in a course:

1. *nevents* – the total number of logged events
2. *nplay_video* – number of clicks "Play" on videos (a proxy measure of user engagement with videos)
3. *nproblem_check* – number of problem submits (a proxy measure of user engagement with assessment)
4. *nchapters* – number of course chapters visited (a proxy measure of course exploration)

We combine these into a single composite variable, denoted *activity index*. This variable has continuous numeric values. We are also interested in a discrete variable *activity level.* We prepare it by clustering users by their activity values, treating each course separately. The number of clusters we use is 5, thus we group users into four levels of activity: 0,1,2,3,4. We suggest naming these activity levels "none", "low", "moderate", "consistent", and "high".

If necessary, the activity levels can be further collapsed to just two, thus forming a Boolean variable: for instance, the users in the top two activity levels are labeled "active". The percentage of active users will vary across courses. Based on courses from 2015 and 2016, we estimate that it is (21±8)% of registered users.

Solution details

Denote $x_i = \ln(1 + V_i)$, where V_i are the above mentioned variables (V_1 is the number of logged events, V_2 is the number of "Play" clicks on videos, etc.). The reason for the logarithmic transformation of variables is that without it the distribution of users in these variables is highly skewed.

If we change the list of participating variables V_i in the future (e.g. add more variables to it) – the procedure below will remain the same.

Forming the *activity index*:

1. Remove the users for whom all V_i are zero. These users have no activity at all and will be treated separately (the activity index value for them will be missing or, if you prefer, set

it to negative infinity). Normalize all x_i to variance 1 and mean 0. If for one or more variables the variance is zero, drop this variable from the set.
2. Perform the EFA (exploratory factor analysis) routine with 1 factor on the remaining variables x_i. In the rare event that this does not converge, substitute EFA with PCA, i.e. perform instead the principal component rotation on the user data and take the user scores along the principal component with the largest eigenvalue. Multiply the obtained factor (whether from EFA or from PCA) by the sign of the sum of its loadings, to make sure the vector direction is not flipped. To make it comparable across courses, normalize these user scores to variance 1 and center to mean 0, and call the resulting user variable *activity index*.
3. The squares of loadings can be viewed as the *weights* of the original variables V_i in *activity index*. They add up to 1.

Forming the categorical *activity level* variable:

1. Cluster users in a course by their activity value, using the x-means algorithm, which chooses the optimal (in some commonly accepted sense) number of clusters K as dictated by the data. Typically, this yields a low number of clusters (for 70% of HarvardX MOOCs from 2015 and 2016, $K \leq 7$, the median number is 6). This convinces us that it is reasonable to set $K = 4$ across all courses, as it is a nicely interpretable number. Thus, we now use the k-means algorithm to split users in each course into 4 clusters: 1,2,3,4 in the order of increasing centroids. The fifth bottom cluster (0) of users with zero in all variables V_i is added to these by hand.

Course examples

Below we use the median time on task as one possible interpretation variable. In practice, we calculate and make available the mean and median values in each activity level for all variables V_i, as well as the time on task.

HarvardX/ENGSCI137x/2T2016:
The activity variable is a mixture of: 29% *nevents*, 24% *nchapters*, 25% *nplay_video*, 22% *nproblem_check*. Breakdown of registered users by activity levels:

Level of activity	0 (None)	1 (Low)	2 (Moderate)	3 (Consistent)	4 (High)
% of users	31%	23%	23%	15%	7%
Median time on task (minutes)	0	0	4	58	512

Harvardx/HLS2X/T12016:
The activity variable is a mixture of: 28% *nevents*, 25% *nchapters*, 24% *nplay_video*, 24% *nproblem_check*. Breakdown of registered users by activity levels:

Level of activity	0 (None)	1 (Low)	2 (Moderate)	3 (Consistent)	4 (High)
% of users	42%	17%	18%	12%	12%
Median time on task (minutes)	0	0.12	11	87	516

HarvardX/PH525.1x/2T2016:
The activity variable is a mixture of: 29% *nevents*, 22% *nchapters*, 25% *nplay_video*, 23% *nproblem_check*. Breakdown of registered users by activity levels:

Level of activity	0 (None)	1 (Low)	2 (Moderate)	3 (Consistent)	4 (High)
% of users	24%	12%	35%	18%	10%
Median time on task (minutes)	0	0.41	5	31	229

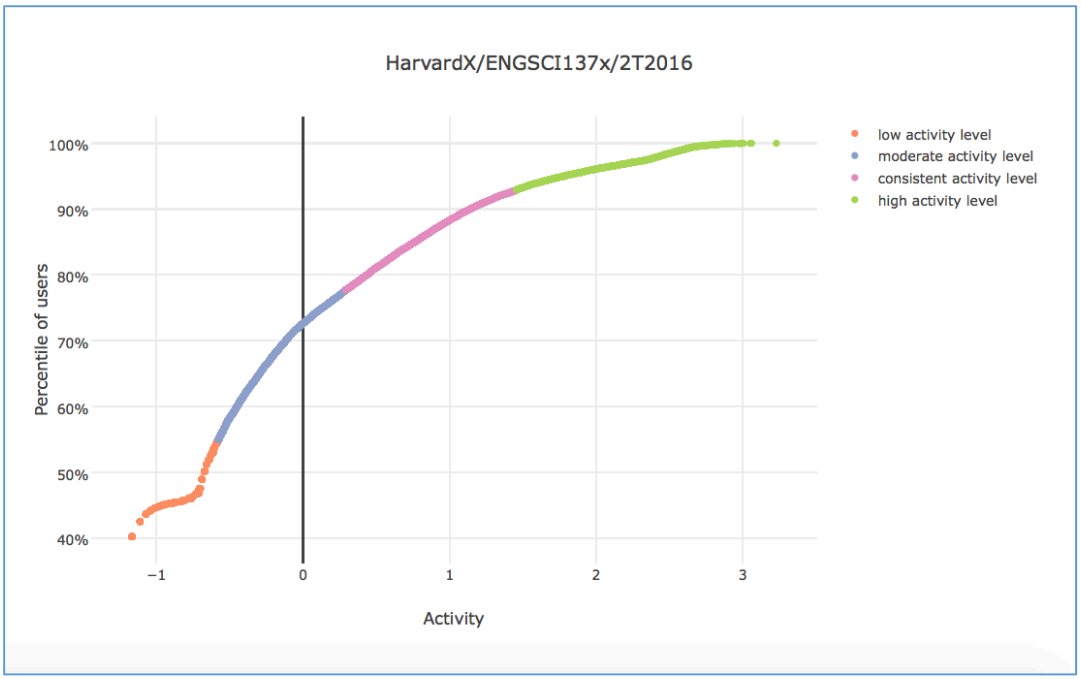

Figure 1.

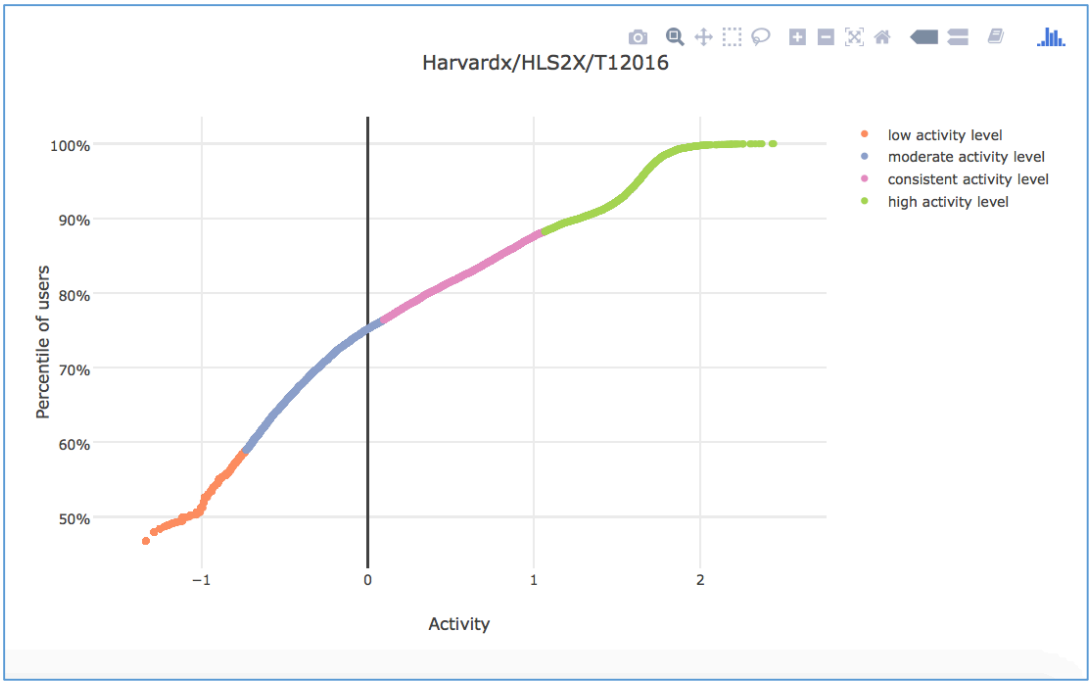

Figure 2

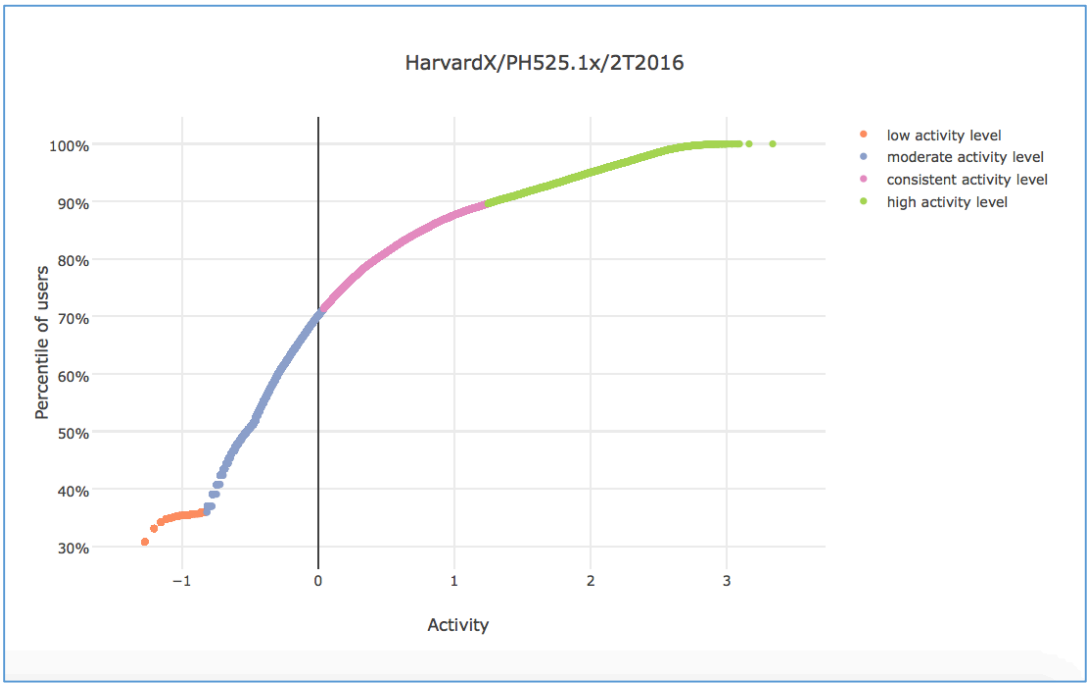

Figure 3