Group: Dinosaur Rodeo (Clowns)
Members: Sungmin Kim, Matt Winkler, Abram Beyer, Hyunyou Choi
September 28, 2016

Cleaning & Transformations on EEG Dataset

Data Collection and Motivation

The dataset was collected in 2011 by a research team at Carnegie Mellon university[1]. Their goal was to better understand how we can measure confusion levels in the brain, as a means of better informing MOOC course development for online education. The authors argue that this knowledge is potentially valuable because of several important differences between online courses and traditional in-person classroom education. Specifically, instructors are able to diagnose confusion levels among students (apart from students asking questions directly) by looking for body language or asking the class questions periodically when lessons are given in-person. In an online setting, however, that's not possible.

The research team sought to use EEG measurements, which have previously been shown to have a relationship with confusion levels, as a means of helping resolve that issue. They note the increased availability of low-cost, mass-produced eeg sensors powerful enough to give accurate feedback on users' mental states. They used the same types of sensors in their study to record EEG measurements while participants watched a series of videos whose content was deliberately made more confusing by being presented out of order. The study has potential to impact the online education domain by closing the gap between how learners experience education in that format versus in-person classes. If MOOC producers are able to user data to understand which parts of their material cause confusion among students, they may be able to address certain shortcomings in the content they produce. According to the authors of the study, previous research indicates that eeg activity has been shown to related to mental states, which also have a relationship to learning.

[1] http://www.cs.cmu.edu/~kkchang/paper/WangEtAl.2013.AIED.EEG-MOOC.pdf

Merging and Cleaning

Our first step was to merge the demographic data for each student with the observational data to create one dataset for analysis. We checked the row counts before and after the merge to ensure that it worked as expected. After the merge, our dataset contained 12,811 observations across 18 variables (2 id variables, 14 predictor variables and 2 target class variables). We didn't detect issues with missing values, or data that appeared to be recorded in the wrong column. The main problem we did identify in the dataset appears to be related to technical problems with the data recording during the experiment. After summarizing the means of the numeric attributes for each subject in the study, we noticed that the values for ID 6 across all attributes are very different from those for the other subjects. Here is a screenshot with means for all numeric features by subject ID:

Out[19]:

	Video ID	Attention	Meditation	Raw	Delta	Theta	Alpha 1	Alpha 2	Beta 1	Beta 2
subject ID										
0	4.313243	48.630452	48.571768	33.042030	684941.555908	165919.786677	35990.755749	24487.685964	20288.437748	37128.914354
1	4.461184	45.561107	57.935434	33.599539	428179.287471	125888.289777	31765.673328	22985.478094	20740.016141	13590.006918
2	4.433801	55.971184	60.865265	33.904984	39549.890966	19263.661994	7781.796729	8248.434579	7975.485202	8410.145639
3	4.432268	40.044140	45.648402	34.219178	780844.464231	221221.465753	52551.698630	34068.783105	23391.176560	24573.475647
4	4.440154	44.895753	52.888803	38.635521	363673.088031	108926.853282	25011.400772	19826.601544	17363.406950	15166.986100
5	4.315372	46.616482	54.213946	34.487322	751906.218700	241902.458003	58220.128368	39022.931854	26944.388273	23847.141838
6	4.408627	0.000000	0.000000	348.143529	748472.609412	251335.407843	66816.250980	78602.474510	61069.448627	160509.882353
7	4.387931	44.844044	50.167712	36.307994	568962.452978	130318.122257	30058.974138	20911.165361	18988.803292	28899.112853
8	4.393916	43.990640	50.177847	35.666147	830712.153666	236931.875195	54706.961778	38513.213729	26461.799532	24249.097504
9	4.313243	42.507534	51.110230	29.070579	870734.973830	180547.519429	51447.114195	45816.708168	20216.413957	46573.854084

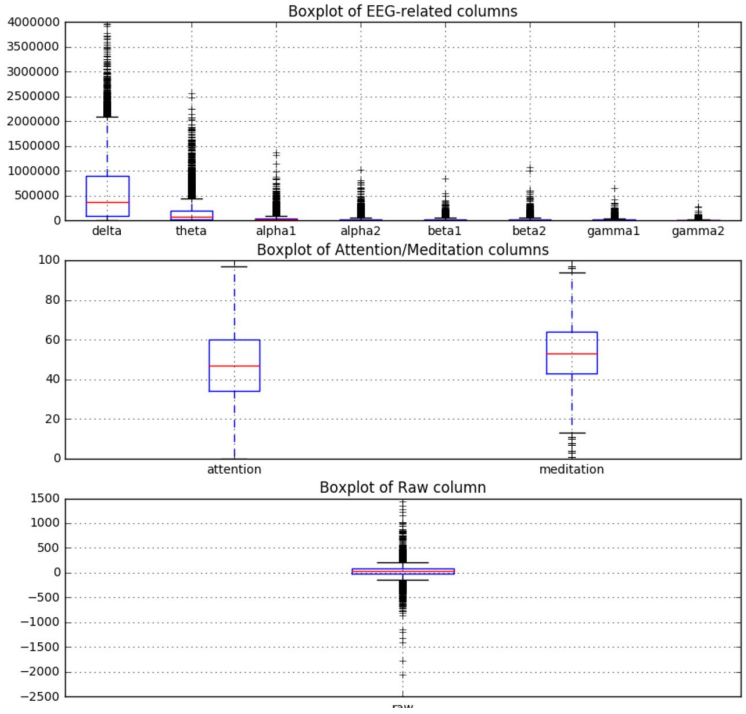
Since there are only ten subjects in the experiment, we thought about trying to replace the faulty data by modelling the good data from the rest of the subjects. However, we think that the EEG readings are likely to be highly individualized, to a degree that inferring the values for a given student based on the others didn't make sense. Another problem with that approach is that we only have 10 unique individuals in the original dataset, so generating a solid model to predict any one student's values for the 9 eeg-related columns based on some demographic information seems like a difficult task. So, we decided to drop all records for subject ID 6 (Male, 24 yrs old, Han Chinese) from the dataset.

Sampling

We also noticed in our exploration that there are different numbers of observations for each subject in the study. The student with the fewest observations makes up 1,261 cases total. So, we sampled the original dataset to include 1,261 observations for each of the nine participants (down from ten with the removal of subject ID 6), giving us a total of 11,349 in the initial analysis data. After sampling the data, we discretized the categorical features (gender and ethnicity), and then looked at the distributions across each of the numeric features to gauge their skew and scale.

Transformations and Normalization

On initial inspection, we noticed that the eeg-related features are on a much different scale (tens of thousands) than the attention, meditation, and raw features. Box plots of the eeg features, attention, meditation, and raw are shown here:

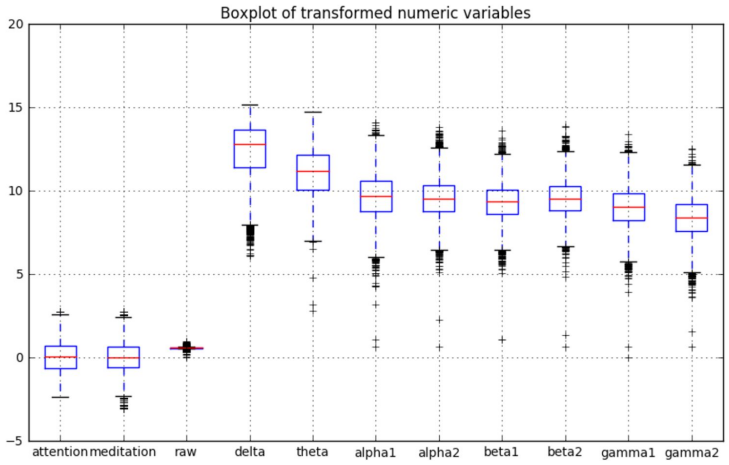


Beyond the difference in scale, it's apparent that the eeg variables all tend to be right-skewed, while attention / meditation are much more normal. Raw follows a distribution with a very high kurtosis.

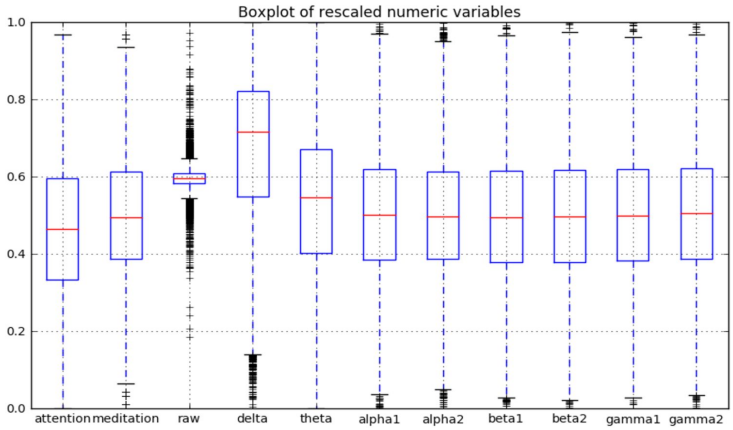
So, it was apparent that we needed to apply different transformations depending on the variable. Based on this analysis, we applied the following:

- **EEG variables**: log transform
- **Raw:** Min / max normalization on a 0-1 interval
- **Attention / Meditation**: Z score normalization

After those transformations, the scales of each feature were brought much closer together, though we still saw some issues in that respect. There are clearly also outliers in the eeg features.



Our next step was to remove those outliers. To do that, we dropped observations which fell greater than 3 standard deviations above / below the mean of the transformed values for each feature obtained in the previous step. After removing outliers, we further normalized all of the numeric features using min-max scaling on a 0-1 interval. The results are shown in the plot below:



We still feel that there may be more work to do on some of the features. The raw variable still has high kurtosis, and the delta variable's mean is higher than the rest of the data. In general, we are also curious as to whether applying these transformations and normalizations will improve our ability to use the data for prediction. Specifically, it seems possible that eeg data behaves differently from other datasets we've encountered in the past, such that we may do better by focusing on the extreme values in the readings, such as the top and bottom eeg readings. It seems plausible that a spike or trough in neural activity as measured by the eeg recording device is more indicative of a change in mental state than averaged values.

Aggregation

Our initial plan for analyzing this dataset was to use the results of the steps described above, where each observation is a (transformed and normalized) reading from the eeg sensor + demographic information. However, after discussing the dataset more, we decided to compare the performance of our models using two different levels of the data. One dataset contains the normalized and transformed values discussed above. The second is an aggregation of the first, where we're grouping all the data from part one at the subject ID / video ID level. We are curious to see how aggregating the dataset affects the accuracy of our models compared with the dataset at the more granular level. There may also be opportunities to sample from the original dataset to focus on the more extreme values in the eeg readings.

Other concerns

Looking at the breakdowns in demographics, there are definitely some aspects of the dataset which may make it difficult to generalize the results of our analysis. They are:

- There were 8 men and 2 women included in the raw dataset
- 7 of 10 subjects were Han Chinese.
- All subjects are in their 20s and early 30s

These aren't things about the data that we're able to change, but they will impact the applicability of the results to some degree. Another concern we have is whether it might be necessary to adjust for effects of certain subjects and/or videos on the eeg readings or confusion levels. Certain subjects may be more prone to perceived confusion than others given the same eeg readings, so we may need to adjust for that effect to improve the reliability of our estimates. Certain videos may also provoke more confusion than others.