Assignment 2 (W4242)

This assignment is due on Wednesday, October 16, 6:10pm before class. You can discuss with other people, but make sure write down their names on your homework.

Q1. (Note, for this question you need to form a group of three to four people, any smaller or larger is not acceptable.. Form a group as soon as possible. If you do not have a group you will not be able to complete this question.)

You are working at a hot tech company, and your R&D department has just developed a new brain-computer interface. The device reads how people are feeling and is small enough to be built into a laptop. It can distinguish between four states: happy, sad, frustrated and bored. Unfortunately, it doesnt work perfectly. They give you a confusion matrix of how well it would work.

	predicted happy	predicted sad	predicted frustrated	predicted bored
actually happy	80%	5%	5%	10%
actually sad	5%	60%	30%	5%
actually frustrated	5%	25%	65%	5%
actually bored	2%	2%	1%	95%

a.(group) Design a product that leverages this new brain-computer interface. In 1-2 paragraphs, describe your target audience and your project and a high level idea of how it works.

b.(individual) Each group member must come up with two storyboards that describe the experience of using your product, each storyboard should have one paragraph explaining it. Your product must be complex enough to have enough meaningful storyboards.

For each storyboard, pick a primary task to be done with your product - make it a task complex enough to fill up a storyboard. Each storyboard should require 5-8 panels, so in total you will have 15-24 panels to turn in. Each storyboard should fit on two 8.5” × 11” sheets of paper and be drawn with a thick pen like a Sharpie. Using a thick pen limits the amount of detail that you can add, forcing you to only draw the most important elements of scenario, user, and interface that communicate your ideas. Clarity, communicativeness, and innovation are more important than aesthetics here.

Your storyboard doesn't need to include the minutia of your interface unless it's important to the task and what is novel about your interface (e.g., dont show people clicking on File → Open). Your storyboard should have enough detail (e.g., dont say Bill was unhappy then he used our awesome tool and now he is happy). You will be graded based on the criteria shown in table 1. Here are some guides to understanding storyboards:

`http://www.usabilitybok.org/storyboard`

`http://hci.stanford.edu/courses/cs147/2009/assignments/storyboard_notes.pdf`

c.(group) Evaluate the designs of your colleagues using peer analysis techniques discussed in class. Have team members come up with questions about your storyboard and your product. For your interface pick 5 questions from each team member and write 1-2 sentence answers. As a team decide on a joint design, and repeat the process in a) as a group this time presenting three storyboards for the group.

d.(group) Your R&D department tells you that if you can collect data from the user, it would increase the performance of the system so that it is nearly flawless. You need one minute of brain signal for each state. How would your teams design change? What are the tradeoffs? Add a storyboard and write a couple of paragraphs describing the change, if any.

Criteria	Guiding Questions	Bare Minimum	Satisfactory	Above and Beyond
Task choice (20%)	Do your storyboards clearly communicate a user's real problem or need? Convince us that this problem needs to be solved!	Task is vague, or ill-specified. Storyboards do not demonstrate the need for such a task.	Storyboards communicate an authentic need and the task effectively.	Task is unique and addresses a real need. Storyboards clearly convince reader of the task's authenticity.
Design alternatives (40%)	Do your storyboards communicate multiple significantly different aspects of your product? Do you demonstrate how your idea solves the user's problem or desire? Generate as many as you can and show us!	Little variation among each storyboard of either interface or scenario. Designs do not convincingly accomplish the task at hand.	Storyboards show significant variation in interface or scenario. Designs solve problem to a degree.	Storyboards demonstrate deep thought about multiple design alternatives, Utility of designs is shown clearly & elegantly.
Clarity (40%)	Are your design ideas communicated clearly? Are the important aspects of your interface illustrated? Do your storyboards give a decent understanding of how your interface words? We are not looking for artistry, just good communication!	Storyboards poorly communicate design ideas. Lacks key elements necessary to establish scenario and design solution.	Storyboards communicate design ideas effectively, using a solid mix of illustrations and words to focus on key elements of story.	Illustrates ideas intelligently, focusing on important scenarios and interface elements, Relies less on labels for explanation.

Table 1: Criteria for product design

Q2. This is a simulated social network dataset of 10000 users (think Facebook or Google Plus). There are 10 files in total. You can find them in CourseWorks. These files provide snapshots of the network from Monday (2013-10-01) to Sunday (2013-10-07) in 7 csv files respectively. Each file contains the number of users visiting the social network site, the number of posts, time spent on website, and new friends made on that day (if users didn't visit the site that day, then there is no record in the corresponding file). There is a csv file about users profile. More details can be found in "README.txt". There is also a big csv file (a 10000×10000 matrix) representing all the friend links between the 10000 users. If users i and user j are friends, then the (i,j) entry will be 1, otherwise it's 0. Note that the matrix is symmetric.

(a). Join the 7 csv files on the users actions. Note that most users don't visit the site everyday. You are supposed to calculate the total number of visits, the total number of posts, the total time spent, and the total number of new friends made during this week for each user. You should also extract the total number of friends for each user from the friendship csv file. Add all these statistics into the users' profile dataset. The profile should be finally be a data frame of 10000 rows and 15 columns.

(b). Perform a basic exploratory data analysis to better under understand this social network. Here are some questions you should answer/guidelines for exploration:

- What's the distribution of the number of friends?
- What's the relation between the number of friends and age?
- Does age affect whether users move away from their hometown to live another city?
- Characterize the relation between age and relationship status.
- Draw scatter plots between the number of total actions, visits, posts and time spent.
- How does the sign up date influence the number of new friends they made?
- Is there any relation between age and users' actions?
- Do earlier users have more friends?

(Some reference to help with plotting in R: `http://flowingdata.com/category/tutorials/`). Keep a log of your exploration and turn it in as part of your assignment.

(c). Think of the social network is a graph where each user is a node and a friendship is a edge between nodes. An interesting character of a graph is the average degree. Let G be the graph, N be the number of nodes, and d_i be the number of edges that node i has. The average degree of the graph can be computed as follows:

$$\eta(G) = \frac{1}{N}\sum_{i=1}^{N} d_i$$

For a graph of 10000 you might be able to compute average degree. But for real social networks the number of users (i.e., N) is more like 10^9 instead of 10^4. It is then computationally infeasible to get an exact answer. However, we can sample. Let's consider two sampling methods to estimate average degree. In both methods, we begin with a uniformly random sample without replacement, let's call this subgraph $\tilde{G}$.

Method I: for each sampled user, we observe all the edges directly linked to that user.

Method 2: an edge is observed if and only if both nodes of that edge have been sampled.

Based on the subgraph $\tilde{G}$ we sampled, calculate the estimates using the same formula above. Take the sample size as 3000, repeat this process for 10000 trials. Draw the distributions of the two estimates (using histograms). Also calculate the true average degree. How do the two estimates compare to the true average degree? Which sampling method is better? Why? Can you come up with an easy adjustment to correct the worse one?

(d). Based on your analysis in (b),

1. build a model to predict the number of total actions. Begin with a simple model and build up to make more complex models by adding additional predictors, interactions between predictors and transforming your predictors. Use cross-validation to evaluate the models.

2. segment your users based first on reasonable heuristics and then using k-means. Try varying k=2,3,4,5. What is the optimal k? Why? Find a way to characterize your clusters using summary statistics and/or visualizations.

3. Once you've settled on your user segments, count the number of edges within each cluster of users and between each pair of clusters. Now imagine you didn't have access to the dataset that contains user profile information, is it possible to detect these same clusters of users using only the friendship matrix?

Q3.(individual, no teamwork) Look at the blog post on Data Products in the

Wild: `http://columbiadatascience.com/2013/09/29/data-products-in-the-wild/`. In the comments section, respond to the question at the end of the blog post, or respond to the responses of any of your classmates. Write 1-2 paragraphs.

Q4.(individual, no teamwork) Getting started on the Kaggle competition:

- Go to our inClass Kaggle competition and download the data.
- Submit your first individual entry
- Document your process and thinking.
- Tell us what user name you are using