CS 443 – Introduction to Data Science
Spring 2015 – Homework 1 – 100 points
Due 11:59 PM (Eastern) February 23, 2013

This homework consists of problems covering data types, data visualization, data wrangling / pre-processing, and finding similar items.

A few instructions to make life easier for all of us:

- Turn in your answers in a single file, called <Last-Name><First-Name>HW1.pdf, by uploading it into Sakai.
- Please write concisely and clearly. There are points for intermediate steps, but not in "talking problems to death."
- Have the following at the top of every page: <Last-name>, <First-name> [<page-number>/<number-of-pages>] Example: Smith, John [1/10]
- Your solution should have a cover-page that provides the following information:

<First-name> <Last-name>

Problem Number	Solution Page
1a	
1b	
1c	
…	

If you did not answer a problem, then enter "not done" instead of the solution page.

- It is highly recommended that you turn-in a typed copy of your homework (as opposed to scanned hand-written copy). This is especially true for equations and plots.
- As always, our TA is here to help you with any doubt or confusion that you may have.

Q1. [50 points total] The file "laptops.csv" contains review data about laptops scraped from Amazon on 01/01/2012 and 01/02/2012. The file has a header line plus 24,735 lines each corresponding to information about a review for a laptop. This data file has already been cleaned up. The columns correspond to features of each review. These include:

- `producer`: This is the brand of the laptop (e.g., Apple, DELL, HP).
- `release_to_review_time`: This is the time interval between when the product was released and the first posted review. A negative value indicates a review that was posted before the product was released. The unit of time is not important here.
- `used_real_name`: This feature is 1 if the reviewer used his/her "real" name; otherwise it is 0.
- `verified_purchase`: This feature is 1 if the reviewer purchased the product before reviewing it; otherwise, it is 0.
- `rating`: This is the rating that the product received. It is one of {1, 2, 3, 4, 5}.
- `helpfulness`: This feature is defined as the number of votes marked helpful divided by the total number of votes. It is always between 0 and 1, inclusive. If the total number of votes is 0, then its value is set to 0.
- `number_of_votes`: This is the total number of votes that the product received.
- `length_of_review_text`: This feature reports the number of characters in the review.

1a. [18 points; 3 points per cell] Fill the following table by inserting the type of feature (nominal, ordinal, binary, interval, or ratio) next to each feature. I have already filled two of the rows for you.

Feature Name	Feature Type
`producer`	
`release_to_review_time`	
`used_real_name`	binary
`verified_purchase`	binary
`rating`	
`helpfulness`	
`number_of_votes`	
`length_of_review_text`	

1b. [3 points] Which brand is the *mode* `producer` in this data?

1c. [3 points] Of the reviews whose reviewers used their real names, what percentage had a verified purchase?

1d. [3 points] Of the reviews that had a verified purchase, what percentage had a reviewer who used his/her real name?

1e. [4 points] For `release_to_review_time`, what are its minimum, Q1, median, Q3, maximum, and interquartile range values? What is the appropriate plot to show these numbers?

1f. [6 points] Plot the histogram for `length_of_review_text`. Use bins of size 200.

1g. [6 points] Is the distribution of `length_of_review_text` skewed? Are their outliers in the histogram for `length_of_review_text`? If so, give the data vectors associated with the top 3 outliers.

1h. [4 points] Compute Pearson Correlation between `length_of_review_text` and `helpfulness`. Briefly explain the association between these two variables based on the correlation value you computed.

1i. [3 points] Plot the scatter plot for `length_of_review_text` and `helpfulness`.

Q2. [8 points total] Suppose we have an LSH family h of (d_1, d_2, 0.6, 0.4) hash functions. We can use three functions from h and the AND-construction to form a (d_1, d_2, w, x) family; and we can use two functions from h and the OR-construction to form a (d_1, d_2, y, z) family. Calculate w, x, y, and z. Briefly explain how you computed these values.

Q3. [8 points total] Consider the following three vectors u, v, w in a 6-dimensional space:

$$u = [1, 0.25, 0, 0, 0.5, 0]$$
$$v = [0.75, 0, 0, 0.2, 0.4, 0]$$
$$w = [0, 0.1, 0.75, 0, 0, 1]$$

Suppose we construct 3-bit sketches of the vectors by the random hyperplane method, using the randomly generated normal vectors r_1, r_2, and r_3 in that order:

$$r_1 = [1,-1,1,-1,1,-1]$$
$$r_2 = [-1,-1,1,1,-1,1]$$
$$r_3 = [1,1,1,1,1,1]$$

3a. [6 points] Construct the sketches of the three vectors u, v, and w. Briefly describe how you constructed the sketches.

3b. [2 points] Estimate the pairwise cosine similarity of u with v and w from their 3-bit sketches.

Q4. [6 points total] Recall that the Mahalanobis distance between two data rows is

$$d_{MH}(x,y) = \left((x-y)^T \Sigma^{-1}(x-y)\right)^{\frac{1}{2}}$$

Evaluates to a scalar distance

Vector difference in d-dimensional space

Inverse covariance matrix

4a. [3 points] When does the Mahalanobis distance reduce to the Euclidean distance? Briefly explain your answer either algebraically or in words.

4b. [3 points] When does the Mahalanobis distance reduce to the normalized Euclidean distance? Briefly explain your answer either algebraically or in words.

Q5. [15 points total] Suppose you are given the following element-by-document matrix:[1]

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

5a. [6 points] Compute the minhash signature for each column with the following three hash functions:

$h_1(x) = 2x + 1 \bmod 6$
$h_2(x) = 3x + 2 \bmod 6$
$h_3(x) = 5x + 2 \bmod 6$

5b. [3 points] Which of these hash functions are true permutations?

5c. [6 points] How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

Q6. [8 points total] You are given the following two sentences:

Sentence A:[2] "Mr. and Mrs. Dursley of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much."

[1] From MMDS Chapter 3, Exercise 3.3.3.
[2] First sentence of *Harry Potter and the Sorcerer's Stone* by JK Rowling.

Sentence B:[3] "Not for the first time, an argument had broken out over breakfast at number four, Privet Drive."

Use these stop-words {and, of, to, that, for, the, an, at} to find the stop-word based shingles of the above sentences. Here a shingle is defined as a stop word followed by the next two words. Compute the Jaccard Similarity of the two sentences based on these shingles.

Q7. [5 points total] Given a variable p in [0,1], the *logit* function is defined as follows:

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

When would you use the logit function to transform data?

End of homework.

[3] First sentence of *Harry Potter and the Chamber of Secrets* by JK Rowling.