FALL 2019: CAP 5768 – Intro to Data Science
[Exam Review]

Problems

1. (Lec 2) How is a Data Frame different from a two-dimensional array?

2. (Lec 4) Explain how the following Python code is equivalent to a Database join:

```python
unames = ['user_id', 'gender', 'age', 'occupation', 'zip']
users = pd.read_table('users.dat', sep='::', header=None, names=unames, engine='python')
rnames = ['user_id', 'movie_id', 'rating', 'timestamp']
ratings = pd.read_table('ratings.dat', sep='::', header=None, names=rnames, engine='python')
pd.merge(movies, ratings, on="movie_id")
```

3. Make sure you understand in what context we used the following discrete distributions – uniform, binomial, negative binomial, geometric and poisson, or their corresponding continuous distributions.

4. What does the law of large numbers say about the relationship between the sample mean and the population mean?

5. Explain a clustered bar chart, stacked bar chart and bar chart with whiskers.

6. What is a histogram and a violin plot?

7. What is a pie chart?

8. What is linear regression and Pearson Correlation Coefficient? When are two variables said to be positively correlated?

9. What is the difference between a t-test and a paired t-test?

10. What is a one-sided error?

11. What is a mode and a bimodal distribution?

12. What do the acronyms TF and IDF stand for?

13. (Lec 7) Explain in some detail how matrix-vector multiplication is handled using MapReduce.

14. (Lec 9) Under what conditions would you have a memory problem when running the APRIORI algorithm for computing frequent itemsets?
15. Explain the principle of monotonicity exploited in the APRIORI algorithm.

16. Differentiate between support and confidence in the APRIORI algorithm.

17. (Lec 10) Explain the relationship between MinHash and Jaccard similarity.

18. (Lec 10-11) What properties must a distance function satisfy? Define one well-known distance function other than the Euclidean distance function.