FALL 2018: 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*:

```
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 disributions.
4. What does the *law of large numbers* say about the relationship between the sample mean and the population mean?
5. What do the acronyms *TF* and *IDF* stand for?
6. (Lec 7) Explain in some detail how matrix-vector multiplication is handled using MapReduce.
7. (Lec 9) Under what conditions would you have a memory problem when running the Apriori algorithm for computing *frequent itemsets*?
8. Explain the *principle of monotonicity* exploited in the Apriori algorithm.
9. Differentiate between *support* and *confidence* in the Apriori algorithm.
10. (Lec 10) Explain the relationship between MinHash and Jaccard similarity.
11. (Lec 11) Explain how to use *Bloom Filters*.
12. Write down pseudocode for applying *Bloom Filters* for set membership.
13. (Lec 13) What properties must a distance function satisfy? Define one well-known distance function other than the Euclidean distance function.
14. (Lec 14) State one consequence of the *curse of dimensionality* and explain it.

15. (Lec 13-14) Which is top-down, hierarchical clustering or K-means? Why?

16. (Lec 19) Explain the connection between *PageRank* and *random walks*.

17. (Lec 23) Explain the concept of *moving averages* and how it helps to reduce the mean square error.