Janine eiser
Midterm Corrections

3. positive skew
source:http://jretz.github.io/datamining290/slides/2014-02-06-Probability.htm#5
4.negative skew
source: http://jretz.github.io/datamining290/slides/2014-02-06-Probability.html#6

5.Extract transform load
source:http://en.wikipedia.org/wiki/Extract,_transform,_load

6.Draw the star schema

ad type tbl
adtypeid
ad_type
size

bid per impression tbl
bpimid
bidperimpression

fact tbl
factid
adtypeid
bpimid
viewid
pgaid
display_timestamp

number of views tbl
viewid
ad_id
num_of_views

pages tbl
pgaid
page
pageid

Source:
http://jretz.github.io/datamining290/slides/2014-02-13-Data-Warehouse.html#28

7.1Rollup
A roll up summarize data along fewer dimensions. Example: What types of ads are being are displayed most often?
Source:http://jretz.github.io/datamining290/slides/2014-02-13-Data-Warehouse.html#36

7.3 Slice and Dice: I got slice wrong. Example fo slice would be: We want to find the number of ad views and the bid price paid for ads shown in the past month so that we can find the total ad revenue for the month.
Source: http://jretz.github.io/datamining290/slides/2014-02-13-Data-Warehouse.html#36

9. Describe recall and precision:
Recall means you haven't missed anything in your classification but you may have a lot of useless results to sift through.
Precisionmeans that everything returned was a relevant result, but you may not have found all the relevant items
Source: http://en.wikipedia.org/wiki/Precision_and_recall

11. For this question, we should change the learning rate. The learning rate of the gradient is too high, meaning the steps we are taking are too big. As a result, the program just runs forever. To fix this, we

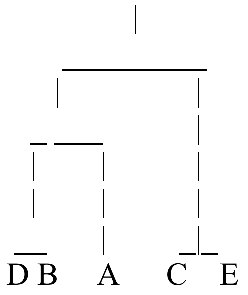
should lower the learning rate and make the gradient takes smaller steps. With smaller steps, we can eventually converge on a solution.
Source: http://jretz.github.io/datamining290/slides/2014-02-27-SVM.html#16

13. The part of this questions I got wrong was, “how do we know when the know when the weights are right?”

We know when the weights are right when: As part of the nueral network trainer, we create a fitness function that measures the error of the weights of the connection. Then, take take the derivative the of the sigmoid → $O_j(1 – O_j)$ aka, taking the gradient, and take a step in the right direction at whatever the learning rate is set to. Also, we need to adjust our weights based on the amount of incorrectness in the system, and try again. The weights are right when we have values error of weights is less than the value of the error that we set as an acceptable amount of error in our fitness trainer.

Source: http://jretz.github.io/datamining290/slides/2014-02-27-Neural-Network.html#29

14.



Source: http://jretz.github.io/datamining290/slides/2014-03-06-Hierarchical.html#8

15.Corrected map reduce job:
mapper: get reviews: none: record ---> category, (#number of reviews, star rank)
reducer: get_biz_category_avg:
category, (#number of reviews, star rank) ---> category, sum(star rank) / category, sum(#number of views)

16. answer: intesection/union = 2/7. I mistakenly counted the midterm twice.