

Data-Mining 290 Midterm

86/90

On my honor as a student, I have neither given nor received aid on this exam.

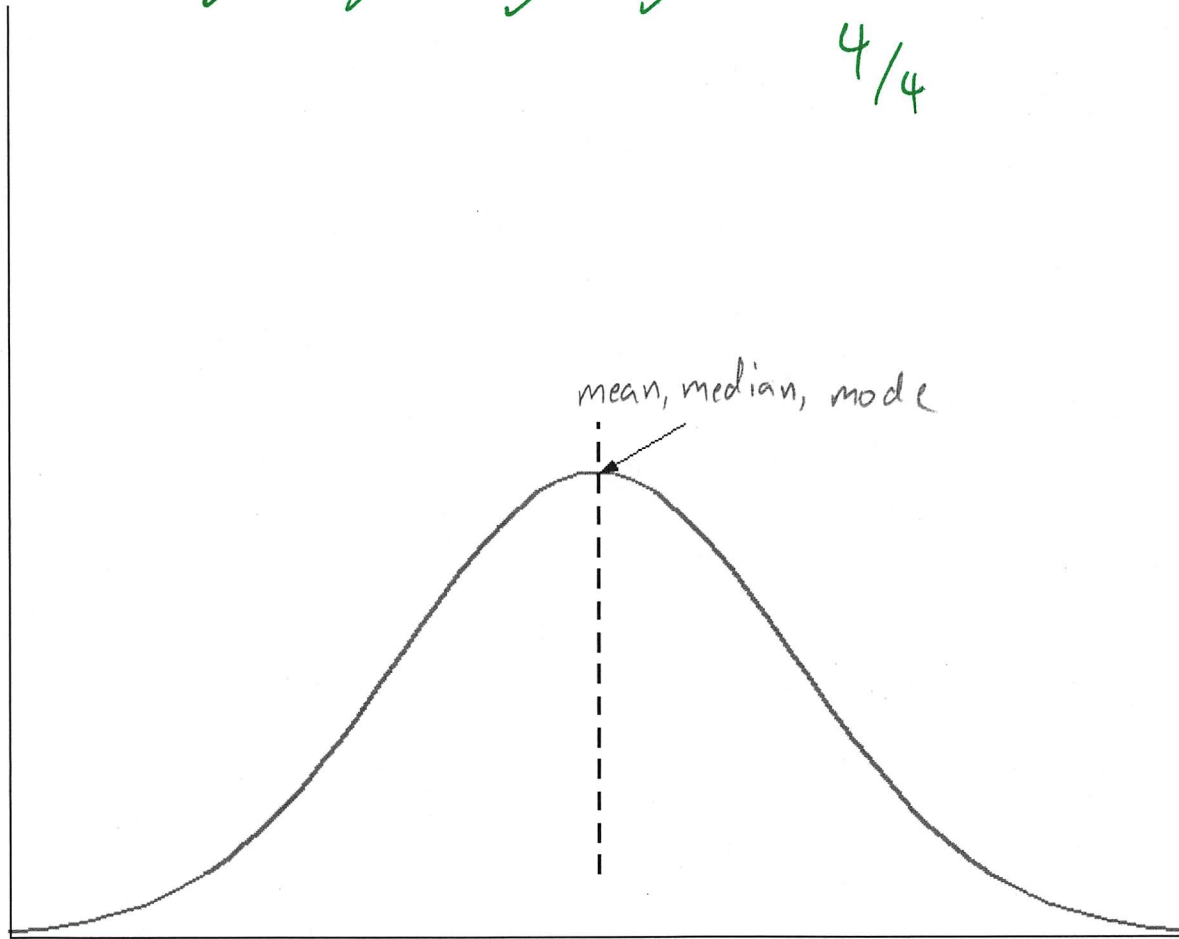
Name: Cory Schillaci

Signature: *Cory Schillaci*

1 Label mean, median, mode, skew

✓ ✓ ✓ ✓

4/4

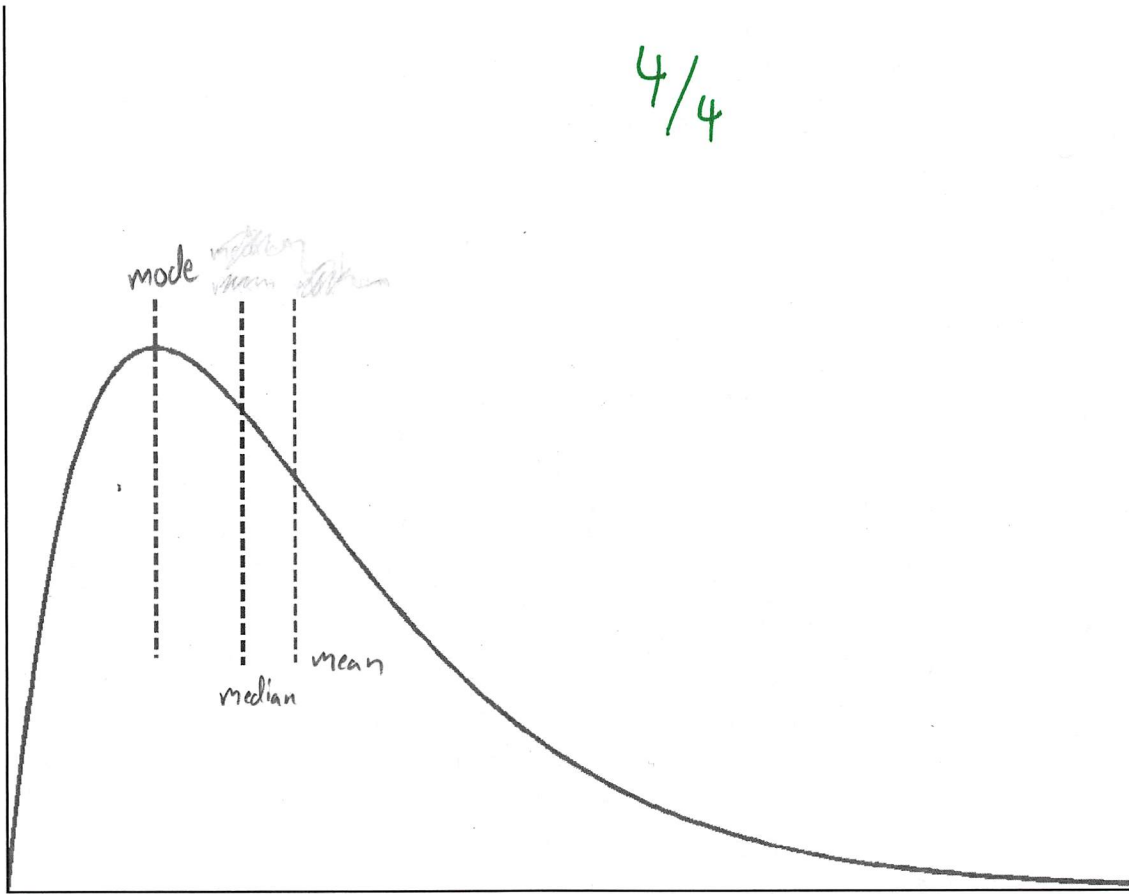


Zero skew

2 Label mean, median, mode, skew



4/4



positive skew (mean > mode)

3 Label mean, median, mode, skew

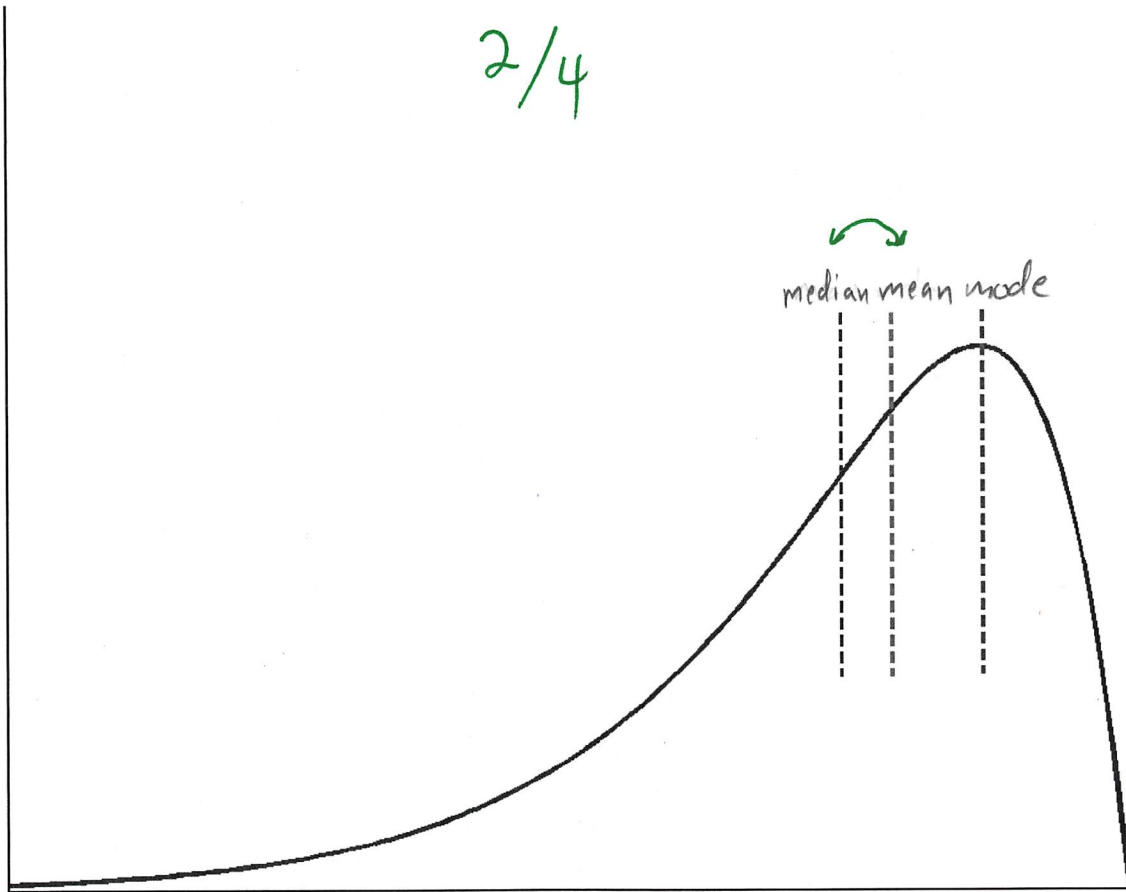
x

x

✓

✓

2/4



negative skew (mean < mode)

4 Normalize the following feature values

Use min-max [0-1] normalization, output fractions or 4 decimal places:

[5, 25, 6, 12, 15]

min = 5
max = 25

$$x_{[0,1]} = \frac{x - \min}{\underbrace{\max - \min}_{20}} = [0, 1, 1/20, \overset{x}{3/5}, \overset{x}{3/4}]$$

$5/5$

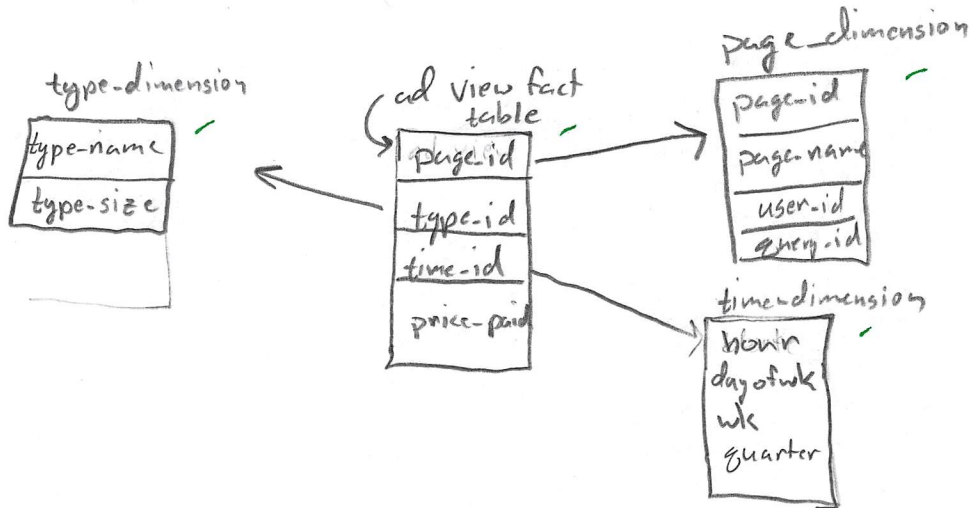
5 What does ETL stand for?

Extract, Transform, Load

$6/6$

6 Draw a star schema for the following information needs:

- The fact table tracks online advertising views
- Advertisements can be different types. Types have names (eg. banner, text, search) and sizes (eg. 728x90, 478x60).
- Advertisements can appear on different pages. Pages have names (eg. home, search, or login) and IDs (eg. the user_{id}, query ID).
- The highest time resolution we need is 1 hour, but we want the ability to summarize by week and quarter
- Every advertisement view is also associated with bid price paid for the impression.



7 Using the advertising schema above, give concrete examples for the following operations

7.1 Rollup

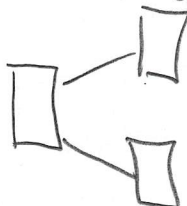
~~Summarize price of all ads with type "banner" and~~
~~(size "478x60")~~

2/2

Look at ads of any type, and size.

by

size.



7.2 Drill-down ✓

Start by looking at data from quarter 1, 2011
then look at data from different weeks
in that quarter.

2/2

7.3 Slice & Dice ✓

Look only at ads with type banner (slice),
or only ads with type banner or type text
(dice).

2/2

7.4 Pivot ^f

I don't really know this one

0/2

8 Put the following steps for building a classifier in the correct chronological order

1. prepare training data
2. teach classifier
3. output guesses on new data
4. input unlabeled new data into model
5. set aside hold-out set
6. save model parameters
7. verification with testing data

6/6

Correct ordering: 1, 5, 2, 7, 6, 4, 3

9 Describe precision and recall in English

4/4

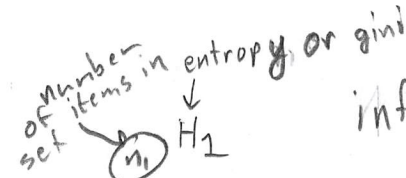
precision = $\frac{TP}{TP+FP}$ = Fraction of positive identifications which are correct

recall = $\frac{TP}{P}$ = Fraction of actual positives which are correctly identified.

10 What is information gain and how is it used to build decision trees?

★ Information gain is the decrease in an impurity measure (Gini index, entropy, etc) after making a split.

5/5



$$\text{info gain} = H_1 - \frac{n_2 H_2 + n_3 H_3}{n_2 + n_3}$$

★ At each step, we make the split which maximizes info gain

11 Fixing gradient descent

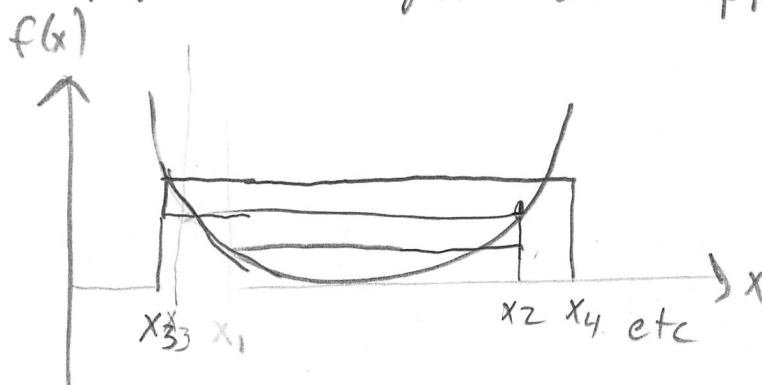
We are training a model using linear regression. However, instead of converging on a solution, our program seems to keep running forever without output. We add a print statement to our program that outputs the RMSE found at each iteration and the result is:

2.25280
4.82905
10.9052
22.9302
51.8141
117.8202

What is going on here? Assuming we don't have any bugs in our implementation, what parameter can we change to fix this problem?

Most likely the optimization algorithm is failing. If using gradient descent, we need to reduce the learning rate. Linear regression is a convex

Problem so the optimum should be unique. Roughly what might be happening?



5/5

12 When implementing a neural network trainer, what are our terminating conditions?

What are (at least) three ways we can choose to stop iterating?

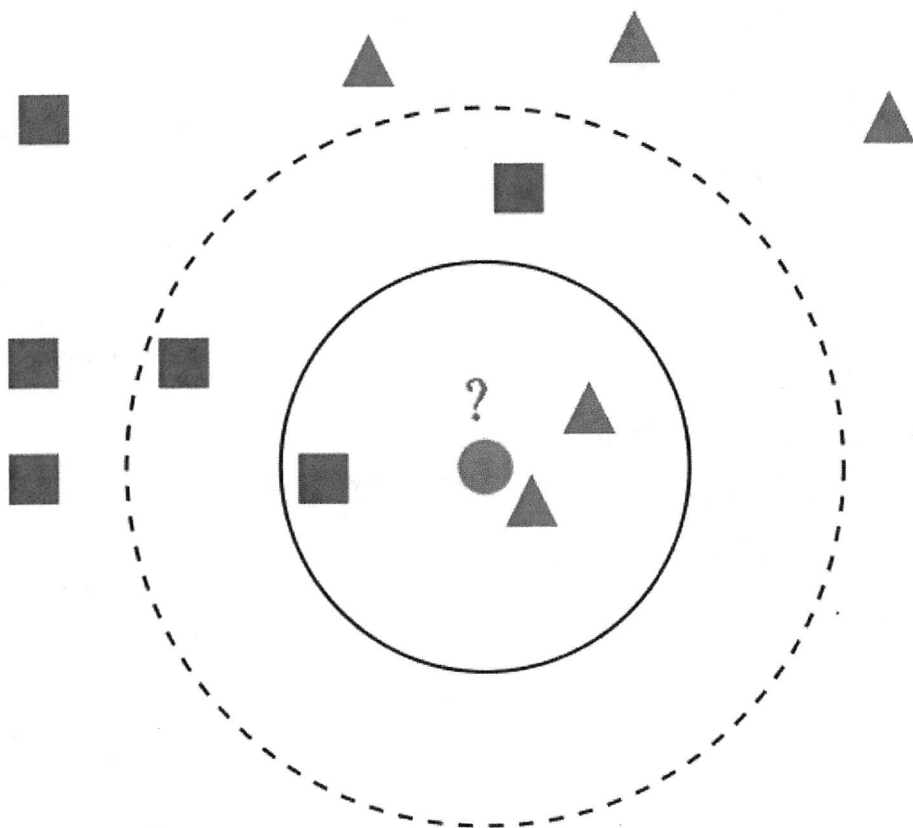
- Desired accuracy reached
- Max # of iterations
- Updates to weights smaller than some threshold



• Max wall-clock training time

6/6

13 Square or Triangle?

Label the unknown circle point using k-nearest neighbor



- Where $k=1$: 
- Where $k=5$: 

4/4

14 Draw a dendrogram showing clustering for these points

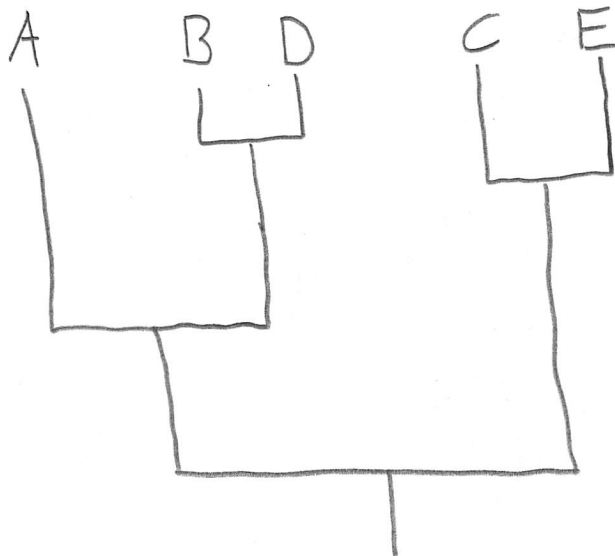
Points are agglomeratively clustered using L_2 norm, minimum cluster distance.

Ⓐ

Ⓑ Ⓓ

Ⓒ

Ⓔ



5/5

15 Write MapReduce inputs/outputs

We'd like to use the MapReduce paradigm to answer the question: What is the average **review** star rating for a business category? You're using the business record type, which has

- average number of stars
- list of categories
- number of reviews

Write the key/value signatures for the MapReduce steps you'd write (you may optionally explain what each step does in English or pseudo-code). Key/Value signatures are:

map|reduce step_name : type_of_key, type_of_value => type_of_key, type_of_value

For example, in a word count job, part of the solution might be:

map get_words : None, record => review_id, number of words

map get-categories: None, record => category, [average star rating]
reduce calc-average: category, review_id => category, n_reviews = sum

map get-categories: None, record => category, stars
Note that this should yield ^{the value} n_reviews times for each record
reduce calc-average: category, stars => category, mean(stars)

example data cat1

Turn page over, I didn't read carefully

map get_categories None, record:

for cat in category-list:

for i in range(0, n-reviews):

yield category, ave-stars

reduce ave-stars category, stars

yield category, mean(stars)

yield 'category, star
pair the # times
rest was reviewed
for each category
it is in

10/10



average the stars

16 What is the Jaccard index between these two sets:

- {good luck on the midterm} $\equiv A$
- {the midterm is over} $\equiv B$

$$\text{size}(A \cup B) = \text{size}(\text{good, luck, on, the, midterm, is, over})$$
$$= 7$$

$$A \cap B = \{\text{the, midterm}\} \quad (\text{size} = 2)$$

$$\text{Jaccard index} = 2/7$$

5/5