CSE 190, Fall 2015: Midterm

Name:	Student ID:

Instructions

The test will start at 5:10pm. Hand in your solution at or before 6:10pm. Answers should be written directly in the spaces provided.

Do not open or start the test before instructed to do so.

Note that the final page contains some algorithms and definitions. Total marks = 25

Section 1: Regression

Feature design

Suppose we collected the following data about businesses from *Google Local*:

ID	Name	Av. Rating	Price	Address	latitude/longitude	hours
0	T C's Referee Sports Bar	5.0	$$	Sioux Falls, SD 57106	43.529, -96.792	m-f/11am-10pm, s-s/11am-1am
1	Old Chicago	3.0	$$	Beaverton, OR 97006	45.535, -122.862	m-f/11am-1am
2	Sabatinos Italian Kitchen	4.0	$$$	Arlington, MA 02474	42.406, -71.143	m-f/10am-10pm, s-s/10am-9pm
3	Oakville Grocery	4.5	$	Healdsburg, CA 95448	25.063, 121.524	mon-sun/9am-5pm
4	Hog Wild Pit BBQ	3.5	$$	Wichita, KS 67213	37.681, -97.389	mon-sun/11am-8pm

1. What would be the mean squared error (MSE) of a trivial model that simply predicted the average rating for all items (1 mark)? A:

2. Suppose we wanted to train a personalized model that predicted the rating I would give to a business based on the population-level average and price, i.e.,

$$\text{my rating} = \theta_0 + \theta_1[\text{average rating}] + \theta_2[\text{price}].$$

   Write down the complete feature matrix (in the space below) that you would use to solve the above equation (1 mark):

$$y = \begin{bmatrix} & \\ & \\ & \\ & \\ & \end{bmatrix} \theta$$

3. Write down the predictions that would be obtained for the five businesses if using the features above if the parameters were $\theta = [0.1, 1.0, -0.2]^T$ (1 mark)

ID	Name	Predicted Rating	Q4 answer	Q5 answer
0	T C's Referee Sports Bar			
1	Old Chicago			
2	Sabatinos Italian Kitchen			
3	Oakville Grocery			
4	Hog Wild Pit BBQ			

4. The opening hours might also influence my preferences. How would you construct useful features for the above businesses, if I have a preference toward *businesses that are open late*? Using your representation write down (in the table above) the features corresponding to each business (1 mark):

A:

5. How would you incorporate opening-hour features if my preferences are toward *businesses that are open outside of work hours (i.e., mon-fri/9am-5pm)*? Write down the features corresponding to each business (1 mark):

A:

6. Finally, suppose I want to model how people's preferences change as a function of geography (based on 1,000 U.S. businesses including the five above), i.e.,

$$\text{average rating} = x(\text{geographical features}) \cdot \theta$$

How might you use the features available above (e.g. address or latitude/longitude) to model such geographical trends (1 mark)? (describe your solution, rather than writing down the actual features)

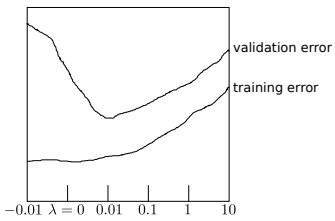
A:

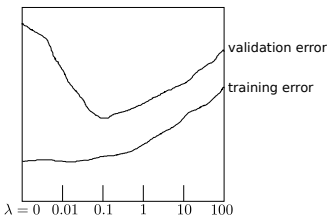
Diagnostics

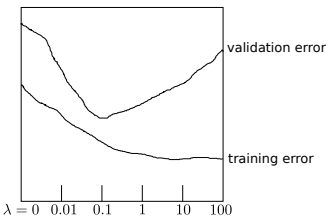
7. Suppose we trained our model above by minimizing the *regularized mean squared error*, i.e.,

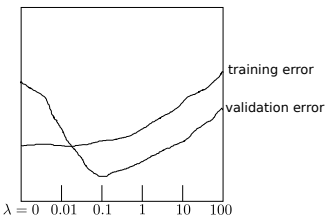
$$\operatorname*{argmin}_{\theta} \|y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

Suppose that we split our data into training, validation, and test sets (and that we do so randomly, given plenty of data). Which of the plots below could correspond to the performance (i.e., MSE) on the training and validation sets? For each that could *not*, briefly explain why below (2 marks).









Section 2: Classification

The following is a list of Vin Diesel's films:

No.	Title	Year	IMDB rating	MPAA rating	length (minutes)
1	The Last Witch Hunter	2015	6.3	PG-13	106
2	Furious 7	2015	7.4	PG-13	137
3	Guardians of the Galaxy	2014	8.1	PG-13	121
4	Riddick	2013	6.4	R	119
5	Fast & Furious 6	2013	7.2	PG-13	130
6	Fast Five	2011	7.3	PG-13	131
7	Fast & Furious	2009	6.6	PG-13	107
8	The Fast and the Furious: Tokyo Drift	2006	6.0	PG-13	104
9	The Pacifier	2005	5.5	PG	95
10	The Chronicles of Riddick	2004	6.7	PG-13	119
11	xXx	2002	5.8	PG-13	124
12	The Fast and the Furious	2001	6.7	PG-13	106
13	Pitch Black	2000	7.1	R	109
14	The Iron Giant	1999	8.0	PG	86
15	Saving Private Ryan	1998	8.6	R	169

You hear a rumor that Vin Diesel has a new film coming out that is (A) Over two hours long (B) Rated PG-13 (C) Has the word "Furious" in the title. Let's try to estimate the probability that it will (D) have an IMDB rating of 7.0 or above.

8. Based on the data above (and not making any other assumptions) write down the probability

$$p(D|A \wedge B \wedge C)$$

(1 mark) A:

9. The above probability may be unreliable as it is based on very few observations that exhibit the required features. So, we'll try to decide whether D is likely to be true or not following the Naïve Bayes assumption. Write down all of the terms involved and finally the probability ratio, and the conclusion you draw as a result (3 marks).

A:

Evaluation measures

Suppose we are performing a ranking task to try and identify pages that are relevant to some particular search query, and that we achieve this by building a logistic regressor that outputs a score indicating the probability that a page is relevant. Suppose the scores we obtain are the following:

page id	score	actually relevant?
0	0.78	yes
1	0.25	no
2	0.36	yes
3	0.18	no
4	0.01	no
5	0.95	yes
6	0.92	yes
7	0.11	no
8	0.20	no
9	0.56	no

10. Write down the number of true positives, true negatives, false positives, and false negatives of our logistic classifier (2 marks).

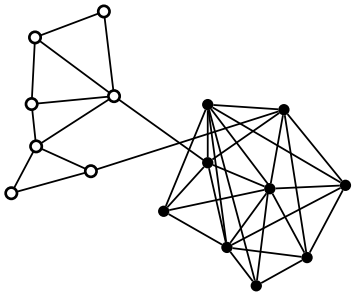
TP:	TN:	FP:	FN:

11. Complete the table below by ranking pages in decreasing order of confidence (3 marks).

page id	confidence	actually relevant?	precision@k	recall@k
5	0.95	yes		

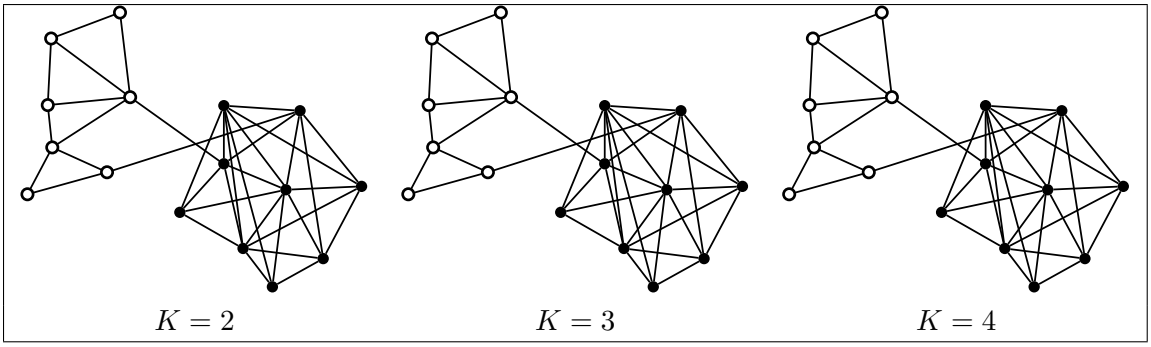
Section 3: Communities & clustering

12. Suppose a social network is divided into the two communities shown below (filled vs. unfilled nodes). If we wanted an algorithm to find these communities automatically, which of *ratio cuts* versus *normalized cuts* would be more appropriate and why (1 mark)?



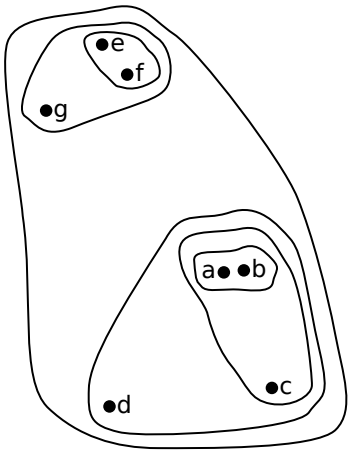
A:

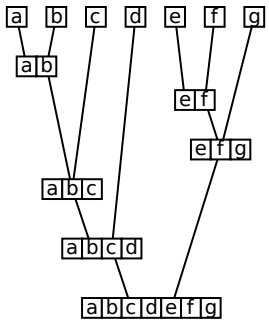
13. What would be the result of running *clique percolation* on the graphs below (3 marks)? Circle the communities that would be found directly on the graphs.



14. Suppose you ran *hierarchical clustering* on the points below, resulting in the dendrogram shown in the center. How would you use the output of this algorithm (i.e., the clusters/dendrogram) to generate useful feature representations for the original points? Write your features for the 7 points below (1 mark).

A:





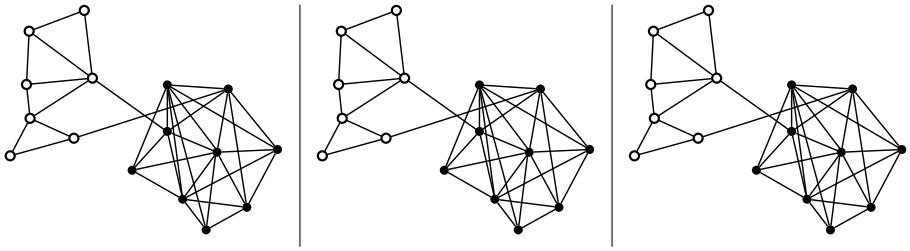
$X =$

Algorithm design

15. Suppose you wanted to design a system to estimate what tip a prospective fare would give for a taxi ride in San Diego. Describe below what data and features you would collect to estimate this value, and what techniques you would use to solve the task (3 marks).

A:

Here are a few more graphs in case you need to re-write your clique-percolation solutions:



Write any additional answers/corrections/comments here:

Algorithm 1 Ratio cut

Choose communities $c \in C$ that minimize $\frac{1}{2} \sum_{c \in C} \frac{\overbrace{cut(c, \bar{c})}^{\text{edges in cut}}}{\underbrace{|c|}_{\text{size of community}}}$

Algorithm 2 Normalized cut

Choose communities $c \in C$ that minimize $\frac{1}{2} \sum_{c \in C} \frac{\overbrace{cut(c, \bar{c})}^{\text{edges in cut}}}{\underbrace{\sum \text{degrees in } c}_{\text{sum of node degrees in community}}}$

Algorithm 3 Clique percolation with parameter k

Initially, all k-cliques in the graph are communities
while there are two communities that have a $(k-1)$-clique in common **do**
  merge both communities into a single community

Algorithm 4 Hierarchical clustering

Initially, every point is assigned to its own cluster
while there is more than one cluster **do**
  Compute the center of each cluster
  Combine the two clusters with the nearest centers

Naïve Bayes:

$$p(\text{label}|\text{features}) \simeq \frac{p(\text{label}) \prod_i p(\text{feature}_i|\text{label})}{p(\text{features})}$$

Precision:

$$\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall:

$$\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$