CSE 190, Spring 2015: Midterm

Name:	Student ID:

Instructions

Hand in your solution at or before 7:45pm. Answers should be written directly in the spaces provided. **Do not open or start the test before instructed to do so.**

Section 1: Regression

Q1: Restaurants & ratings (10 marks)

Suppose we collected the following data about restaurants from *Yelp!*:

Name	Average Rating	Takes reservations?	Take-out?	Price	Good for
Oceana Coastal Kitchen	4.5	Yes	No	$$$	Breakfast
Beyer Deli	5.0	No	Yes	$	Lunch
Werewolf	4.5	Yes	Yes	$$	Brunch
C Level	4.0	No	Yes	$$	Lunch, Dinner
Cucina Urban	4.5	Yes	Yes	$$	Dinner

and that from this data we want to estimate

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] + \theta_3[\text{price}]$$

1. What is the average rating across all restaurants (1 mark)? A:

2. What is the Mean Squared Error of the a predictor that just predicts the average rating for all items (1 mark)? A:

3. Suppose we'd like to write down the above expression for the rating in the form $y \simeq X\theta$. Complete the following equation to do so:

$$\begin{bmatrix} 4.5 \\ \\ \\ \\ \\ \end{bmatrix} \simeq \begin{bmatrix} 1 & 1 & 0 & 3 \\ & & & \\ & & & \\ & & & \\ & & & \end{bmatrix} \theta$$

   (1 mark)

4. In the expression $y \simeq X\theta$, which term encodes the labels, which term encodes the features, and which term encodes the parameters (1 mark)? labels: features: parameters:

5. Suppose that after fitting our model for the rating we obtain $\theta = [7, 0.5, -1, -1]^T$. What is the interpretation of $\theta_0 = 7$ in this expression (1 mark)?

   A:

6. What is the interpretation of $\theta_3 = -1$ (1 mark)?

   A:

7. Write down the predictions made by the model when $\theta = [7, 0.5, -1, -1]^T$:

$$\text{predictions} = \begin{bmatrix} 4.5 \\ \\ \\ \\ \\ \end{bmatrix}$$

   (1 mark)

8. What is the Mean Squared Error of the predictions you computed above (1 mark)? A:

9. Suppose you wanted to incorporate the 'Good for' field (the last column of the above table) into your model. How would you represent the features in order to do so? Answer this by writing down the model you would use:

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] +$$

$$\theta_3[\text{price}] +$$ A:

and by completing the feature matrix using your representation:

$$X = \left[\begin{array}{cccc} 1 & 1 & 0 & 3 \\ & & & \\ & & & \\ & & & \end{array}\right.$$

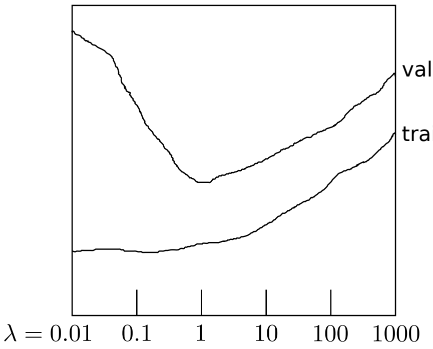
(2 marks)

Q2: Training, testing, & model selection (6 marks)

Suppose we are training regressors to minimize the regularized Mean Squared Error

$$\sum_{(x,y)\in train} \frac{1}{|train|}(y - x \cdot \theta)^2 + \lambda\|\theta\|_2^2.$$

10. Suppose that we fit some model for $\lambda \in \{0.01, 0.1, 1, 10, 100, 1000\}$ and obtain the following performance on the training and validation sets:



Which value of λ would you select based on the results above (1 mark)? $\lambda =$

11. Answer the following questions about training, validation, and test sets:

    (a) What is the role of a validation set (1 mark)?

    A:

    (b) How does the training error typically vary with λ (1 mark)?

    A:

    (c) What is meant by under/over fitting? Which values of λ in the above figure correspond to maximum over/under fitting (1 mark)?

    A:

12. Further suppose that we consider two different feature representations (model X and model Y), and two different regularization parameters ($\lambda = 1$ and $\lambda = 10$) and obtain the following results on the training and validation sets:

model	training error	validation error
model X, $\lambda = 1$	23.34	?
model X, $\lambda = 10$	?	?
model Y, $\lambda = 1$	?	18.32
model Y, $\lambda = 10$	25.98	?

('?' indicates an unknown value).

Assuming that our training/validation/test sets are large, independent samples, is the above information enough to determine which model and which value of λ we would expect to yield the best performance *on the test set*? If so, which model and which value of λ would you expect to perform best and why? Explain your answer (2 marks).

A:

Section 2: Classification

Q3: Fantasy novels (6 marks)

Suppose we have a database consisting of the following books:

Title	Genre	Prediction
The Circle of Sorcerers	Fantasy	True
Knights: The Eye of Divinity	Fantasy	
Superman/Batman: Sorcerer Kings	Graphic Novel	
In the Blood	Mystery	
Remains of the Day	Literature & Fiction	
Blood Song	Fantasy	
Flame Moon	Fantasy	
The Book of The Sword: A History of Daggers	History	
A Storm of Swords	Fantasy	
The Storm Book	Children's	

Further, suppose we are given the following classifier to classify Fantasy vs. non-Fantasy books:

```
if (Title contains 'Sorcerer' or 'Blood' or 'Knights' or 'Moon' or 'Storm'):
  return True
else:
  return False
```

13. What are the predictions made by this classifier? Write your answers in the last column of the table above (1 mark).

14. Of these predictions, what is the number of true positives, true negatives, false positives, and false negatives (1 mark)?

	true positive	true negative	false positive	false negative
A:				

15. What are the true positive rate (hint: TP / (TP + FN)), true negative rate, and balanced error rate (1 mark)?

	true positive rate	true negative rate	balanced error rate
A:			

16. In class we saw three approaches to classification: naïve Bayes, logistic regression, and support-vector machines. Describe one benefit of each approach compared to the other two (3 marks).

naïve Bayes:

logistic regression:

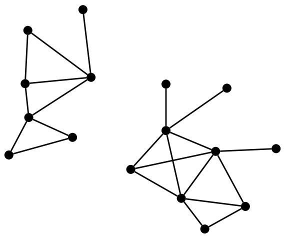
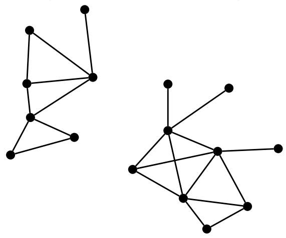
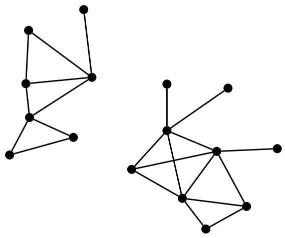
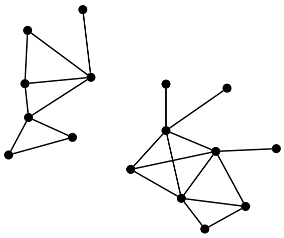
SVM:

Section 3: Communities & clustering

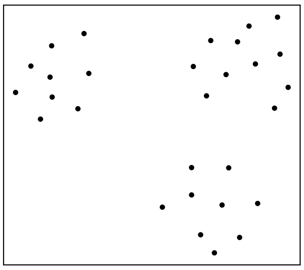
Q4: Algorithms for community detection, dimensionality reduction, and clustering

Recall three algorithms we saw in class to detect communities: connected components, ratio cut, and clique percolation (pseudocode is given as Algorithms 1, 2, and 3 at the end of the test).

17. Identify the communities that would be produced on the graphs below using these three algorithms. Circle the communities directly in the space below (some more graphs are provided on the final page in case you need to re-write your answer):

Connected components	Ratio cut (2 communities)	Clique percolation ($k = 3$)	Clique percolation ($k = 2$)
			
(1 mark)	(1 mark)	(1 mark)	(1 mark)

18. Suppose we are given the following 2-dimensional data X, and wish to cluster it so as to minimize the reconstruction error ($\sum_{x \in X} \|\bar{x} - x\|_2^2$). Separate the points into three clusters such that the reconstruction error (when replacing each point by its cluster centroid) would be minimized. Draw the clusters directly in the space below (1 mark):

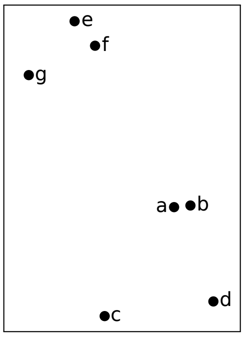


19. By replacing each point with one of the three centroids above, we have effectively 'compressed' the data, since each (2-d) point is replaced by a (1-d) integer. Another way to compress the data would be to perform Principal Component Analysis, and discard the lowest variance dimension, which would also result in a 1-d representation of the data. Out of these two possible compressed representations, which one would result in the lower reconstruction error on the above data, and why (1 mark)?

A:

20. In class we saw *hierarchical clustering*, an algorithm that works by successively joining clusters whose centroids are nearest. Psuedocode is given in Algorithm 4 over the page.

Suppose you are given the following set of points:



Step	Clusters merged	List of clusters
0	(initialization)	$\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}$
1	$\{a\}$ merges with $\{b\}$	$\{a, b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}$
2		
3		
4		
5		
6		$\{a, b, c, d, e, f, g\}$

If we were to perform hierarchical clustering on this data, in what order would the clusters be joined? Answer this question by completing the table above (2 marks).

Algorithm 1 Connected components

```
Two nodes $a$ and $b$ should be assigned to the same community if and only if $a$ is reachable from $b$, and $b$ is reachable from $a$
```

Algorithm 2 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 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
```

Write any additional answers/corrections/comments here:

Here are a few more graphs in case you need to re-write your solutions to Q17: