Name:
Email:
Signature:

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
Total	

Lehman College, CUNY
MAT 456-01: Topics Course: Data Science
SAMPLE FINAL EXAM
Spring 2016

1. What will the following code draw:

```
import numpy as np
import matplotlib.pyplot as plt

def f(t):
    return np.cos(2*np.pi*t)/(t+1)

t1 = np.arange(0.0, 5.0, 0.1)
t2 = np.arange(0.0, 5.0, 0.02)

plt.figure(1)
plt.subplot(211)
plt.plot(t1, f(t1), ’bo’, t2, f(t2), ’k’)

plt.subplot(212)
plt.plot(t2, np.cos(2*np.pi*t2), ’r--’)
plt.show()
```

Output:

2. For each of the regular expressions, give a string that will matches it:

   (a) `(\d\w){2}\d`

   (b) `[bch][ai]+[tT]+`

   (c) `((Roosevelt)|(Long)|(Ellis)) Island`

   (d) `\d{3}-\d{2}-\d{4}`

   (e) `\w+@\w+\.\w+`

3. The New York City Open Data project contains all motor vehicle collisions reported to the New York Police Department. The data can be downloaded as CSV files with the following format:

```
DATE,TIME,BOROUGH,ZIP CODE,LATITUDE,LONGITUDE,LOCATION,ON STREET NAME,CROSS STREET NAME,OFF STREET
02/01/2016,0:09,BRONX,10465,40.8341548,-73.8174815,"(40.8341548, -73.8174815)",BARKLEY AVENUE,DEAN
```

   All lines are formatted similarly: they start with the date, then time, the borough, zip code, latitude and longitude, and also include cross streets, types of vehicles involved, number of injuries/fatalities, and possible cause. The first line of the file gives the entries in the order they occur in the rows.

   Write a program that takes a file, `bronxCollisions.csv`, and prints out all the zip codes for crashes occur in the Bronx:

4. The Center for Disease Control (CDC) provides data on the number of occurrences of Lyme Disease. Assuming you have the data stored:

   ```
   years = [2003,2004,2005,2006,2007,2008,2009,2010,2011]
   ny = [5399,5100,5565,4460,4165,5741,4134,2385,3118]
   nj = [2887,2698,3363,2432,3134,3214,4598,3320,3398]
   ct = [1403,1348,1810,1788,3058,2738,2751,1964,2004]
   ```

   Write a program that will plot the state numbers as well as the total Lyme Disease occurrence (i.e. the sum of the values for the three states for each year).

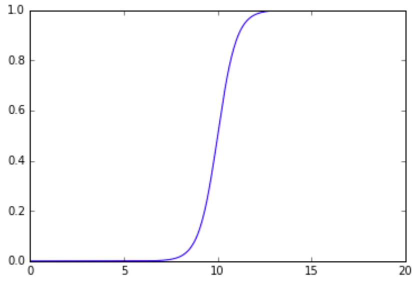
5. A popular business analytics company provides this example of A/B test:

   - A: 300 visitors, 15 conversions, and
   - B: 400 visitors, 20 conversions.

   What is the A/B test statistic for Option A and Option B (i.e. the difference of the estimated means divided by the square root of the sum of the variances)?

6. You culled 10,000 311 records from the NYC Open Data and measured how many days it took to resolve the complaint lodged by the caller. Using this training data, you fit the following logistic function:

$$f(x) = \frac{1}{1 + e^{20-2*x}}$$



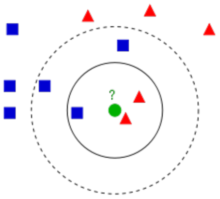
(a) With what probability would you expect that a call recorded 5 days ago is resolved? Explain your answer.

(b) With what probability would you expect that a call recorded 10 days ago is resolved? Explain your answer.

(c) With what probability would you expect that a call recorded 15 days ago is resolved? Explain your answer.

(d) Say you sampled data from a single month and found that average time to resolve a complaint was 2.73 days with a standard deviation of 0.25 day. Does that fit with your prediction function above? Why or why not?

7. (a) In this image from wiki (by Antti Ajanki), all points are being classified as triangles or squares. Using the k-Nearest Neighbor Algorithm for $k = 3$, what shape should be assigned to the mystery circle?



(b) Using the k-Nearest Neighbor Algorithm for $k = 5$, what shape should be assigned to the mystery circle?

(c) Write a function that takes as input:

- (x,y): coordinates of new point
- a value k, and
- a list of tuples, `(x,y,shape)`, of the location and shape of each point.

and returns the predicted shape of the new point using the k-Nearest Neighbor approach.

8. Assume that you have that people move from the two states with the following probabilities:

75% New Yorkers stay in NY	25% of New Yorkers move to California
10% Californians move to NY	90% of CA stay in CA

which can be viewed as the Markov Chain: $\begin{pmatrix} \text{New York}_{t+1} \\ \text{California}_{t+1} \end{pmatrix} = \begin{pmatrix} 0.75 & 0.10 \\ 0.25 & 0.90 \end{pmatrix} \cdot \begin{pmatrix} \text{New York}_t \\ \text{California}_t \end{pmatrix}$

(a) Assume the initial populations are California has 40 million and New York has 20 million. What are the populations for each state:

- in 1 year?

- in 2 years?

- in 3 years?

(b) Compute the eigenvalues and eigenvectors for the transition matrix:

(c) What is the steady state for the population (i.e. the same number of people move in each direction, and the populations stay the same forever)?

Hint: Think about the eigenvalues. There's a value of v_t for which $Av_t = \lambda v_t = 1v_t = v_t$.

9. Given the points: $(1, 1)$, $(1, 3)$, and $(5, 2)$.

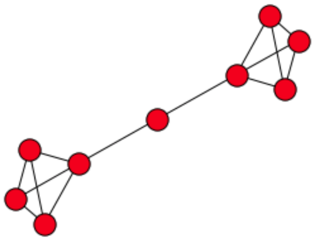
   (a) Compute the midpoint of the three points under the Euclidean (traditional) distance.

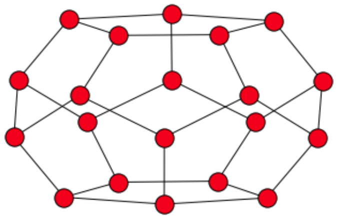
   (b) Draw the Voronoi Diagram for the Euclidean distances.

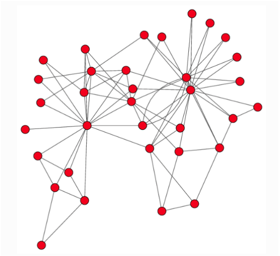
   (c) Compute the midpoint of the three points under the Manhattan distance.

   (d) Draw the Voronoi Diagram for the Manhattan distances.

10. In network analysis, the node most “central” to the graph plays a crucial role.







(a) There are several variations on centrality. Define what the “central” node means to use.

(b) Which node is central in each graph above? Explain your answer.

(c) Design a program that takes as input a graph, and returns the most central node: