CSCI4390/6390 – Data Mining
Fall 2011, Exam I
Total Points: 100 + 10 (bonus)

1. (30 points) Consider the "mixed" data given in Table 1. Here X_1 is a numeric attribute and X_2 is a categorical one. Assume that the domain of X_2 is given as $dom(X_2) = \{a, b\}$.

X_1	X_2
0.3	a
-0.3	b
0.44	a
-0.60	a
0.40	a
1.20	b
-0.12	a
-1.60	b
1.60	b
-1.32	a

Table 1: Mixed Data

Answer the following questions

(a) (15 points) Assuming that $a = 1$ and $b = -1$ for attribute X_2 compute the covariance between X_1 and X_2?

```
Dataset:
X1    X2
 0.30  1
-0.30 -1
 0.44  1
-0.60  1
 0.40  1
 1.20 -1
-0.12  1
-1.60 -1
 1.60 -1
-1.32  1

mean = (0, 0.2)

After centering:
X1    X2
 0.30  0.8
-0.30 -1.2
 0.44  0.8
-0.60  0.8
 0.40  0.8
 1.20 -1.2
-0.12  0.8
-1.60 -1.2
 1.60 -1.2
-1.32  0.8


cov = X1 dot X2 / 10 = -1.8/10 = -0.18
```

(b) (15 points) Assuming that X_1 is discretized into three bins, as follows:

$$b_1 = (-2, -0.5]$$
$$b_2 = (-0.5, 0.5]$$
$$b_3 = (0.5, 2]$$

Construct the contingency table between the discretized X_1, and X_2, and determine whether they are dependent or not at the 5% significance level, using the chi-squared table below.

Chi-Square Probabilities:*p-values* for different Chi-Square values are given for various degrees of freedom *df*. For example for $df = 5$, a chi-Square value of $\chi^2 = 11.070$ has a *p-value* of 0.05.

p-value	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.02 5	0.01	0.005
df=1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
df=2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
df=3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
df=4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
df=5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
df=6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548

Answer:

The contingency table is given as:

```
        a b
  b1 2 1 | 3
  b2 4 1 | 5
  b3 0 2 | 2
  -------
     6 4
```

The expected counts are:

```
    a   b
b1 1.8 1.2
b2 3.0 2.0
b3 1.2 0.8
```

The difference between observed and expected values is follows:

```
       a    b
  b1  0.2 -0.2
  b2  1.0 -1.0
  b3 -1.2  1.2
```

The final chi-square value is $\chi^2 = 3.89$. With 2 degrees of freedom, at 5% significance level, we have a χ^2 value of 5.99. This the value 3.89 is well within the acceptance region for the null hypothesis that the variables are independent. Put differently, there is nothing surprising about the value 3.89. Thus we conclude that the variables are not dependent, but rather they are independent.

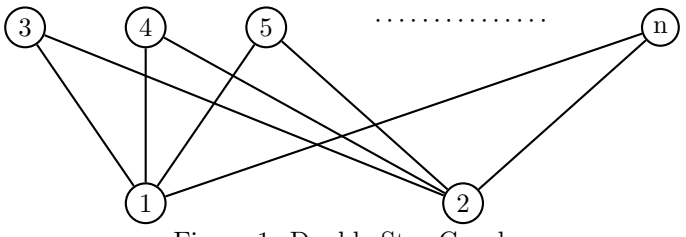


Figure 1: Double Star Graph

2. (30 points+ 10 Bonus) Consider the double star graph given in Figure 1 with n nodes, where only the nodes 1 and 2 are connected to all other vertices, and there are no other links. Answer the following questions (treating n as a variable).

(a) (15 pts) What is the degree distribution for this graph? What is the mean degree?

(b) (15 pts) What is the clustering coefficient $C(G)$ for the entire graph? What happens to the clustering coefficient as $n \to \infty$?

(c) (**Bonus: 10 pts**) What is the degree variance for the graph?

Answer:

The degree distribution $f(k)$ is given as

k	n(k)	f(k)
2	n-2	1 - 2/n
n-2	2	2/n

Here $n(k)$ is the number of nodes with degree k, whereas $f(k)$ is the probability mass function.

The mean or average degree is given as:

$$2(1-2/n) + (n-2)(2/n) = 2 - 4/n + 2 - 4/n = 4 - 8/n$$

The clustering coefficient for n-2 nodes is 0, and the clustering coefficient for nodes 1 and 2 is also 0. The average clustering coefficient for the graph is thus 0. As $n \to \infty$, $C(G) \to 0$.

The variance for the degree can be computed as follows:

$$E[X^2] = 2^2(1-2/n) + (n-2)^2 \cdot 2/n = 4 - 8/n + 2n - 8 + 8/n = 2n - 4$$

$$E[X]^2 = (4 - 8/n)^2 = 16 - 64/n + 64/n^2$$

$$var(X) = E[X^2] - E[X]^2 = 2n - 20 + 64/n - 64/n^2$$

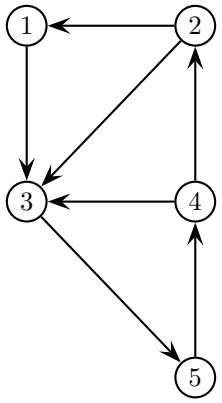


Figure 2: Directed Graph

3. (25 points) Consider the graph in Figure 2. Compute the authority scores for each node in the graph (based on the HITS approach). What is the dominant eigenvalue? Which are the high authority nodes? Can you explain why? (Note: you should scale the vector in each iteration to easily see convergence, which should happen in a few iterations).

The adjacency matrix A is given as:

0	0	1	0	0
1	0	1	0	0
0	0	0	0	1
0	1	1	0	0
0	0	0	1	0

The authority matrix $A^T A$ is given as:

```
1    0    1    0    0
0    1    1    0    0
1    1    3    0    0
0    0    0    1    0
0    0    0    0    1
```

Starting with the initial vector $\mathbf{x}_0 = (1, 1, 1, 1, 1)^T$, successive iterations give:

$\mathbf{x}_1 = (2, 2, 5, 1, 1)^T$ or $(0.4, 0.4, 1, 0.2, 0.2)^T$ after scaling by dividing all entries by 5

$\mathbf{x}_2 = (1.4, 1.4, 3.8, 0.2, 0.2)^T$ or $(0.37, 0.37, 1, 0.053, 0.053)^T$ after scaling by dividing all entries by 3.8

$\mathbf{x}_3 = (1.37, 1.37, 3.74, 0.053, 0.053)^T$ or $(0.37, 0.37, 1, 0.014, 0.014)^T$ after scaling by dividing all entries by 3.74

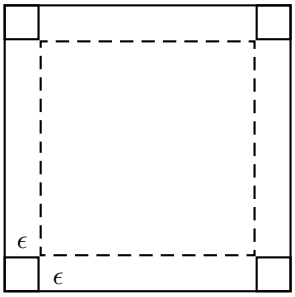
$\mathbf{x}_4 = (1.37, 1.37, 3.74, 0.014, 0.014)^T$

Normalizing $\mathbf{x}_4$ we get the final authority score vector:
$(0.33, 0.33, 0.89, 0, 0)^T$.

The eigenvalue is 3.74.

The high authority node is 3, since several of the other nodes point to it.

4. (15 points) Consider the corner hypercubes of length $\epsilon < 1$ inside a unit-hypercube. The 2D case is shown in the figure below:



   (a) (7.5 points) Derive an expression for the volume occupied by all of the corner hypercubes of length $\epsilon < 1$, as a function of the dimension d. What happens to the fraction of the volume in the corners as $d \to \infty$.

   (b) (7.5 points) What is the fraction of volume occupied by the thin hypercube shell of width $\epsilon < 1$ as a fraction of the total volume of the outer (unit) hypercube, as $d \to \infty$ (e.g., in the 2D case, the thin shell is the space between the outer square (solid) and inner square (dashed)).

a) In d-dimensions, there are 2^d corners, and the volume of each corner is ϵ^d. Thus the fraction of the volume in the corners is $(2\epsilon)^d$. As $d \to \infty$, we can see that if $\epsilon \leq 0.5$, then the volume goes to 0, otherwise, it increases without bound.

b) The volume in the thin shell is given as:

$$1 - (1 - 2\epsilon)^d$$

When $\epsilon \leq 0.5$, then the volume of the shell approaches 1, i.e., it contains all of the volume of the shell. If $\epsilon > 0.5$, given the fact that $\epsilon < 1$, even though for smaller dimensions the volume fluctuates, with volume less than 1 for even dimensions and more than one for odd dimensions, as $d \to \infty$, the volume still approaches 1.