DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Question 1:

Given two strings s1 = abcdefabcdf, and s2 = bcdefaddcfa, compute their shingle vectors (using 3-shingles). Compute the Hamming distance and Jaccard similarity between the strings using their shingle vectors. Show the steps.

Answer:

a b c d e f a b c d f

b c d e f a d d c f a

Shingle vectors:

S1 = { abc, bcd, cde, def, efa, fab, cdf } count = 7
S2 = { bcd, cde, def, efa, fad, add, ddc, dcf, cfa } count = 9

Union of shingle vectors:

$S1 \cup S2$ = { abc, bcd, cde, def, efa, fab, fad, cdf, add, ddc, dcf, cfa }

Intersection of shingle vectors:

$S1 \cap S2$ = { bcd, cde, def, efa }

DATA MINING

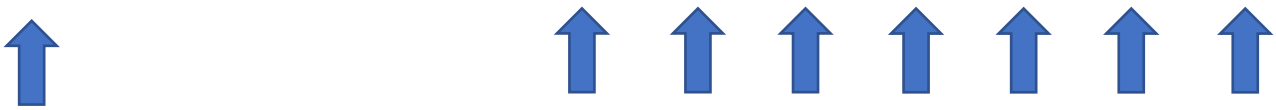
Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Hamming Distance:

Hamming distance between two shingle vectors of a string can be calculated by summing the number of unmatching elements from both the shingle vector sets.

	abc	bcd	cde	def	efa	fab	fad	cdf	add	ddc	dcf	cfa
S1	1	1	1	1	1	1	0	1	0	0	0	0
S2	0	1	1	1	1	0	1	0	1	1	1	1



Hence Hamming Distance = 8

Jaccard Similarity:

$| S1 \cup S2 | = 12$ $| S1 \cap S2 | = 4$

Jaccard Similarity $= \frac{| S1 \cap S2 |}{| S1 \cup S2 |} = \frac{4}{12} = \frac{1}{3} = 0.334$

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Question 2:

Given a shingle(word)-document matrix as below,

Shingle ID	*d1*	*d2*	*d3*	*d4*	*d5*	*d6*
0	0	1	0	1	1	0
1	0	0	1	0	0	0
2	1	1	0	0	0	0
3	1	1	0	1	1	1
4	1	0	0	1	1	1
5	0	1	0	0	1	1

(a) Compute the Jaccard similarity for all pairs of the six documents.

$$\text{Jaccard Similarity} = \frac{|\,S1 \cap S2\,|}{|\,S1 \cup S2\,|}$$

$$|\,S1 \cup S2\,| = f_{01} + f_{10} + f_{11} \qquad |\,S1 \cap S2\,| = f_{11}$$

1) **d1 and d2 :**

$|\,d1 \cup d2\,| = 5$ $|\,d1 \cap d2\,| = 2$

$$\text{Jaccard Similarity} = \frac{|\,d1 \cap d2\,|}{|\,d1 \cup d2\,|} = \frac{2}{5} = 0.4$$

2) **d1 and d3:**

$|\,d1 \cup d3\,| = 4$ $|\,d1 \cap d3\,| = 0$

$$\text{Jaccard Similarity} = \frac{|\,d1 \cap d3\,|}{|\,d1 \cup d3\,|} = \frac{0}{4} = 0$$

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

3) **d1 and d4:**

$| d1 \cup d4 | = 4$ $| d1 \cap d4 | = 2$

Jaccard Similarity = $\frac{| d1 \cap d4 |}{| d1 \cup d4 |} = \frac{2}{4} = 0.5$

4) **d1 and d5:**

$| d1 \cup d5 | = 5$ $| d1 \cap d5 | = 2$

Jaccard Similarity = $\frac{| d1 \cap d2 |}{| d1 \cup d2 |} = \frac{2}{5} = 0.4$

5) **d1 and d6:**

$| d1 \cup d6 | = 4$ $| d1 \cap d6 | = 2$

Jaccard Similarity = $\frac{| d1 \cap d6 |}{| d1 \cup d6 |} = \frac{2}{4} = 0.5$

6) **d2 and d3:**

$| d2 \cup d3 | = 5$ $| d2 \cap d3 | = 0$

Jaccard Similarity = $\frac{| d2 \cap d3 |}{| d2 \cup d3 |} = \frac{0}{5} = 0$

7) **d2 and d4:**

$| d2 \cup d4 | = 5$ $| d2 \cap d4 | = 2$

Jaccard Similarity = $\frac{| d2 \cap d4 |}{| d2 \cup d4 |} = \frac{2}{5} = 0.4$

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

8) **d2 and d5:**

$| d2 \cup d5 | = 5$ $| d2 \cap d5 | = 3$

Jaccard Similarity = $\frac{| d2 \cap d5 |}{| d2 \cup d5 |} = \frac{3}{5} = 0.6$

9) **d2 and d6:**

$| d2 \cup d6 | = 5$ $| d2 \cap d6 | = 2$

Jaccard Similarity = $\frac{| d2 \cap d6 |}{| d2 \cup d6 |} = \frac{2}{5} = 0.4$

10) **d3 and d4:**

$| d3 \cup d4 | = 4$ $| d3 \cap d4 | = 0$

Jaccard Similarity = $\frac{| d3 \cap d4 |}{| d3 \cup d4 |} = \frac{0}{4} = 0$

11) **d3 and d5:**

$| d3 \cup d5 | = 5$ $| d3 \cap d5 | = 0$

Jaccard Similarity = $\frac{| d3 \cap d5 |}{| d3 \cup d5 |} = \frac{0}{5} = 0$

12) **d3 and d6:**

$| d3 \cup d6 | = 4$ $| d3 \cap d6 | = 0$

Jaccard Similarity = $\frac{| d3 \cap d6 |}{| d3 \cup d6 |} = \frac{0}{4} = 0$

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

13) **d4 and d5:**

$| d4 \cup d5 | = 4$ $| d4 \cap d5 | = 3$

Jaccard Similarity = $\frac{| d4 \cap d5 |}{| d4 \cup d5 |} = \frac{3}{4} = 0.75$

14) **d4 and d6:**

$| d4 \cup d6 | = 4$ $| d4 \cap d6 | = 2$

Jaccard Similarity = $\frac{| d4 \cap d6 |}{| d4 \cup d6 |} = \frac{2}{4} = 0.5$

15) **d5 and d6:**

$| d5 \cup d6 | = 12$ $| d5 \cap d6 | = 4$

Jaccard Similarity = $\frac{| d5 \cap d6 |}{| d5 \cup d6 |} = \frac{3}{4} = 0.75$

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

(b) Compute the minhash signatures for each column (document) if the following hash functions are used: h1(x) = (2x + 1)%6; h2(x) = (3x + 2)%6; h3(x) = (5x + 2)%6; and h4(x) = (7x + 3)%6.

Min Hash Signatures:

h1(x) = (2x + 1)%6

h1(0) = (2 x 0 + 1) % 6 = 1
h1(1) = (2 x 0 + 1) % 6 = 3
h1(2) = (2 x 0 + 1) % 6 = 5
h1(3) = (2 x 0 + 1) % 6 = 1
h1(4) = (2 x 0 + 1) % 6 = 3
h1(5) = (2 x 0 + 1) % 6 = 5

h2(x) = (3x + 2)%6

h2(0) = (3 x 0 + 2) % 6 = 2
h2(1) = (3 x 0 + 2) % 6 = 5
h2(2) = (3 x 0 + 2) % 6 = 2
h2(2) = (3 x 0 + 2) % 6 = 5
h2(2) = (3 x 0 + 2) % 6 = 2
h2(2) = (3 x 0 + 2) % 6 = 5

h3(x) = (5x + 2)%6

h1(0) = (5 x 0 + 2) % 6 = 2
h1(1) = (5 x 0 + 2) % 6 = 1
h1(2) = (5 x 0 + 2) % 6 = 0
h1(3) = (5 x 0 + 2) % 6 = 5
h1(4) = (5 x 0 + 2) % 6 = 4
h1(5) = (5 x 0 + 2) % 6 = 3

h4(x) = (7x + 3)%6

h1(0) = (7 x 0 + 3) % 6 = 3
h1(1) = (7 x 0 + 3) % 6 = 4
h1(2) = (7 x 0 + 3) % 6 = 5
h1(2) = (7 x 0 + 3) % 6 = 0
h1(2) = (7 x 0 + 3) % 6 = 1
h1(2) = (7 x 0 + 3) % 6 = 2

Shingle_ID	**h1(x)**	**h2(x)**	**h3(x)**	**h4(x)**
0	1	2	2	3
1	3	5	1	4
2	5	2	0	5
3	1	5	5	0
4	3	2	4	1
5	5	5	3	2

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

h1(x) = (2x + 1)%6

d1 = 0 0 1 1 1 0	h1(d1) = min(5, 1, 3) = 1
d2 = 1 0 1 1 0 1	h1(d2) = min (1, 5, 1, 5) = 1
d3 = 0 1 0 0 0 0	h1(d3) = min(3) = 3
d4 = 1 0 0 1 1 0	h1(d4) = min(1, 1, 3) = 1
d5 = 1 0 0 1 1 1	h1(d5) = min(1, 1, 3, 5) = 1
d6 = 0 0 0 1 1 1	h1(d6) = min(1, 3, 5) = 1

h2(x) = (3x + 2)%6

d1 = 0 0 1 1 1 0	h2(d1) = min(2, 5, 2) = 2
d2 = 1 0 1 1 0 1	h2(d2) = min (2, 2, 5, 5) = 2
d3 = 0 1 0 0 0 0	h2(d3) = min(5) = 5
d4 = 1 0 0 1 1 0	h2(d4) = min(2, 5, 2) = 2
d5 = 1 0 0 1 1 1	h2(d5) = min(2, 5, 2, 5) = 2
d6 = 0 0 0 1 1 1	h2(d6) = min(5, 2, 5) = 2

h3(x) = (5x + 2)%6

d1 = 0 0 1 1 1 0	h3(d1) = min(0, 5, 4) = 0
d2 = 1 0 1 1 0 1	h3(d2) = min (2, 0, 5, 3) = 0
d3 = 0 1 0 0 0 0	h3(d3) = min(1) = 1
d4 = 1 0 0 1 1 0	h3(d4) = min(2, 5, 4) = 2
d5 = 1 0 0 1 1 1	h3(d5) = min(2, 5, 4, 3) = 2
d6 = 0 0 0 1 1 1	h3(d6) = min(5, 4, 3) = 3

h4(x) = (7x + 3)%6

d1 = 0 0 1 1 1 0	h4(d1) = min(5, 0, 1) = 0
d2 = 1 0 1 1 0 1	h4(d2) = min (3, 5, 0, 2) = 0
d3 = 0 1 0 0 0 0	h4(d3) = min(4) = 4
d4 = 1 0 0 1 1 0	h4(d4) = min(3, 0, 1) = 0
d5 = 1 0 0 1 1 1	h4(d5) = min(3, 0, 1, 2) = 0
d6 = 0 0 0 1 1 1	h4(d6) = min(0, 1, 2) = 0

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Signature Matrix:

	d1	**d2**	**d3**	**d4**	**d5**	**d6**
h1(x)	1	1	3	1	1	1
h2(x)	2	2	5	2	2	2
h3(x)	0	0	1	2	2	3
h4(x)	0	0	4	0	0	0

(c) Compute the similarities for all pairs of the six documents using the signatures.

1) **d1 and d2 :**

d1 = 1 2 0 0
d2 = 1 2 0 0

similarity(d1, d2) = $\frac{4}{4}$ = 1

2) **d1 and d3 :**

d1 = 1 2 0 0
d3 = 3 5 1 4

similarity(d1, d3) = $\frac{0}{4}$ = 0

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

3) **d1 and d4 :**

d1 = 1 2 0 0
d4 = 1 2 2 0

similarity(d1, d4) = $\frac{3}{4}$ = 0.75

4) **d1 and d5 :**

d1 = 1 2 0 0
d5 = 1 2 2 0

similarity(d1, d5) = $\frac{3}{4}$ = 0.75

5) **d1 and d6 :**

d1 = 1 2 0 0
d6 = 1 2 3 0

similarity(d1, d6) = $\frac{3}{4}$ = 0.75

6) **d2 and d3 :**

d2 = 1 2 0 0
d3 = 3 5 1 4

similarity(d2, d3) = $\frac{0}{4}$ = 0

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

7) **d2 and d4 :**

d2 = 1 2 0 0
d4 = 1 2 2 0

similarity(d2, d4) = $\frac{3}{4}$ = 0.75

8) **d2 and d5 :**

d2 = 1 2 0 0
d5 = 1 2 2 0

similarity(d2, d5) = $\frac{3}{4}$ = 0.75

9) **d2 and d6 :**

d2 = 1 2 0 0
d6 = 1 2 3 4

similarity(d2, d6) = $\frac{3}{4}$ = 0.75

10) **d3 and d4 :**

d3 = 3 5 1 4
d4 = 1 2 2 0

similarity(d3, d4) = $\frac{0}{4}$ = 0

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

11) **d3 and d5 :**

d3 = 3 5 1 4
d5 = 1 2 2 0

similarity(d3, d5) = $\frac{0}{4}$ = 0

12) **d3 and d6 :**

d3 = 3 5 1 4
d6 = 1 2 3 0

similarity(d3, d6) = $\frac{0}{4}$ = 0

13) **d4 and d5 :**

d4 = 1 2 2 0
d5 = 1 2 2 0

similarity(d4, d5) = $\frac{4}{4}$ = 1

14) **d4 and d6 :**

d4 = 1 2 2 0
d6 = 1 2 3 0

similarity(d4, d6) = $\frac{3}{4}$ = 0.75

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

15) **d5 and d6 :**

d5 = 1 2 2 0
d6 = 1 2 3 0

similarity(d4, d5) = $\frac{3}{4}$ = 0.75

(d) Does minhash provide good signatures for computing the document similarity for this example?

No, MinHash provides good similarity calculations only when the documents are completely different, i.e. there is no similarity between 2 docs.

As you can see, minhash signatures give us similarity values for d1 and d2 as 1 although the documents are not completely similar as indicated by the Jaccard Similarity. So there is some amount of ambiguity present in the similarity calculations using minhash.

(e) Can you design another hash function that provides true permutation?

h(x) = (11x + 2) % 6

h(0) = (11 x 0 + 2) % 6 = 2
h(1) = (11 x 1 + 2) % 6 = 1
h(2) = (11 x 2 + 2) % 6 = 0
h(3) = (11 x 3 + 2) % 6 = 5
h(4) = (11 x 4 + 2) % 6 = 4
h(5) = (11 x 5 + 2) % 6 = 3

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Question 3:

Credit card fraud detection using KNN. Please use this code as the start code.

1) Before calling KNN for classification, were there data processing steps applied in the start code? What distance metric was used in KNN in the start code?

Yes, before calling the KNN for classification, StandardScalar is fitted to our column 'Amount' to normalize columns' values into a new column 'AmountNormalized'.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import metrics
data['AmountNormalized'] = StandardScaler().fit_transform(data['Amount'].values.reshape(-1, 1))
data['AmountNormalized'].describe()
```

Distance metric used is Minkowski with p value = 2 which is the Eucledian distance metric.

Minkowski - sum(|x - y|^p)^(1/p)
p=2 so sum(|x - y|^2)^(1/2) which is the Eucledian distance metric.

```
#KNN
from sklearn.neighbors import KNeighborsClassifier
#train
knn = KNeighborsClassifier(n_neighbors=5, metric= 'minkowski', p=2)
```

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

2) Using “copy and edit” function on Kaggle to create your only KNN classifier. Try different settings, including different k values and different distance metrics and report how the classifier’s performance changes.

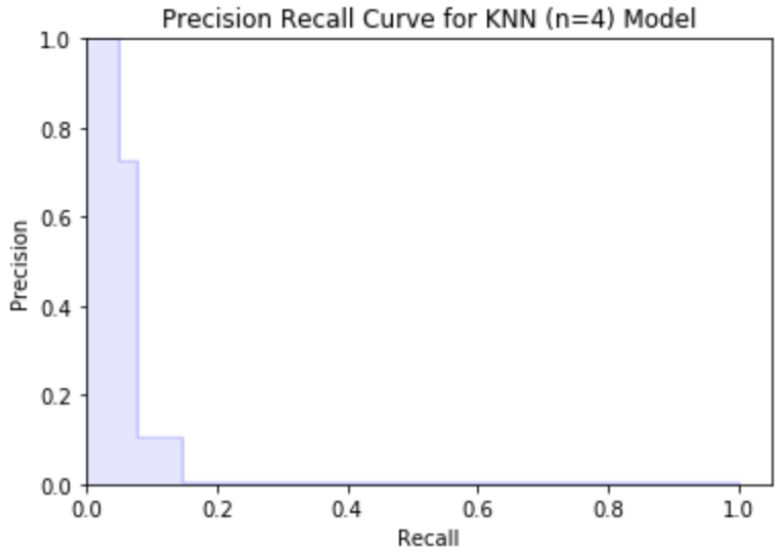
For 4 neighbors, n=4

```
Confusion Matrix for KNN (n=4) Model
[[56861     0]
 [   96     5]]
Classification Report for KNN (n=4) Model
              precision    recall  f1-score   support

       False   0.998315  1.000000  0.999157     56861
        True   1.000000  0.049505  0.094340       101

    accuracy                       0.998315     56962
   macro avg   0.999157  0.524752  0.546748     56962
weighted avg   0.998318  0.998315  0.997552     56962

Area under under ROC curve for KNN (n=4) Model
0.5732184146819036
```



DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

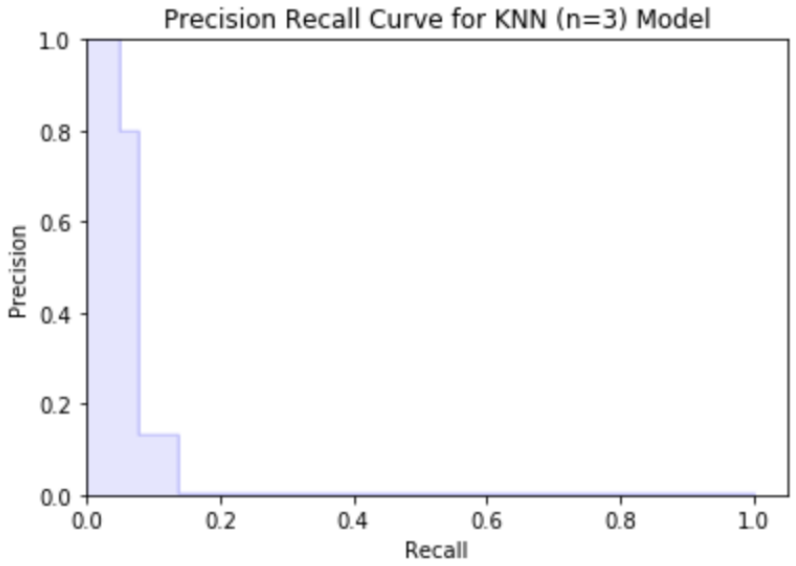
For 3 neighbors, n=3

```
Confusion Matrix for KNN (n=3) Model
[[56859     2]
 [   93     8]]
Classification Report for KNN (n=3) Model
              precision    recall  f1-score   support

       False   0.998367  0.999965  0.999165     56861
        True   0.800000  0.079208  0.144144       101

    accuracy                       0.998332     56962
   macro avg   0.899184  0.539586  0.571655     56962
weighted avg   0.998015  0.998332  0.997649     56962

Area under under ROC curve for KNN (n=3) Model
0.5685685485240106
```



DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

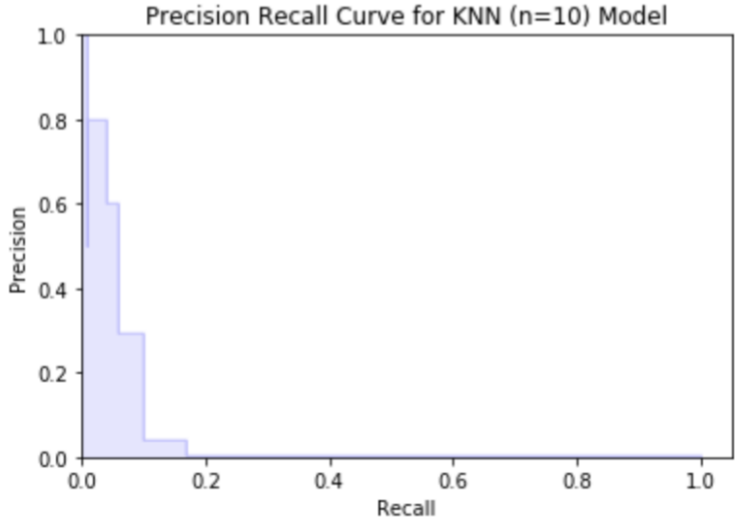
For 10 neighbors, n=10

```
Confusion Matrix for KNN (n=10) Model
[[56861     0]
 [  100     1]]
Classification Report for KNN (n=10) Model
              precision    recall  f1-score   support

       False   0.998244  1.000000  0.999121     56861
        True   1.000000  0.009901  0.019608       101

    accuracy                       0.998244     56962
   macro avg   0.999122  0.504950  0.509365     56962
weighted avg   0.998248  0.998244  0.997385     56962

Area under under ROC curve for KNN (n=10) Model
0.580813016142718
```



In the above observations, we can see that as we decrease the number of neighbors from 4 to 3, there is an increase in false positives and decrease in false negatives. Hence, having a higher number of neighbors could help us with increasing the accuracy. But, When n = 10, false positives are 0, and false negatives are 100 which increased from 93 which is not ideal either, hence, having a really high amount of neighbors also doesn't indicate better accuracy. But area under roc curve increases from n = 4 to n = 10 which indicates model is able to classify better when n is increased. So based on these observations and we can identify the ideal n value

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

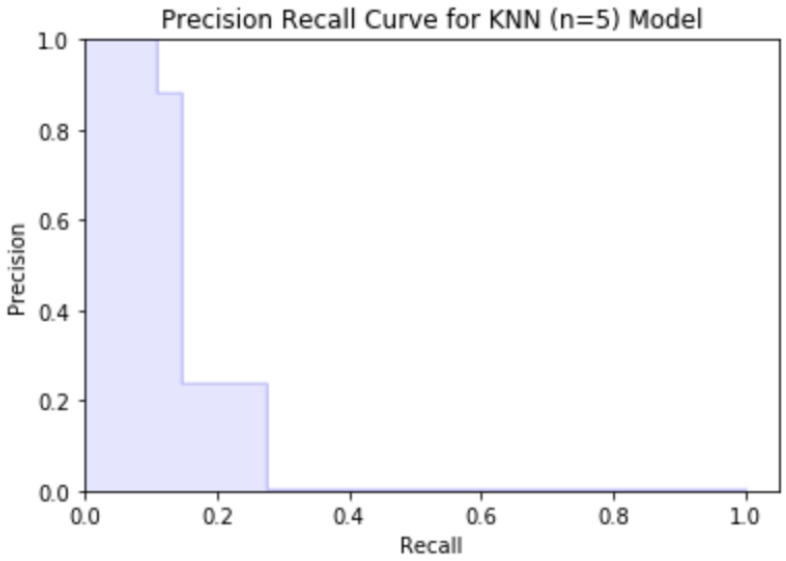
For Manhattan distance, metric = minkowski p =1 and n = 5 neighbors

```
Confusion Matrix for KNN (n=5) Model
[[56861     0]
 [   90    11]]
Classification Report for KNN (n=5) Model
              precision    recall  f1-score   support

       False   0.998420  1.000000  0.999209     56861
        True   1.000000  0.108911  0.196429       101

    accuracy                       0.998420     56962
   macro avg   0.999210  0.554455  0.597819     56962
weighted avg   0.998422  0.998420  0.997786     56962

Area under under ROC curve for KNN (n=5) Model
0.6379445202570591
```



DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

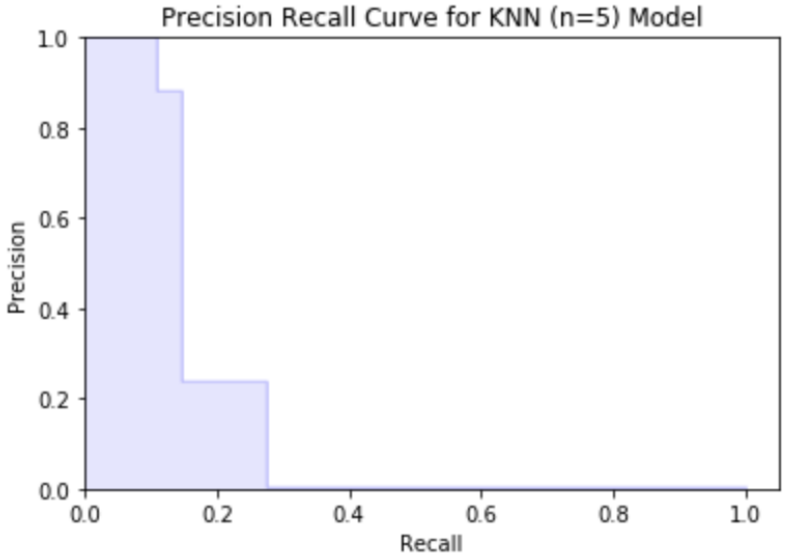
For metric = Manhattan and n= 5 neighbors

```
Confusion Matrix for KNN (n=5) Model
[[56861     0]
 [   90    11]]
Classification Report for KNN (n=5) Model
              precision    recall  f1-score   support

       False   0.998420  1.000000  0.999209     56861
        True   1.000000  0.108911  0.196429       101

    accuracy                       0.998420     56962
   macro avg   0.999210  0.554455  0.597819     56962
weighted avg   0.998422  0.998420  0.997786     56962

Area under under ROC curve for KNN (n=5) Model
0.6379445202570591
```



From the above two observations, we can see that we have used metric as Manhattan for one and metric as minkowski and p value as 1 which is ideally the Manhattan distance itself for n = 5 neighbors.

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

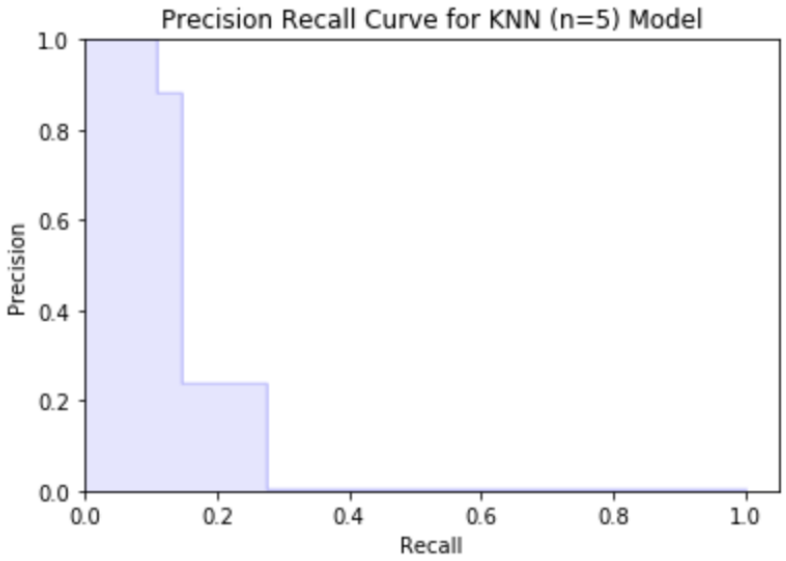
For metric = Manhattan and n= 5 neighbors

```
Confusion Matrix for KNN (n=5) Model
[[56861     0]
 [   90    11]]
Classification Report for KNN (n=5) Model
              precision    recall  f1-score   support

       False   0.998420  1.000000  0.999209     56861
        True   1.000000  0.108911  0.196429       101

    accuracy                       0.998420     56962
   macro avg   0.999210  0.554455  0.597819     56962
weighted avg   0.998422  0.998420  0.997786     56962

Area under under ROC curve for KNN (n=5) Model
0.6379445202570591
```



DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

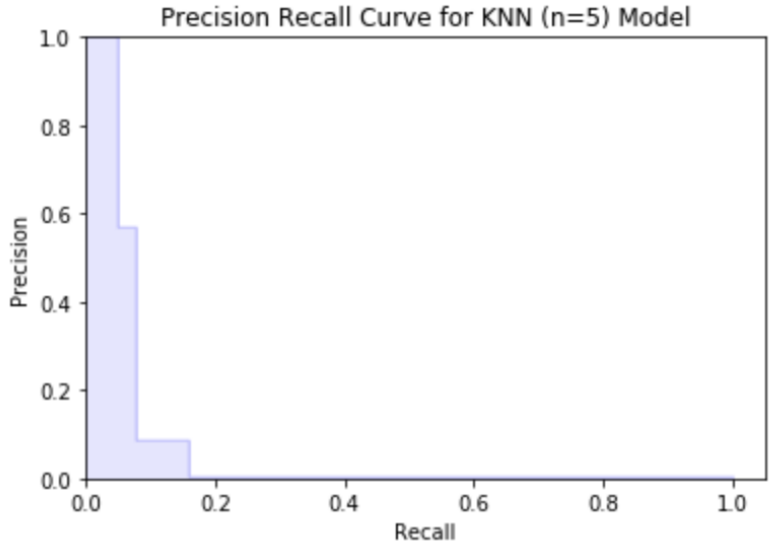
For metric = Euclidean and n= 5 neighbors

```
Confusion Matrix for KNN (n=5) Model
[[56861     0]
 [   96     5]]
Classification Report for KNN (n=5) Model
              precision    recall  f1-score   support

       False   0.998315  1.000000  0.999157     56861
        True   1.000000  0.049505  0.094340       101

    accuracy                       0.998315     56962
   macro avg   0.999157  0.524752  0.546748     56962
weighted avg   0.998318  0.998315  0.997552     56962

Area under under ROC curve for KNN (n=5) Model
0.5777933195088735
```



For n = 5, changing the metric from Manhattan to Euclidean gives us a lower AUC ROC. Though, we can see that true negatives have increased and false negatives are decreased, which tell us that Euclidean distance may be an ideal distance metric to use with 5 neighbors.

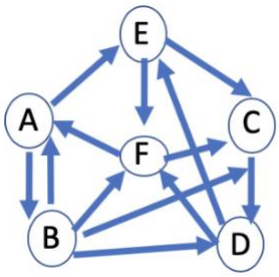
DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Question 4:

Link Analysis



(a) Given the above toy web (with 6 web pages, A, B, ⋯, F), derive its transition probability matrix.

Transition probability matrix:

	A	**B**	**C**	**D**	**E**	**F**
A	0	0.5	0	0	0.5	0
B	0.25	0	0.25	0.25	0	0.25
C	0	0	0	1	0	0
D	0	0	0	0	0.5	0.5
E	0	0	0.5	0	0	0.5
F	0.5	0	0.5	0	0	0

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

(b) Assume a surfer is on web page A, what's the probability that the person will be next visiting B and then D (5 points)?

Probability that a surfer goes from Page A to Page B = P(A->B) = 0.5

Probability that a surfer goes from Page B to Page D = P(B->D) = 0.25

Probability that a surfer goes from Page A to Page B and then B to D
= P(A->B->D)
= P(A->B) * P(B->D)
= 0.5 * 0.25
= 0.125

(c) Implement the PageRank algorithm that uses power iteration. Your program takes a matrix of web links as the input, and computes the ranks of the web pages. Test your program using the matrix your derived in (a). Try different initial distributions and see if the result changes or not.

```
import numpy as np

trans_prob = np.array([[0,0.5,0,0,0.5,0],[0.25,0,0.25,0.25,0,0.25],[0,0,0,1,0,0],[0,0,0,0,0.5,0.5],[0,0,0.5,0,0,0.5],[0.5,0,0.5,0,0,0]])

damp = np.array([0.16,0.16,0.16,0.16,0.16,0.16])

prev = np.dot(trans_prob.T,damp)

def pageRank(trans_prob, prev):
    count=0
    while True:
        mul = np.dot(trans_prob.T,prev)
        count+=1
        if np.array_equal(np.around(mul, decimals = 3),np.around(prev, decimals = 3)):
            print("Final Page Rank:")
            print(list(prev))
            break
        else:
            prev = mul
            print("Page Ranks for iteration : {}".format(count))
            print(list(prev))
            print()

pageRank(trans_prob, prev)
```

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

```
Page Ranks for iteration : 1
[0.12000000000000001, 0.06, 0.19999999999999998, 0.22, 0.16, 0.2]

Page Ranks for iteration : 2
[0.115, 0.060000000000000005, 0.195, 0.21499999999999997, 0.17, 0.20500000000000002]

Page Ranks for iteration : 3
[0.11750000000000001, 0.0575, 0.2025, 0.21000000000000002, 0.16499999999999998, 0.2075]

Page Ranks for iteration : 4
[0.118125, 0.058750000000000004, 0.20062499999999997, 0.216875, 0.16375, 0.201875]

Page Ranks for iteration : 5
[0.115625, 0.0590625, 0.19749999999999998, 0.21531249999999996, 0.1675, 0.20500000000000002]

Page Ranks for iteration : 6
[0.11726562500000001, 0.0578125, 0.20101562500000003, 0.21226562499999999, 0.16546875, 0.206171875]

Page Ranks for iteration : 7
[0.1175390625, 0.058632812500000006, 0.2002734375, 0.21546875000000004, 0.164765625, 0.2033203125]

Page Ranks for iteration : 8
[0.116318359375, 0.05876953125, 0.19870117187499997, 0.214931640625, 0.16650390625000003, 0.20477539062500003]

Page Ranks for iteration : 9
[0.11708007812500001, 0.0581591796875, 0.20033203125000001, 0.21339355468749996, 0.165625, 0.20541015625000003]

Page Ranks for iteration : 10
[0.11724487304687502, 0.05854003906250001, 0.20005737304687501, 0.214871826171875, 0.16523681640625, 0.20404907226562496]

Page Ranks for iteration : 11
[0.11665954589843748, 0.05862243652343751, 0.1992779541015625, 0.21469238281250003, 0.16605834960937502, 0.2046893310546875]

Page Ranks for iteration : 12
[0.11700027465820313, 0.05832977294921874, 0.2000294494628906, 0.21393356323242185, 0.16567596435546875, 0.2050309753417969]

Page Ranks for iteration : 13
[0.11709793090820314, 0.058500137329101566, 0.1999359130859375, 0.2146118927001953, 0.16546691894531249, 0.20438720703125]

Page Ranks for iteration : 14
[0.11681863784790039, 0.05854896545410157, 0.19955209732055662, 0.21456094741821288, 0.1658549118041992, 0.20466444015502927]

Page Ranks for iteration : 15
[0.11696946144104003, 0.058409318923950196, 0.19989691734313964, 0.21418933868408202, 0.16568979263305664, 0.20484517097473143]

Final Page Rank:
[0.11696946144104003, 0.058409318923950196, 0.19989691734313964, 0.21418933868408202, 0.16568979263305664, 0.20484517097473143]
```

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Different initial distributions:

```
import numpy as np

trans_prob = np.array([[0,0.5,0,0,0.5,0],[0.25,0,0.25,0.25,0,0.25],[0,0,0,1,0,0],[0,0,0,0,0.5,0.5],[0,0,0.5,0,0,0.5],[0.5,0,0.5,0,0,0]])

damp = np.array([0.25,0.25,0.25,0.25,0.25,0.25])

prev = np.dot(trans_prob.T,damp)

def pageRank(trans_prob, prev):
    count=0
    while True:
        mul = np.dot(trans_prob.T,prev)
        count+=1
        if np.array_equal(np.around(mul, decimals = 3),np.around(prev, decimals = 3)):
            print("Final Page Rank:")
            print(list(prev))
            break
        else:
            prev = mul
            print("Page Ranks for iteration : {}".format(count))
            print(list(prev))
            print()

pageRank(trans_prob, prev)
```

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

```
Page Ranks for iteration : 1
[0.1875, 0.09375, 0.3125, 0.34375, 0.25, 0.3125]

Page Ranks for iteration : 2
[0.1796875, 0.09375, 0.3046875, 0.3359375, 0.265625, 0.3203125]

Page Ranks for iteration : 3
[0.18359375, 0.08984375, 0.31640625, 0.328125, 0.2578125, 0.32421875]

Page Ranks for iteration : 4
[0.1845703125, 0.091796875, 0.3134765625, 0.3388671875, 0.255859375, 0.3154296875]

Page Ranks for iteration : 5
[0.1806640625, 0.09228515625, 0.30859375, 0.33642578125, 0.26171875, 0.3203125]

Page Ranks for iteration : 6
[0.1832275390625, 0.09033203125, 0.3140869140625, 0.3316650390625, 0.258544921875, 0.3221435546875]

Page Ranks for iteration : 7
[0.18365478515625, 0.09161376953125, 0.31292724609375, 0.336669921875, 0.2574462890625, 0.31768798828125]

Page Ranks for iteration : 8
[0.1817474365234375, 0.091827392578125, 0.3104705810546875, 0.3358306884765625, 0.260162353515625, 0.3199615478515625]

Page Ranks for iteration : 9
[0.1829376220703125, 0.09087371826171875, 0.313018798828125, 0.33342742919921875, 0.2587890625, 0.320953369140625]

Page Ranks for iteration : 10
[0.1831951141357422, 0.09146881103515625, 0.3125896453857422, 0.3357372283935547, 0.2581825256347656, 0.31882667541503906]

Page Ranks for iteration : 11
[0.1822805404663086, 0.0915975570678711, 0.3113718032836914, 0.33545684814453125, 0.25946617126464844, 0.3198270797729492]

Page Ranks for iteration : 12
[0.18281292915344238, 0.0911402702331543, 0.3125460147857666, 0.3342711925506592, 0.2588686943054199, 0.3203608989715576]

Page Ranks for iteration : 13
[0.18296551704406738, 0.09140646457672119, 0.31239986419677734, 0.3353310823440552, 0.2585420608520508, 0.3193550109863281]

Page Ranks for iteration : 14
[0.18252912163734436, 0.09148275852203369, 0.31180015206336975, 0.33525148034095764, 0.2591482996940613, 0.3197881877422333]

Final Page Rank:
[0.18252912163734436, 0.09148275852203369, 0.31180015206336975, 0.33525148034095764, 0.2591482996940613, 0.3197881877422333]
```

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Question 5:
Write a summary for this review article: Information Security in Big Data: Privacy and Data Mining.

- The length of your summary is about one page.
- Write the summary in your own words; don't copy and paste.

Paper Description:

The paper is concerned about the growing popularity of Data Mining and it's threats to security due to advancements in the field of Big Data and other data related technologies. Focus on privacy preservation with data mining is the central goal of this paper. Studies corresponding to PPDM (Privacy preserving Data Mining) focus on mitigating privacy risks brought by Data Mining operations. Paper focus on ways to mitigate these privacy risks.

Summary:

The paper starts with defining an application scenario with data mining at it's core. A user – role based methodology is established based on KDD (Knowledge discovery using data). Four different types of users – Data Provider, Data Collector, Data Miner, Decision Maker.

Data Provider:
Individual sensitive information may be included in the data acquired from data providers. Data providers' privacy will be violated if data is released directly to the data miner, hence data modification is necessary. On the other side, following alteration, the data must still be valuable; otherwise, gathering the data is pointless.

Data Collector:
The primary goal of the data collector is to ensure that the updated data does not include any sensitive information while yet maintaining a high level of utility.

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Data Miner:
The data miner uses mining techniques to extract relevant information from data that has been supplied by the data collector while maintaining anonymity.

Decision Maker:
The data mining findings might be obtained directly from the data miner or via an Information Transmitter. It's conceivable that the information transmitter tampers with the mining results, either purposefully or inadvertently, putting the decision maker at risk. As a result, the decision maker is concerned about the mining findings' credibility.

Different approaches to privacy protection for Data Provider:

1) LIMIT THE ACCESS:
   The method of providing data is discussed here, whether data is provided voluntarily, “Active” way; or data is generated through provider’s routine activities. If the data is provided passively, then efficient measures could be taken by the data provider to limit the access of data for the data collector. Ways like Anti-tracking extensions, ad-blockers, encryption tools like VPN can be used.
2) TRADE PRIVACY FOR BENEFIT:
   Demographic data in return for personalized product recommendation. This is a tradeoff that can be possible. This trade deal is only possible if the returns provided are worth the data provider’s sensitive information.

Once the data is handed over by the provider, there is no guarantee about the safety of provider’s sensitive information.

DATA MINING

Homework 3

Name: Ameya Dalvi mail: abdalvi@iu.edu

Different approaches to privacy protection for Data Collector:

1) BASICS OF PPDP:
   Study of anonymization approaches for publishing sensitive data while preserving privacy. The concept here is table anonymization to prevent potential tampering by adeversaries.
2) PRIVACY-PRESERVING PUBLISHING OF SOCIAL NETWORK DATA:
   In the context of social networks, PPDP is primarily concerned with anonymizing graph data, which is far more difficult than anonymizing relational table data.
3) ATTACK MODEL:
   To de-anonymize people and discover linkages between de-anonymized persons, adversaries frequently rely on prior knowledge. T he Seed-and-Grow method for identifying people from an anonymised social graph simply based on graph topology. The method first finds a seed sub-graph, which is either planted by an attacker or revealed by a small number of users' cooperation, and then expands the seed bigger depending on the adversary's prior knowledge of users' social relationships.
4) PRIVACY-PRESERVING PUBLISHING OF TRAJECTORY DATA:
   Commercial (e.g., a telecommunications company) and public (e.g., a transportation business) organizations gather enormous amounts of individuals' trajectory data, i.e., sequences of successive location readings with time stamps, in order to deliver location-based services.