Data Mining Assignment 1:

Due on Saturday, 3rd February, 2018

Dr. Predrag Radivojac

Arnav Arnav (aarnav)

February 3, 2018

Contents

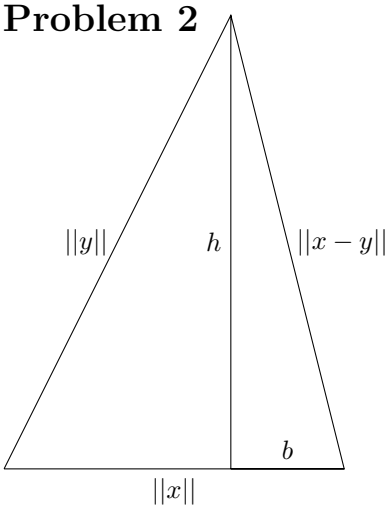
Problem 1

a. The term frequency-inverse document frequency encoding weights frequency each term in the document such that terms that are very common words and appear frequently across all documents have a lower weights, and rare words have more weight. This is helpful in text mining as there are a lot of common English terms that may not be useful for classifying documents in various groups. It helps us identify the words that can help us in identifying or classifying the documents.

b. The TF-IDF encoding weighs terms gives smaller weights to common terms in two documents. This may be a problem in a few cases. First, Consider that we have a corpus with only one document. Then the TF-IDF encoding of the document would be a vector of zeros, since the inverse document frequency would be zero for all the terms because they appear in all the documents. Thus we can not identify important terms for the document. Using only term frequency or count of terms, we will not have this problem.
Suppose, we have two documents of the same size such that their first K terms are same in the two documents and all other terms are different. We can assume that no two terms in a document are same for simplicity. Thus, the TF-IDF encoding of the K terms that occur in both the documents will be zero. and the rest of the terms will have same non-zero TF-IDF value. Thus even though these two documents are only partly similar, the TF-IDF encoding for both of the documents will be the same. Therefore, again looking at the encodings we can not differentiate between the two documents.

c. If a term occurs in only one document, then the inverse document frequency weight would be $log(n)$. Thus if a term appears only in one document, the TF-IDF encoding of the term will be large. This means that if a term appears only in one document then, it is more important than other terms to differentiate this document with other documents.
If a term appears in every document, then the inverse document frequency weight of the term would be $log(1)$, which is zero. This means that if a term appears in all the documents then it is a common term in the language and its occurrence does not help us differentiate a document with any other document.

Problem 2



Referring to the diagram above, we can say that:
Let θ be the angle between the vectors $\vec{x}$ and $\vec{y}$

Then, $h = ||\vec{y}||cos(\theta)$
and, $b = ||\vec{x}|| - ||\vec{y}||sin(\theta)$

Problem 2 continued on next page...

$=> (||\vec{x} - \vec{y}||)^2 = h^2 + b^2$
$=> (||\vec{x} - \vec{y}||)^2 = ||\vec{y}||^2 cos^2(\theta) + (||\vec{x}|| - ||\vec{y}||sin(\theta))^2$
$=> (||\vec{x} - \vec{y}||)^2 = ||\vec{y}||^2 cos^2(\theta) + ||\vec{y}||^2 sin^2(\theta) + ||\vec{x}||^2 + 2||\vec{x}||.||\vec{y}||cos(\theta)$
$=> (||\vec{x} - \vec{y}||)^2 = ||\vec{x}||^2 + ||\vec{y}||^2 + 2||\vec{x}||.||\vec{y}||cos(\theta)$ – (1)

Also since x and y are vectors, we know that,
$(||\vec{x} - \vec{y}||)^2 = (\vec{x} - \vec{y}).(\vec{x} - \vec{y})$
$=> (||\vec{x} - \vec{y}||)^2 = \vec{x}.\vec{x} + \vec{y}.\vec{y} + 2.\vec{x}.\vec{y}$
$=> (||\vec{x} - \vec{y}||)^2 = ||\vec{x}||^2 + ||\vec{y}||^2 + 2.\vec{x}.\vec{y}$ – (2)

Now, from equations (1) and (2), we have
$||\vec{x}|| + ||\vec{y}|| + 2||\vec{x}||.||\vec{y}||cos(\theta) = ||\vec{x}||^2 + ||\vec{y}||^2 + 2.x.y$
$=> ||\vec{x}||.||\vec{y}||cos(\theta) = x.y$
$=> cos(\theta) = \frac{x.y}{||\vec{x}||||\vec{y}||}$

We know that: $\vec{x}.\vec{y} = x^T y$, Therefore,
$cos(\theta) = \frac{x^T y}{||\vec{x}||.||\vec{y}||}$

Therefore, the cosine of the angle (cos(x,y)) between two vectors x and y is given by
$cos(x, y) = \frac{x^T y}{||x||||y||}$

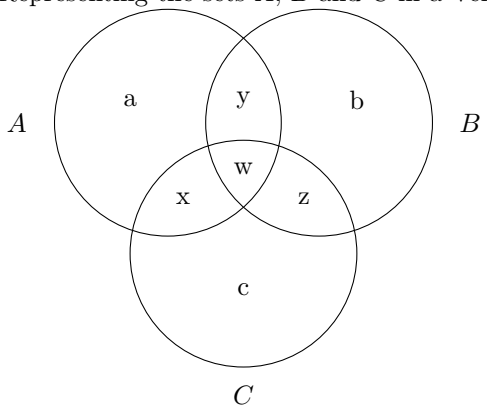
Problem 3

a. $d_1(A, B) = |A - B| + |B - A|$
we have,
$d_1(A, B) = |A - B| + |B - A|$
$=> d_1(A, B) = |A| - |A \cap B| + |B| - |A \cap B|$
$=> d_1(A, B) = |A| + |B| - 2|A \cap B|$ –(1)
For a distance to be a metric the following four properties must be satisfied:

1. Property 1: $d_1(A, B) \geq 0$, for all A and B
   $=> d_1(A, B) = |A| + |B| - 2|A \cap B| \geq 0$, for all A and B
   $=> |A \cup B| - |A \cap B| \geq 0$ –(2)
   This is always true since $|A \cup B| \geq |A \cap B|$, for all A and B
   Therefore, the first property is satisfied.
2. Property 2: $d_1(A, B) = 0$ if $A = B$
   from (2), we have,
   $d_1(A, A) = |A \cup A| - |A \cap A|$
   $=> d_1(A, A)|A| - |A| = 0$
   Therefore, the second property is satisfied
3. Property 3: $d_1(A, B) = d_1(B, A)$, for all A and B
   from (2), we have,
   $d_1(A, B) = |A \cup B| - |A \cap B|$
   $=> d_1(A, B) = |B \cup A| - |B \cap A|$
   $=> d_1(A, B) = d_1(B, A)$
   Therefore, the third property is satisfied

4. Property 4: $d_1(A,C) \leq d_1(A,B) + d_2(B,C)$, for all A, B and C
   from (1), we have,
   $d_1(A,B) = |A| + |B| - 2|A \cap B|$
   Therefore
   $|A| + |C| - 2|A \cap C| \leq |A| + |B| - 2|A \cap B| + |B| + |C| - 2|B \cap c|$
   $=> |A| + 2|B| + |C| - 2|A \cap B| - 2|B \cap c| - |A| - |C| + 2|A \cap C| \geq 0$
   $=> 2|B| + 2|A \cap C| - 2|A \cap B| - 2|B \cap C| \geq 0$
   $=> |B| + |A \cap C| - |A \cap B| - |B \cap C| \geq 0$

   Representing the sets A, B and C in a Venn diagram, we get:



   $=> (b + w + y + z) + (w + x) - (w + y) - (w + z) \geq 0$
   $=> b + x \geq 0$
   This is always true since $b \geq 0$ and $e \geq 0$
   Therefore, the fourth property is satisfied. Therefore $d_1(A,B)$ is a metric.

b. $d_2(A,B) = \dfrac{|A - B| + |B - A|}{|A - B|}$

   We have,
   $d_2(A,B) = \dfrac{|A - B| + |B - A|}{|A \cup B|}$
   $=> d_2(A,B) = \dfrac{|A \cup B| - |A \cap B|}{|A \cup B|}$
   $=> d_2(A,B) = 1 - \dfrac{|A \cap B|}{|A \cup B|}$ –(1)

1. Property 1: $d_2(A,B) \geq 0$, for all A and B
   The second term in (1) can never be greater than 1 The minimum value of $d_2(A,B)$ is obtained when the second term
   $\dfrac{|A \cap B|}{|A \cup B|} = 1$
   This happens when $|A \cap B| = |A \cup B|$, or $A = B$
   then, the minimum value of $d_4(A,B)$ is
   $d_2(min) = d_2(A,A) = 1 - \dfrac{|A \cap A|}{|A \cup A|}$ –(2)
   $=> d_2(min) = d_2(A,A) = 1 - 1 = 0$
   therefore, the minimum value of $d_2(A,B)$ is 0, and $d_2(A,B) \geq 0$
   Therefore, the first property is satisfied
2. Property 2: $d_2(A,B) = 0$, when A = B

when A = B,

$d_2(A, A) = 1 - \frac{|A \cap A|}{|A \cup A|}$

It follows form (2) that $d_2(A, A) = 0$.

Therefore, property 2 is satisfied.

3. Property 3: $d_2(A, B) = d_2(B, A)$, for all A and B

   We have from (1)

   $d_2(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$

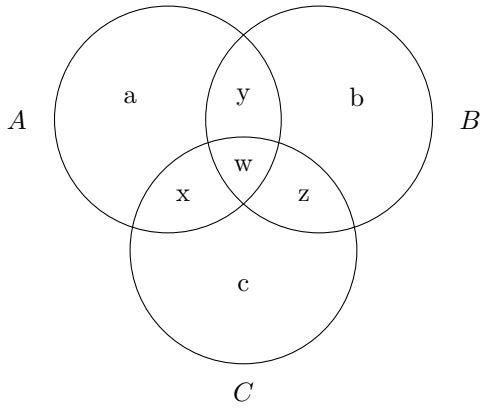
   $=> d_2(A, B) = 1 - \frac{|B \cap A|}{|B \cup A|}$

   $d_2(A, B) = d_2(B, A)$

   Therefore, the third property is satisfied.

4. Property 4: $d_2(A, C) \leq d_2(A, B) + d_2(B, C)$

   Consider the Venn diagram for the sets A B and C as follows:



then,

let $k = a + b + c + x + y + z + w$

then

$d_2(A, B) = 1 - \frac{y + w}{k - c}$

$d_2(B, C) = 1 - \frac{w + z}{k - a}$

$d_2(A, C) = 1 - \frac{x + w}{k - b}$

then

$d_2(A, B) + d_2(B, C) \geq d_2(A, C)$

$=> 1 - \frac{y + w}{k - c} + 1 - \frac{w + z}{k - a} \geq 1 - \frac{x + w}{k - b}$

$=> 1 - \frac{w + z}{k - a} - \frac{y + w}{k - c} \geq -\frac{x + w}{k - b}$

$=> \frac{w + z}{k - a} + \frac{y + w}{k - c} - \frac{x + w}{k - b} \leq 1$

$=> \frac{w + z}{k - a} + \frac{yk - by + wk - bw - (wk + xk - wc - cx)}{(k - b)(k - c)} \leq 1$

$=> (w + z)(K^2 - bk - ck + bc) + (yk - by - bw + xk - wc - cx)(k - a) \leq (k - a)(k - b)(k - c)$

$=> wk^2 - bwk - cwk + wbc + zk^2 - bzk - czk + bzc + yk^2 - byx - bwx - xk^2 + wck + cxk - ayk + aby + abw + axk - awk - acx \leq k^3 - bk^2 - ck^2 + bck - ak^2 + abk + ack - abc$

on simplifying the above inequality we get

$2xk^2 + bwc + bzk + czk + byk + bwk + ayk + cxk - axk + abk + bck + cak - wbc - aby - bcz - abw + awc + acx - abc \geq 0$

$=> kx(2k-a) + bw(k-c-a) + ab(k-y) + bc(k-z) + zc(k-b) bzk + czk + byk + ayk + cxk_a wc + ack \geq 0$

Now, we see that each term on the left hand side of the inequality is non negative.

Therefore, the inequality always holds, and property 4 is satisfied.

Therefore, d_2 satisfies all the properties and is a metric

c. $d_3(A,B) = 1 - \left(\frac{1}{2}.\frac{|A \cap B|}{|A|} + \frac{1}{2}.\frac{|A \cap B|}{|B|}\right)$

1. property 1: $d_3(A,B) \geq 0$, for all A and B
   We know, $|A \cap B| \leq |A|$ and $|A \cap B| \leq |B|$
   the maximum value of the second therm in the d_3 occurs when A = B
   Therefore, the minimum value of d_3 is
   $d_3(min) = d_3(A,A) = 1 - \left(\frac{1}{2}.\frac{|A \cap A|}{|A|} + \frac{1}{2}.\frac{|A \cap A|}{|B|}\right)$
   $=> d_3(min) = 1 = 1 = 0$
   –(1)
   Therefore, the first property is satisfied.
2. Property 2: $d_3(A,B) = 0$, if A = B
   It follows from (1) that $d_3(min) = d_3(A,A) = 0$
   Therefore the second property is satisfied.
3. Property 3: $d_3(A,B) = d_3(B,A)$, for all A and B
   we have, $d_3(A,B) = 1 - \left(\frac{1}{2}.\frac{|A \cap B|}{|A|} + \frac{1}{2}.\frac{|A \cap B|}{|B|}\right)$
   $=> d_3(A,B) = 1 - \left(\frac{1}{2}.\frac{|B \cap A|}{|A|} + \frac{1}{2}.\frac{|B \cap A|}{|B|}\right)$
   $=> d_3(A,B) = 1 - \left(\frac{1}{2}.\frac{|B \cap A|}{|B|} + \frac{1}{2}.\frac{|B \cap A|}{|A|}\right)$
   $=> d_3(A,B) = d_3(B,A)$
   Therefore the third property is satisfied
4. Property 4: $d_3(A,C) \leq d_3(A,B) + d_3(B,C)$
   This can be shown to be not true using a counter example.

   Consider sets A, B and C such that.
   $A = \{a,b,c\}, C = \{d,e,f\}, B = \{a,b,c,d,e,f\}$, then
   $|A| = 3$,
   $|B| = 6$, $|C| = 3$,
   $|A \cap B| = 3$,
   $|B \cap C| = 3$,
   $|A \cap C| = 0$

   Then,
   $d_3(A,B) = 1 - \left(\frac{1}{2}.\frac{|A \cap B|}{|A|} + \frac{1}{2}.\frac{|A \cap B|}{|B|}\right)$

Problem 3 continued on next page...

$$=> d_3(A,B) = 1 - \left(\frac{1}{2}.\frac{3}{3} + \frac{1}{2}.\frac{3}{6}\right)$$
$$d_3(A,B) = 1 - \frac{3}{4} = \frac{1}{4}$$

similarly, $d_3(B,C) = 1/4$
and

$$d_3(A,C) = 1 - \left(\frac{1}{2}.\frac{|A \cap C|}{|A|} + \frac{1}{2}.\frac{|A \cap C|}{|C|}\right)$$
$$=> d_3(A,C) = 1 - \left(\frac{1}{2}.\frac{0}{|A|} + \frac{1}{2}.\frac{0}{|B|}\right)$$
$$=> d_3(A,C) = 1 - 0 = 1$$

Therefore,
$$d_3(A,B) + d_3(B,C) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \ngeq d_3(A,C) = 1$$

Therefore, d_3 does not satisfy the fourth property and is not a metric

d. $$d_4(A,B) = 1 - \left(\frac{1}{2}.\frac{|A|}{|A \cap B|} + \frac{1}{2}.\frac{|B|}{|A \cap B|}\right)^{-1}$$

1. property 4: $d_4(A,C) \leq d_4(A,B) + d_4(B,C)$
   It can be shown that d_4 does not satisfy this property.
   Consider sets A, B and C such that.
   $A = \{a,b,c\}, C = \{d,e,f\}, B = \{a,b,c,d,e,f\}$, then
   $|A| = 3$,
   $|B| = 6$, $|C| = 3$,
   $|A \cap B| = 3$,
   $|B \cap C| = 3$,
   $|A \cap C| = 0$

   Then
   $$d_4(A,B) = 1 - \left(\frac{1}{2}.\frac{|A|}{|A \cap B|} + \frac{1}{2}.\frac{|B|}{|A \cap B|}\right)^{-1}$$
   $$=> d_4(A,B) = 1 - \left(\frac{1}{2}.\frac{3}{3} + \frac{1}{2}.\frac{6}{3}\right)^{-1}$$
   $$=> d_4(A,B) = 1 - \left(\frac{1}{2} + 1\right)^{-1}$$
   $$=> d_4(A,B) = 1 - \left(\frac{3}{2}\right)^{-1} => d_4(A,B) = 1 - \frac{2}{3} = 1/3$$
   Similarly
   $$d_4(B,C) = \frac{1}{3}$$
   and
   $$d_4(A,C) = 1 - \left(\frac{1}{2}.\frac{|A|}{|A \cap C|} + \frac{1}{2}.\frac{|C|}{|A \cap C|}\right)^{-1}$$
   $$=> d_4(A,C) = 1 - \left(\frac{1}{2}.\frac{|A| + |C|}{|A \cap C|}\right)^{-1}$$
   $$=> d_4(A,C) = 1 - 2\frac{|A \cap C|}{|A| + |C|}$$

$=> d_4(A,C) = 1 - 2\frac{0}{|A|+|C|} = 1$

Then

$d_4(A,B) + d_4(B,C) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \not\geq d_4(A,C) = 1$

Therefore the fourth property is not satisfied, and d_4 is not a metric

Problem 4

$$d(x,y) = \left(\left(\sum_{i:x_i > y_i} (x_i - y_i)^p \right) + \left(\sum_{i:y_i > x_i} (y_i - x_i)^p \right) \right)^{1/p}$$

1. Property 1: $d(x,y) \geq 0$, for all x and y

   We see that for any vectors x and y, we are individually take the difference between the elements such that both the summation terms are positive. These terms are raised to a power of p and summed and raised to a power $\frac{1}{p}$.

   Therefore, $d(x,y) \geq 0$ for all vectors x and y

   Therefore, the first property is satisfied

2. Property 2: $d(x,y) = 0$ if x = y,

   When x = y, $d(x,y)$ becomes zero since we are only the difference $x_i - y_i = 0$

   Therefore the second property is satisfied

3. Property 3: $d(x,y) = d(y,x)$

   We have,

   $$d(x,y) = \left(\left(\sum_{i:x_i > y_i} (x_i - y_i)^p \right) + \left(\sum_{i:y_i > x_i} (y_i - x_i)^p \right) \right)^{1/p}$$

   $$=> d(x,y) = \left(\left(\sum_{i:y_i > y_i} (y_i - x_i)^p \right) + \left(\sum_{i:x_i > y_i} (x_i - y_i)^p \right) \right)^{1/p}$$

   $=> d(x,y) = d(y,x)$

   Therefore, the third property is satisfied

   The three properties above hold for both the cases, when $0 < p < 1$ and when $p > 1$.

4. Property 4: $d(x,y) + d(y,z) \geq d(x,z)$

Problem 5

a. The graphs for each of the distance metrics for data from a normal distribution are given below:

b. The graphs for each of the distance metrics for data from a normal distribution are given below:

c. We see from the plots that the plot of $r(k)$ is sub linear values. Therefore, the ratio of the spread of distances to the minimum distance decrease exponentially with increase in k. As the number of attributes in the data increases, the points generated in the K-dimensional space are spread far apart. Therefore, the maximum distance between two points ($d_max(k)$) becomes almost same as the minimum distance between two points ($d_min(k)$). Thus the ratio of the spread to the minimum distance ($r(k)$) decreases

Problem 5 continued on next page...

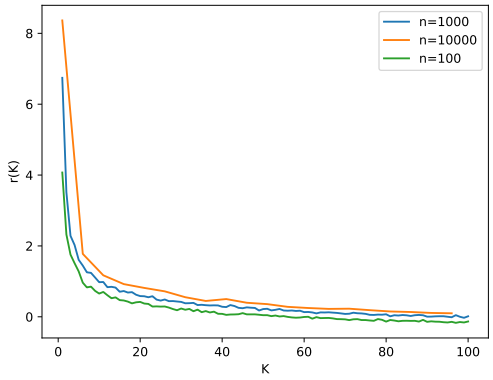


Figure 1: Plot of r(k) vs k for normal data points and euclidean distance

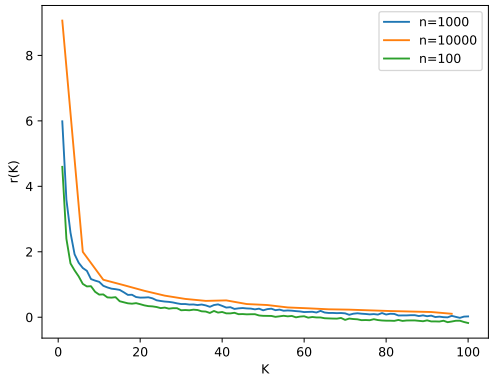


Figure 2: Plot of r(k) vs k for normal data points and problem minkowski distance

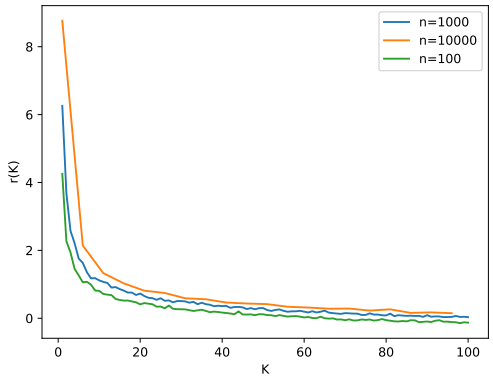


Figure 3: Plot of r(k) vs k for normal data points and problem 4 distance

rapidly. To avoid this we need to increase the number of data points exponentially as the number of attributes increase.

We also see that, the $r(k)$ values for a distance metric for the same values of n and k, are greater for data

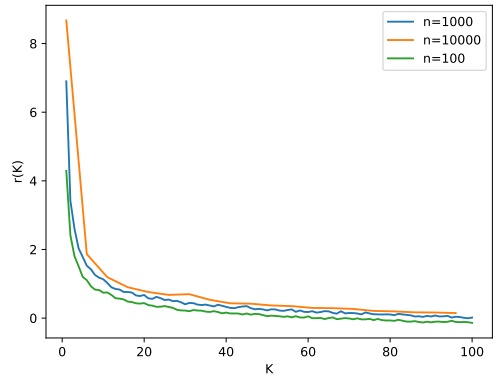


Figure 4: Plot of r(k) vs k for normal data points and problem cityblock distance

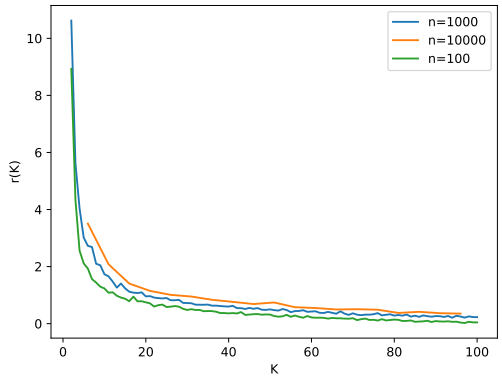


Figure 5: Plot of r(k) vs k for normal data points and problem cosine distance

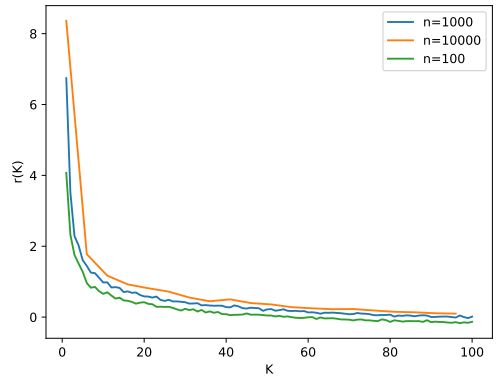


Figure 6: Plot of r(k) vs k for uniform data points and euclidean distance

generated from a normal distribution then from the uniform distribution. This is because in the case of uniform distribution the data points generated are spread across randomly, whereas points from a normal distribution are centered across the mean. Thus the minimum distance between two points $d_min(k)$ is

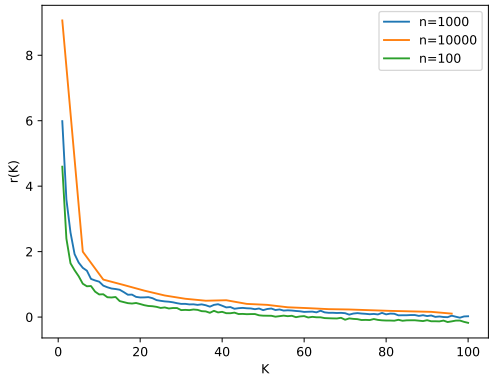


Figure 7: Plot of r(k) vs k for uniform data points and problem minkowski distance

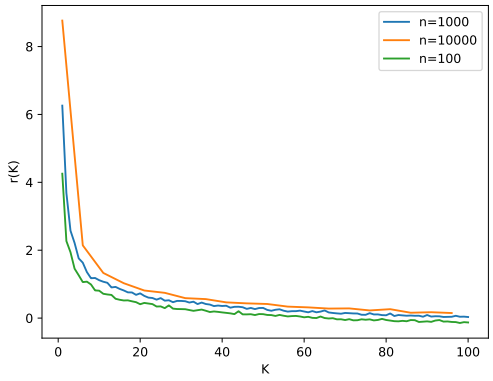


Figure 8: Plot of r(k) vs k for uniform data points and problem 4 distance

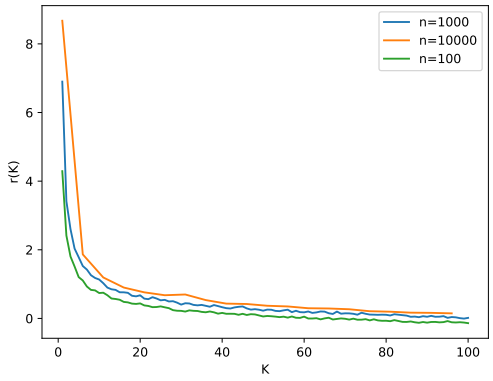


Figure 9: Plot of r(k) vs k for uniform data points and problem cityblock distance

still smaller than the maximum distance $d_max(k)$ in the case of the normal distribution. This can be seen from the plots that the value of $r(k)$ is slightly higher for the normally distributed data.

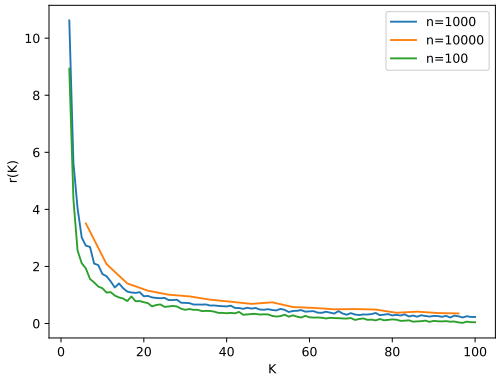


Figure 10: Plot of r(k) vs k for uniform data points and problem cosine distance

Problem 6

a. The code can be found in the $/code/q5/decision_tree.py$ file.

To run the code using GINI index execute:

$pythondecision\_tree.pyGINI < file_name >$

To run the code using information gain execute:

$pythondecision\_tree.pyENTROPY < file_name >$

The code stops splitting the data further if either all the elements in a node belong to the same class or

b. The code randomly splits the data into two halves for training and testing and trains the data on one half and measures the accuracy on the second half. This process is repeated 5 times, and the code returns the average accuracy after 5 iterations. The following table represents the performance on various data sets for both GINI and information gain metrics.

The data sets used are, UCI wine dataset, UCI Iris dataset and UCI Breast Cancer(Diagnostic) dataset

Table 1: Performance evaluation of decision tree

Dataset	**Average Accuracy range (GINI)**	**Average Accuracy range (Information Gain)**
$processedIrisData$	$0.86 - 0.92$	$0.87 - 0.91$
$processedWineData$	$0.88 - 0.92$	$0.86 - 0.90$
$BreastCancerData$	$0.89 - 0.92$	$0.09 - 0.91$

the table represents the range of average accuracy after training and testing on 5 random splits. The average accuracy of 5 runs varies as we are using only half of the randomly selected points to train the data and test the performance on the other half. Since we do not have a large amount of data, this means that on some splits, most points it the training data set can belong to one class, skewing the distribution of class in the entire data , and leading to lower accuracy.

c. We can see from the above table that the the performance of both the GINI index and information gain is almost the same, and that there is not much difference in practice between the two measures of impurity.

References

- https://archive.ics.uci.edu/ml/datasets/wine
- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- https://archive.ics.uci.edu/ml/datasets/iris
- https://proofwiki.org/wiki/Law_of_Cosines
- http://www.p-value.info/2013/02/when-tfidf-and-cosine-similarity-fail.html
- https://en.ryte.com/wiki/TF*IDF
- https://en.wikipedia.org/wiki/Tf
- https://mathoverflow.net/questions/18084/is-the-jaccard-distance-a-distance
- https://www.reddit.com/r/learnmath/comments/2v916l/set_theory_proof_of_triangle_inequality_for/
- https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf
- http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Discussed with classmates Dheeraj (dhsingh) and Pulkit (maloop)