Homework Assignment 1

Rishabh Sinha (rsinha2)
CSC 522 Auto Learn Data Analysis

Q1) (10 points) Answer the following questions
(a) How is data mining different from statistical analysis?

Data Mining is different from statistical Analysis as Statistical Analysis involves making statistical inferences from existing data, While Data mining involves finding previously unknown information or patterns from a dataset.

(b) Differentiate between noise and outliers

A noise is a data value, that is essentially not true, but has crept into the dataset in the process of data collection.
While an outlier is a valid data point, whose value is a lot different from other points in the given dataset. For example given the mean 10 and standard deviation of 3. A point which is more that 2 standard deviations away from the mean is an outlier say value of 17 for the above example.

(c) List 5 unique data characteristics that makes traditional data mining techniques less effective.

Five characteristics of data that made traditional data mining techniques less effective are:
1) Scalability : Data mining algorithms must be scalable so that they can handle massive datasets, that exist today, which may range up to several petabytes.
2) High Dimensionality of Data: With progress in fields such as bio-informatics, the dimensionality of data has radically increased. It is not uncommon to find data sets, having more that a thousand attributes. In old data analysis algorithms, computational complexity increases rapidly with increase in dimensionality of data.
3)Heterogeneous and Complex data: Emergence of complex types of data such as spatio-temporal, web data, DNA data with 3D structure, etc. make traditional data mining techniques obsolete.
4) Data ownership and distribution: Sometimes data maybe not be owned by one organization, but may be distributed among multiple entities. This might lead to issues of data consolidation and Security.
5)Non Traditional Analysis: Traditional approach is based on hypothesize and test paradigm which is very labor intensive.

Q2)(10 points) It is important to define or select similarity measures in data analysis. However, there is no commonly accepted similarity measure. Different similarity measures may come up with different results. And the similarities between data points may or may not change after data transformation. In this question, we want to explore these problems. Suppose we have the following two-dimensional data set:

Data points	x1	x2
p1	0.3	0.8
p2	0.7	0.4
p3	1	0.1

p4	-0.1	-0.3
p5	0.9	0.8

(a) Consider the data as two-dimension data points. Give a new data point, q1 = (0.4,0.2) as a query, rank the data points based on the similarity with the query point using (1) Euclidean distance (2) cosine similarity [round your similarity to three decimal places]

Euclidean Distance measure with 2 attributes is $((x_1-x_2)^2+(y_1-y_2)^2)^{1/2}$ for point (x_1,y_1) and (x_2,y_2).
According to the Euclidean Distance measure, The above points rank as:-
$d(p1,q1) = ((0.4-0.3)^2+(0.2-0.8)^2)^{1/2} = (0.01+0.36)^{1/2} = 0.608$
$d(p2,q1) = ((0.4-0.7)^2+(0.2-0.4)^2)^{1/2} = (0.09+0.04)^{1/2} = 0.361$
$d(p3,q1) = ((0.4-1)^2+(0.2-0.1)^2)^{1/2} = (0.36+0.01)^{1/2} = 0.608$
$d(p4,q1) = ((0.4+0.1)^2+(0.2+0.3)^2)^{1/2} = (0.25+0.25)^{1/2} = 0.707$
$d(p5,q1) = ((0.4-0.9)^2+(0.2-0.8)^2)^{1/2} = (0.25+0.36)^{1/2} = 0.781$

Thus the ranking in ascending order of similarity is :
p5, p4, p3=p1, p2

Cosine similarity for data points is defined as :A.B/|A||B|, where A and B are data vectors.
According to Cosine Similarity measure, The above points rank as:-
$d(p1,q1) = ((0.4*0.3) + (0.2*0.8))/ ((0.4^2+0.2^2)^{1/2}(0.3^2 + 0.8^2)^{1/2}) = 0.733$
$d(p2,q1) = ((0.4*0.7) + (0.2*0.4))/ ((0.4^2+0.2^2)^{1/2}(0.7^2 + 0.4^2)^{1/2}) = 0.36/0.36055512755 = 0.998$
$d(p3,q1) = ((0.4*1) + (0.2*0.1))/ ((0.4^2+0.2^2)^{1/2}(1^2 + 0.1^2)^{1/2}) = .42/.4494441011 = 0.934$
$d(p4,q1) = ((0.4*-0.1) + (0.2*-0.3))/ ((0.4^2+0.2^2)^{1/2}(-0.1^2 + -0.3^2)^{1/2}) = -.1/.14142135624 = -0.707$
$d(p5,q1) = ((0.4*0.9) + (0.2*0.8))/ ((0.4^2+0.2^2)^{1/2}(0.9^2 + 0.8^2)^{1/2}) = .52/0.538516481 = 0.966$

Thus the ranking in ascending order of similarity is :
p4, p1, p3, p5, p2

Thus the order totally changes, p5 is least similar to the query according to euclidean distance similarity while it is very similar according to cosine.

(b) Transform each value in your data set and the query point using the sigmoid function:
y = 1/ (1 + e^(-x))
The new transformed data points are:

Data points	x1	x2
p1	0.574	0.690
p2	0.668	0.599
p3	0.731	0.525
p4	0.475	0.426
p5	0.711	0.690
queryPt	0.599	0.550

And re-rank the data points based on the similarity with the query point using (1) Euclidean distance
According to the Euclidean Distance measure, The above points rank as:-

$d(p1,q1) = ((0.599-0.574)^2 +(0.550 -0.690)^2)^{1/2} = 0.142$

$d(p2,q1) = ((0.599 -0.668)^2 +(0.550 -0.599)^2)^{1/2} = 0.085$

$d(p3,q1) = ((0.599 -0.731)^2 +(0.550 -0.525)^2)^{1/2} = 0.134$

$d(p4,q1) = ((0.599 -0.475)^2 +(0.550 -0.426)^2)^{1/2} = 0.175$

$d(p5,q1) = ((0.599 -0.711)^2 +(0.550 -0.690)^2)^{1/2} = 0.179$

Thus the ranking in ascending order of similarity is :
p5, p4, p1, p3, p2
Due to the transformation, the euclidean distance, that was equal for p3 and p1 from the query, changes and now p3 is closer to query than p1.

(2) cosine similarity [round your similarity to three decimal places]

$d(p1,q1) = ((0.599*0.574) + (0.550*0.690))/ ((0.599^2+0.550^2)^{1/2}(0.574^2 + 0.690^2)^{1/2}) = 0.991$

$d(p2,q1) = ((0.599*0.668) + (0.550*0.599))/ ((0.599^2+0.550^2)^{1/2}(0.668^2 + 0.599^2)^{1/2}) = 1$

$d(p3,q1) = ((0.599*0.731) + (0.550*0.525))/ ((0.599^2+0.550^2)^{1/2}(0.731^2 + 0.525^2)^{1/2}) = .0.993$

$d(p4,q1) = ((0.599*0.475) + (0.550*0.426))/ ((0.599^2+0.550^2)^{1/2}(0.475^2 + 0.426^2)^{1/2}) = 1$

$d(p5,q1) = ((0.599*0.711) + (0.550*0.690))/ ((0.599^2+0.550^2)^{1/2}(0.711^2 + 0.690^2)^{1/2}) = 1$

The Rank of the above points according to cosine similarities in ascending order is:
p1,p3,p5,p2,p4
Here due to transformation to sigmoid function the value of the points hugely changes leading to a huge increase in cosine similarity of all the above points, to the query point. The cosine similarity of the points p2, p5 and p4 are virtually 1 to the query point that is they, are virtually identical

3). (10 points) What is a metric space? What are the conditions a function must satisfy to be called a metric? Explain in full generality.

A metric space is a set X that has a notion of a distance d(x,y) between every pair of points x and y in the space. A metric space may have n dimensions, An example of a metric space is the euclidean space. A function to be called a metric must satisfy the following properties:-

1. d(p, q) >= 0 for all p and q and d(p, q) = 0
only if p = q. (Positive definiteness)
This means that the distance metric in a metric space between any two points is greater than equal to 0, and can only be zero if the two points are identical.

2. d(p, q) = d(q, p) for all p and q. (Symmetry)
This means that the distance remains same whether we measure it from a to b or from b to a.

3. d(p, r) <= d(p, q) + d(q, r) for all points p, q,
and r. (Triangle Inequality).
This means that for any three distinct points in a metric space, the distance between 2 points is less than, that between, the path that goes through the third point.

where d(p, q) is the distance (degree of dissimilarity) between
points (data objects), p and q.

A distance that satisfies these properties is called a metric.

4. (10 points) (a) For each of the following, indicate whether the variable is binary/discrete/continuous, nominal/ordinal/interval/ratio.

Variable	Binary/discrete/continuous	Nominal/Ordinal/interval/ratio	Assumptions(if Any)
Brightness in terms of dark and light	Binary	Nominal	(Assuming only 2 states exist, Dark and Light, and One cant be considered more or better than the other.)
Temperature in Fahrenheit	continuous	Interval	
Barcodes	Discrete	Nominal	
Student Grades	Discrete	Ordinal	
Weight of an Object	Continuous	Ratio	