Homework01_solutions_Elias

September 2, 2019

REI502M - Introduction to Data Mining

Homework 1

Student: Elías Snorrason

Problem 1

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Some notes

For a discrete/continous attribute type, the set of it's possible values can be be either finite or infinite, respectively. A binary attribute type is a special case of a discrete attribute with two possible values.

A rough distinction between nominal and ordinal attribute types is that the former is invariant under permutation of values, while the latter doesn't (has a preferred order).

A key difference between interval- and ratio attribute levels is that a set with a ratio attribute only allows for rescaling of values, not shifting them. An interval attribute allows both value shifting and rescaling (typically both)

1.A Time in terms of AM or PM.

Binary (AM *OR* PM), Qualitative and Ordinal (AM/PM <-> 00:00-23:59).

1.B Brightness as measured by a light meter.

Continuous (Lux meters represent measurements as floating-point variables), Quantitative and Interval (no zero-point).

1.C Brightness as measured by peoples judgments.

Discrete (people tend to think logarithmically), Qualitative and Ordinal (logarithms are monotonic)

1.D Angles as measured in degrees between 0 and 360.

Continous (floating-point...), Quantitative and Ratio (only need to rescale for radians/gradians in the given range)

1.E Bronze, Silver, and Gold medals as awarded at the Olympics.

Discrete (finite set of medal types), Qualitative and Ordinal (ranking performance of athletes)

1.F Height above sea level.

Continuous (length measurements are continuous), Qualitative and Interval (since sea-level is not necessarily an absolute reference height, e.g. shift zero-height to Challenger Deep or Mt. Everest)

1.G Number of patients in a hospital.

Discrete (countable set of patients), Quantitative and Ratio (lower limit is zero patients)

1.H ISBN numbers for books. (Look up the format on the Web.)

Discrete (finite set of integers), Qualitative and Nominal (no preferred ranking as the set integers are unique)

1.I Ability to pass light in terms of the following values: opaque, translucent,transparent.

Discrete (3 possible values), Qualitative and ordinal (No clear cutoff-values for opacity. However: transparent < translucent < opaque).

Problem 2

Distinguish between noise and outliers. Be sure to consider the following questions.

2.A Is noise ever interesting or desirable? Outliers?

Noise in attrbutes is undesirable by default, as it distorts the original attribute values. Outliers can potentially be legitimate objects of data (or values), i.e. identifying them can be the main objective of some data mining tasks. Thus, outliers can potentially be interesting/desirable, but noise is not (by definition).

2.B Can noise objects be outliers?

Noise in attribute values can make the data look more randomized or unusual. Thus, it is possible that some instances in noisy data will appear as outliers.

2.C Are noise objects always outliers?

Noisy data can appear as normal data. So noise objects are not always outliers.

2.D Are outliers always noise objects?

Outliers can be legitimate data objects that appear to not belong in the data set. Those outliers would typically not classify as noise objects.

2.E Can noise make a typical value into an unusual one, or vice versa?

As discussed in previous parts, the source of noise in data can randomly make some values appear as unusual. Or some outliers as typical data objects.

Problem 3

3.A

For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

$x = 0101010001$
$y = 0100011000$

Solution 3.A

The Hamming distance L_1 between two binary vectors can be computed with

$$L1 = f_{10} + f_{01} = f_{10\,\mathrm{OR}\,01}$$

Running x and y through an XOR gate (element-wise) and summing up the terms in the resulting vector is equivalent to computing the Hamming distance.

```
In [1]: x = [0 1 0 1 0 1 0 0 0 1]
        y = [0 1 0 0 0 1 1 0 0 0]

        # Use a XOR gate, element-wise, on x and y
        z = xor.(x,y)
        println("z = $z")
        println("Hamming distance = $(sum(z))")
```

```
z = [0 0 0 1 0 0 1 0 0 1]
Hamming distance = 3
```

The Jaccard similarity coefficient J can be calculated with

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{f_{11}}{n - f_{00}}$$

where n is the number of bits.

Let's use the former expression.

f_{11} can be computed with summing the result of running x and y through an AND gate. $f_{01} + f_{10} + f_{11}$ uses an OR gate.

```
In [2]: # Define function that calculates J
        function jaccard(x,y)
            # Use an AND gate, then sum.
            f11 = sum(x .& y)
            # Use an OR gate, then sum.
            denom = sum(x .| y)
            return f11/denom
        end

        J = jaccard(x,y)
        println("J = $J")
```

```
J = 0.4
```

3.B

Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

Solution 3.B

The Hamming distance explicitly ignores genes that match, so the Jaccard measure is more appropriate for the task.

Problem 4

For the following vectors, x and y, calculate the indicated similarity or distance measures.

4.A

$x = (1,1,1,1), y = (2,2,2,2)$ cosine, correlation, Euclidean

Solution 4.A

Cosine similarity: $\cos(x,y) = \frac{x \cdot y}{\sqrt{(x \cdot x)(y \cdot y)}}$

Correlation: $\rho_{x,y} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

Euclidean distance: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

In [3]:
```
# Define custom method for cos(x,y)
function cossimilarity(x,y)
    A = sum(x .* y)
    B = sqrt(sum(x .*x) * sum(y .* y))
    return A / B
end


# Define custom methods for x
function mean(x)
    n = length(x)
    return sum(x) / n
end

# Define custom method for corr(x,y), note special case for zeros in denominator
function correlation(x,y)
    X = x .- mean(x)
    Y = y .- mean(y)

    A = sum(X .* Y)
    B = sqrt(sum(X .* X))
    C = sqrt(sum(Y .* Y))
    if B == 0 || C == 0
        return NaN
    else
        return A/(B*C)
    end
end

# Define custom method for d(x,y)
function euclidean(x,y)
    Z = (x .- y)
    return sqrt(sum(Z .* Z))
end
```

Out[3]:
```
euclidean (generic function with 1 method)
```

```
In [4]: # Initialize data
        x = [1,1,1,1]
        y = [2,2,2,2]

        #Print statements
        println("x = $x")
        println("y = $y")
        println("cos(x,y) = $(cossimilarity(x,y))")
        println("corr(x,y) = $(correlation(x,y))")
        println("d(x,y) = $(euclidean(x,y))   (Euclidean distance)")
```

```
x = [1, 1, 1, 1]
y = [2, 2, 2, 2]
cos(x,y) = 1.0
corr(x,y) = NaN
d(x,y) = 2.0   (Euclidean distance)
```

4.B

$x = (0,1,0,1), y = (1,0,1,0)$ cosine, correlation, Euclidean, Jaccard

Solution 4.B

Use previous function definitions, including `jaccard(x,y)`.

```
In [5]: x = [0,1,0,1]
        y = [1,0,1,0]

        println("x = $x")
        println("y = $y")
        println("cos(x,y) = $(cossimilarity(x,y))")
        println("corr(x,y) = $(correlation(x,y))")
        println("d(x,y) = $(euclidean(x,y))   (Euclidean distance)")
        println("J(x,y) = $(jaccard(x,y))")
```

```
x = [0, 1, 0, 1]
y = [1, 0, 1, 0]
cos(x,y) = 0.0
corr(x,y) = -1.0
d(x,y) = 2.0   (Euclidean distance)
J(x,y) = 0.0
```

Problem 5

Proximity is typically defined between a pair of objects.

5.A

Define two ways in which you might define the proximity among a group of objects.

Solution 5.A

We could define a geometric center for the entire group of objects x_i, with N objects, each of dimension n.

$$x_0 \equiv \frac{\sum_{i=1}^{N} x_i}{N}$$

The group-proximity might be defined as the greatest Euclidean distance of all the points from x_0, i.e.

$$d_{\text{group}} = \max_j \left(\sum_i^n (x_{j,i} - x_{0,i})^2 \right)$$

or computing some average of these distances:

$$d_{\text{group}} = \frac{1}{N} \sum_j^N \left(\sum_i^n (x_{j,i} - x_{0,i})^2 \right)$$

Another way to define proximity might be to make a multi-dimensional array of all possible pairwise proximity measures. Then you could determine the extrema of those pairwise proximity measures (for minimum of similarity, maximum of dissimilarity).

5.B

How might you define the distance between two sets of points in Euclidean space?

Solution 5.B

As mentioned in the problem text, the distance is typically defined between a pair of objects(points). This means reducing each set to a single point in the same space. A geometric center/centroid is a good candidate. The distance would be measured between the two resulting centroids.

5.C

How might you define the proximity between two sets of data objects? (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.)

Solution 5.C

An estimate for the upper bound on the proximity measure can be made by choosing the maximal proximity measure between an object in set 'A' with an object in set 'B' (pair of objects with minimal distance). Alternatively, minimizing the pairwise proximity measure of two objects gives a lower bound for the proximity (pair of objects with maximal distance). Similarily to the previous part, averaging over all possible pairwise proximity measures might be a more reliable definition of such a distance.