Metric Spaces, Curse of dimensionality, Taxonomic Type of data, Data Cleaning, Noise vs Outliers, Distance metrics

Shivani Pal

May 8, 2023

Problem 1

The following problems have to do with metrics. In each case, prove or disprove the distance is a metric ($\mathbb{R}$ is the set of reals, and $\|X\|$ is the size of a finite set X.)

(a) Let $X \subset \mathbb{R}^n$ for positive integer $n > 0$. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x, y) = max\{|x_i - y_i|\}, \forall i, 1 \leq i \leq n.$$

Ans. If d(x, y) is the distance between two points, x and y, then the following properties hold.
1. Positivity
(a) $d(x, x) \geq 0$ for all x and y
(b) d(x,y)= 0 only if x=y
These properties are satisfied since maximum of absolute is always non-negative for the given constrains.
2. Similarity
d(x,y) = d(y,x) for all x and y
This property is satisfied since $max\{|x_i - y_i|\} = max\{|y_i - x_i|\} \forall i, 1 \leq i \leq n$
3. Triangle Inequality
d(x,z) ≤ d(x,y) + d(y,z) for all points x, y, and z.
$d(x, y) = max\{|x_i - y_i|\}$
$d(y, z) = max\{|y_i - z_i|\}$
$d(x, z) = max\{|x_i - z_i|\}$
We see, d(x,z) ≤ d(x,y) + d(y,z) for all points x, y, and z.

Thus this property is satified.

(b) Let $c : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ be defined as

$$c(x, y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{o.w.} \end{cases}$$

Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x, y) = \sum_{i}^{n} \frac{c(x_i, y_i)}{i}, \forall i \ 1 \leq i \leq n$$

Ans.

1. Positivity
(a) $d(x, x) \geq 0$ for all x and y
(b) d(x,y)= 0 only if x=y
These properties are satisfied d(x,y) alwasys positive since i value in constraint is positive and value of c(x,y) is non-negative.
2. Similarity
d(x,y) = d(y,x) for all x and y
This property is satisfied since the value of c(x,y) will be 0 or 1 irrespective of the order since c(x,y) = c(y,x)
3. Triangle Inequality
d(x,z) ≤ d(x,y) + d(y,z) for all points x, y, and z.
$d(x, y) = \sum_{i}^{n} \frac{c(x_i, y_i)}{i}$

Problem 1 continued on next page...

$$d(y,z) = \sum_{i}^{n} \frac{c(y_i, z_i)}{i}$$

$$d(x,z) = \sum_{i}^{n} \frac{c(x_i, z_i)}{i}$$

We see, d(x,z) $\leq$ d(x,y) + d(y,z) for all points x, y, and z.

Thus this property is satified.

(c) Suppose $d_0,\ d_1$ are metrics.

i. $d_0 \times d_1$

ii. $(d_0 + d_1)/(d_0 d_1)$

iii. $max\{d_0,\ d_1\}$

iv. Let X be a finite set. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as $d(x,y) = \frac{||x \cap y||}{||x \cup y|| + 1}$

Problem 2

For the following data, give the best taxonomic type (interval, ratio, nominal, ordinal):

1. A section of highway on a map. - **Nominal**
   They are names,and can be only distinguished one from another.
2. The value of a stock. - **Ratio**
   Monetary quantitiy, differences and ratios both are valid
3. The weight of a person. - **Ratio**
   Differences and ratios both are valid
4. Marital status. - **Nominal**
   Married/Unmarried, can be only distinguished one from another.
5. Visiting United Airlines (https://www.united.com) the seating is: Economony, Economy plus, and United Business. - **Ordinal**
   United Business is better than Economy plus and Economy plus is better than Economy, hence showing order.

Problem 3

You are datamining with a column that has physical addresses in some city with the same zipcode. For example,

```
55 WEST CIR
2131 South Creek Road
Apt. #1 Fountain Park
1114 Rosewood Cir
1114 Rosewood Ct.
1114 Rosewood Drive
```

What structure would you create to mine these? What questions do you think you should be able to answer?
Ans. The physical address column can be seen as a multidimensional hierarchical data. We can create a "data cube" structure to mine these. Data cube allows us to take a multidimensional viewpoint of data and aggregate data in various ways. We can divide our physical address column into 3 dimensions: Street, Block and Apt.
It can solve multiple questions by calculating various aggregates on the multidimensional data. We can do operations like:

1. Dimentionality reduction/ Aggregation: even though street, block and apt are atomic on their own, they are together identified as an address. We can Aggregate all blocks on a particular street. Or aggregate all apt_nos on a particular block.

2. Slicing and Dicing: We can splice out data based on a particular value of a street or dice the data based on a range of block values.

3. Roll up/Drill Down: We can aggregate cells within a dimension. Even though street, blocks and apt seem to be atomic values, but sometimes they can be drilled down further. Eg Street can have East, West, North and South. We have an option to roll-ip or dill-down based on these subcategories.

Problem 4

For this problem you will be using a data set with total 81 attributes describing every aspect of residential homes of Ames, Iowa. You can download the data from here [link]. The downloadable file already comes with the corresponding names of the attributes. Also a document describing the data is available here [link].

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?
The data set is a housing data describing every aspect of residential homes of Ames, Iowa. It's a regression problem statement wherein SalePrice is the target variable. This prediction can be done by applying regression algorithm. The attributes includse range,inetrval, ordinal and nominal values. Before applying the modelling algorithm, there are multiple missing values to handle as well.

R/Python Code

Using R/Python, show code that answers the following questions:

1. How many entries are in the data set? Write the R or Python code in the box below.
   Ans. 1460

R/Python script

```
shape =  df.shape
print("Entries in data: ", shape[0])
```

2. How many unknown or missing data are in the data set? Write the R or Python code in the box below.
   Ans. 6965

R/Python script

```
# to find null values in each attribute
# print (df.isna().sum())
# to find null values in the whole data set
print ("Total Null values: ",df.isnull().sum().sum())
```

3. Find 10 attributes influencing the target attribute `SalePrice`. Use coherent plotting methods to describe and discuss their relation with `SalePrice`. Place images of these plots into the document. Write the R or Python code in the box below.
Ans. **['YearRemodAdd', 'YearBuilt', 'TotRmsAbvGrd', 'FullBath', '1stFlrSF', 'TotalBsmtSF', 'GarageArea', 'GarageCars', 'GrLivArea', 'OverallQual']**

R/Python script

```
# Correlation between Saleprice and other attributes
corr = df[df.columns[1:]].corr()['SalePrice']

#top 10 attributes influencing target
topten = corr.sort_values().to_frame().tail(11)
col_corr_with_target = topten.index.to_list()[:-1]
print(col_corr_with_target)

for i in col_corr_with_target:
    plt.scatter(x = df[i], y = df['SalePrice'])
    plt.xlabel(i)
    plt.ylabel('SalePrice')
    plt.title(i+" vs "+"SalePrice")
    name = "Scatter_"+i+"_vs_"+"SalePrice"
    plt.savefig(name)
    plt.show()
```

Discussion of Findings

And. Findings and Observation:

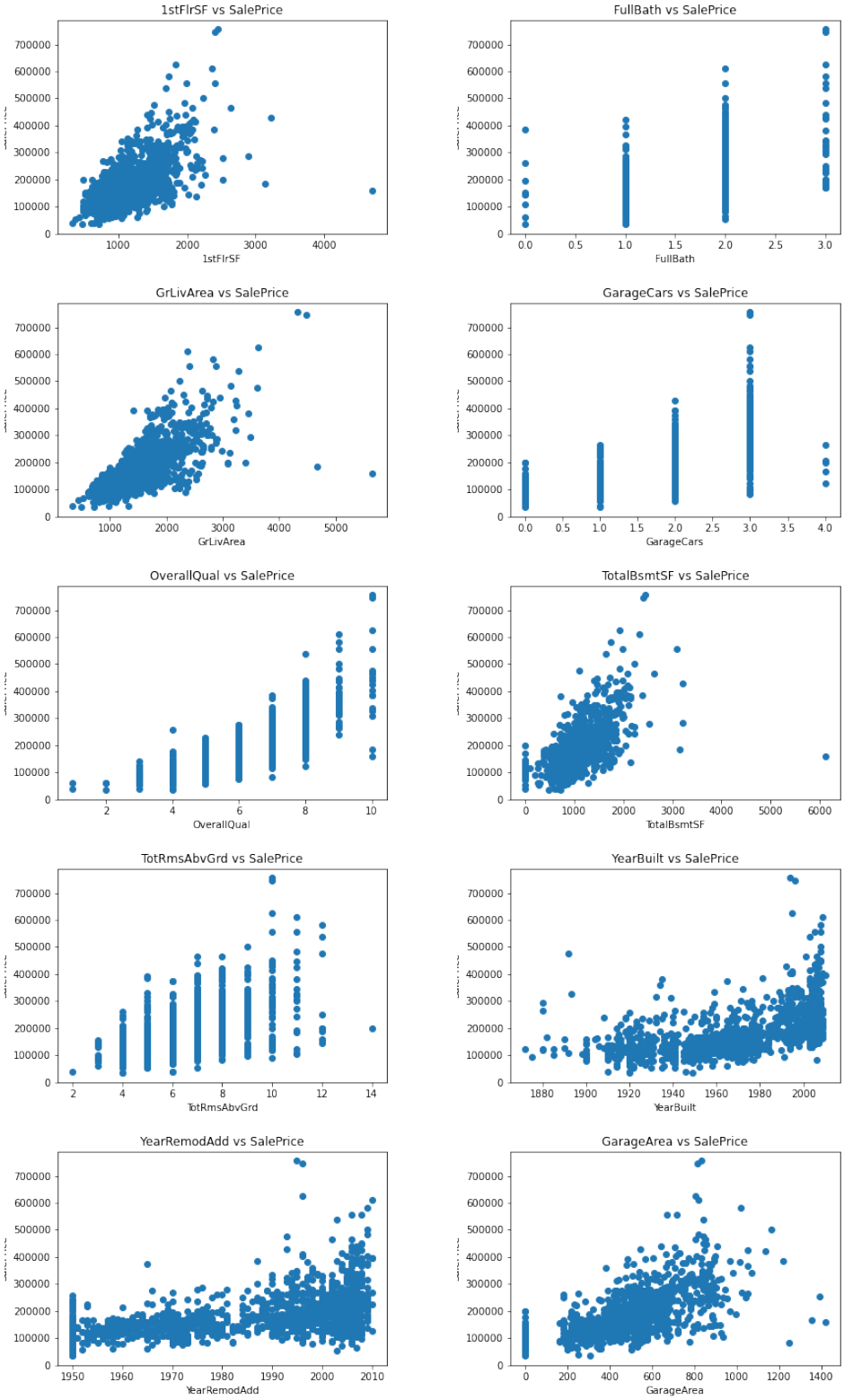
(a) Houses remodelled recently hold a higher value as compared to other houses remodelled before that.

(b) Houses built recently built hold cost higher than ancient houses except for some exceptions where a few of the ancient houses have relatively higher sales price

(c) There are not many houses with more than 10 rooms. But general pattern is similar to the previous trends that when when room increase price increases

(d) The increase in square feet area leads to increase in Sales Price

Plot/s



4. Make a histogram/bar plot for each of those 10 attributes influencing `SalePrice` and discuss the distribution of values, *e.g.*, are uniform, skewed, normal of those attributes. Place images of these

histograms into the document. Show the R/Python code that you used below and discussion below that.

R/Python script

```
for i in col_corr_with_target:
    plt.hist(x = df[i])
    plt.xlabel(i)
    plt.title(i+" vs "+"SalePrice")
    name = "Hist_"+i+"_vs_"+"SalePrice"
    plt.savefig(name)
    plt.show()
```

Discussion of Attributes

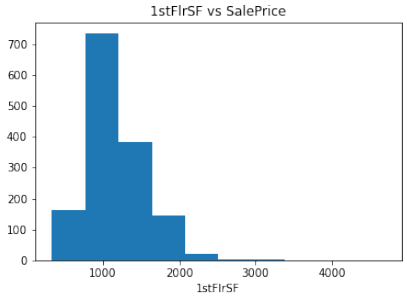
Ans.

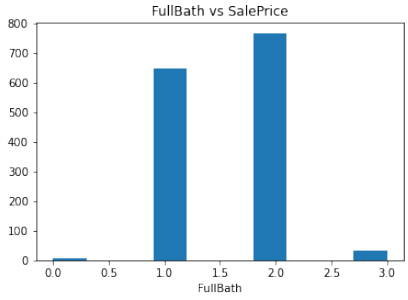
(a) Square Feet Area - first floor, garrage, basement, living room seem to follow a normal distribution

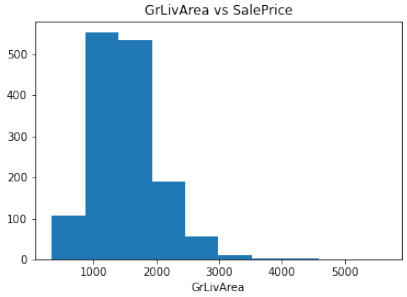
(b) The number of houses built have increased since 1880 and and maximum have been build in 2000

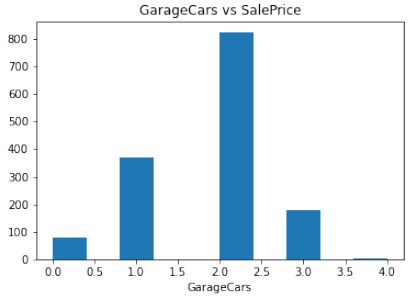
(c) Old houses and the newer houses are remmodelled the most.

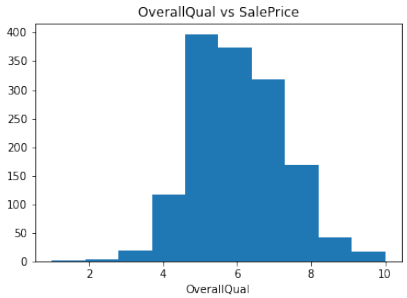
Histograms/Bar Plots

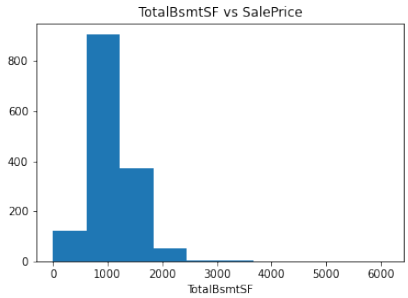


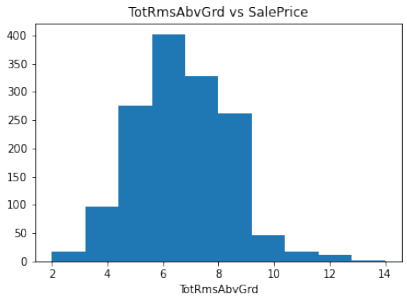


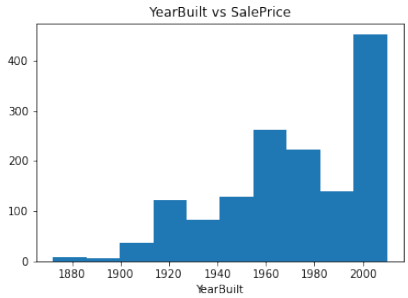


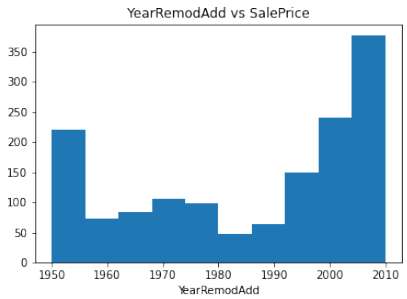


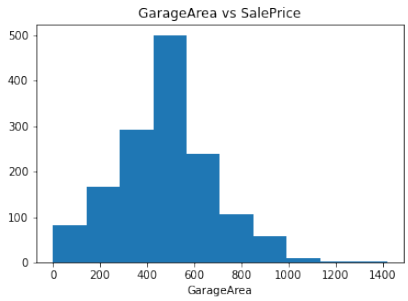












Discussion of simply removing tuples

Quantify the affect of simply removing the tuples with unknown or missing values. What is the cost in human capital?
Ans. There are around 6965 missing values and around 1400 rows of data missing for one of the attribute-Misc Feature. So if we simply remove the tuples with unknown or missing value we would be left with very less amount of data which woyuld not be enough for predicting accurate values for SalePrice.
The human capital loss could be huge since the predictions will be incorrect and misleading based on the given parameter.

Problem 5

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

1. Is noise ever interesting or desirable? Outliers?

   Ans. Noise is not interesting or desirable, however outliers might carry a meaning or reason for being an outlier to give further insight about the problem statement. Eg In fraud detection, outliers represent unusual events which are of great interest.

2. Can noise objects be outliers?

   Ans. Noise objects are usually random component of measurement error which might involve a distortion of value. Some of the noise may sometimes look like outliers, but at the same time some of them may camouflage in data.

3. Are noise objects always outliers?

   Ans. No, outliers are usually legitimate data which have some characteristics unusual with respect to typical values of data. On the other hand, noise is a random unwanted error in measurement.

4. Are outliers always noise objects?

   Ans. No, as said earlier, outliers are legitimate data whereas noise is not. Outlier is considered as an anomalous object rather than an unwanted random error.

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

   Ans. It depends on the type of noise. If it's just a small error in measurement, it might not make a typical value into an unusual one. But, if it's a random large error in a few measurements, they look unusual.

Problem 6

You are given a set of m objects that is divided into K groups, where the i^{th} group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following sampling schemes? (Assume sampling with replacement.)

1. We randomly select $n * m_i/m$ elements from each group.

2. We randomly select n elements from the data set, without regard for the group to which an object belongs.

Problem 6 continued on next page...

Ans. The first method selects n samples from each group of size m_i. This thus has elements from every group without missing any articular group. Additionally the elements chosen from each group are proportion to m_i, ie, the size of the i^th group. This implies that the sample represents the population better in that sense.
In the second method, the sampling is done randomly instead of doing it in a stratified way as done in the first. This does not guarantee that elements from each group are sampled, which would imply that it wouldn't represent the data distribution accurately.
Example: While analysing employee attrition, if we sample randomly (like mentioned in the second point), we might not retain the original data distribution of data with respect to gender. The sampled data might have 20:80 ratio while the actual data might have a 50:50. This could hamper the observations/analysis performed on that sample.

Problem 7

Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf_{ij}^{'} = tf_{ij} * \log \frac{m}{df_i},$$

where df_i is the number of documents in which the i^{th} term appears, which is known as the document frequency of the term. This transformation is known as inverse document frequency transformation.

1. What is the effect of this transformation if a term occurs in one document? In every document?

   Ans. If a term occurs only in one document, m = m and $df_i = 1$, this maximises the weight ie,

   $$tf_{ij}^{'} = tf_{ij} * log(m)$$

   On the contrary, if a term occurs in all the documents, then m = m and df_i = m, this makes the weight 0, ie,

   $$tf_{ij}^{'} = tf_{ij} * log(\frac{m}{m}) = tf_{ij} * log(1) = tf_{ij} * 0 = 0$$

2. What might be the purpose of this transformation?
   Ans. As seen in the previous question, the words which occur in all documents have weight 0, which mean that they are not very use-full in determining if the documents in question are different from each other. Whereas, the word those occur in fewer documents have higher weights can be used as a distinguishing factor among the documents.
   So, this transformation gives additional weight to unique words to capture more information of a document.

Problem 8

This question compares and contrasts some similarity and distance measures.

1. 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

- $\mathbf{y} = 0100011000$

Ans.
$\mathbf{x} = 0101010001$
$\mathbf{y} = 0100011000$
Hamming Distance $L_1 = f_{01} + f_{10} = 12 = 3$
Jaccard Coefficient $\mathbf{J} = \frac{f_{11}}{f_{00}+f_{01}+f_{10}+f_{11}} = \frac{2}{2+2+2} = \frac{2}{5} = 0.4$

2. Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming distance is a distance, while the other three measures are similarities, but don't let this confuse you.)
Ans. Hamming Distance is more similar to Simple Matching Coefficient whereas Jaccard Coefficient can be considered more similar to the cosine measure. This is because both Simple Matching Coefficient considers absences and presences equally and finds how similar two vectors and Hamming distance finds how dissimilar two vectors are.
Hamming Distance $L_1 = f_{01} + f_{10}$
SMC = $\frac{f_{11}+f_{00}}{f_{00}+f_{01}+f_{10}+f_{11}}$
We notice that, $\frac{L_1}{Total_number_of_bits} = SMC - 1$
Jaccard Coefficient on the other hand, gives importance to matching presences. Cosine similarly ignores the 0-0 matches like Jaccard does with addition to handling the non-binary vectors.

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

$Cos(x, y) = \frac{x \cdot y}{||x||||y||}$

3. 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. (Note: 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.)
Ans. Since we're comparing the genes that the organisms "share", jaccard would be a better option since it priortizes presense while finding how simmilar two vectors are. 0-0 matches are ignored. While in hamming distance it solely tells us the dissimilarity.

4. If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note: Two human beings share > 99.9% of the same genes.)
Ans. In this case, calculating Hamming distance would make more sense. Hamming distance would be able to tell the dissimilarity ie, 0-1 and 1-0. Also, since Two human beings share > 99.9% of the same genes, the 1-1 match would overpower. Hence, using Jaccard Distance would be a bad idea.

Problem 9

For the following vectors, **x** and **y**, calculate the indicated similarity and distance measures. Show detailed calculations/steps.

1. $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$ cosine, correlation, Euclidean.
Ans.

(a) Cosine Similarity

$$cos(x, y) = \frac{x \cdot y}{||x||||y||}$$

where

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$

$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2}$$

$$||y|| = \sqrt{\sum_{k=1}^{n} y_k^2}$$

$$x \cdot y = 1 \times 2 + 1 \times 2 + 1 \times 2 + 1 \times 2 = 8$$
$$||x|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = \sqrt{4} = 2$$
$$||y|| = \sqrt{2^2 + 2^2 + 2^2 + 2^2} = \sqrt{16} = 4$$
$$\implies cos(x, y) = \frac{8}{2 \times 4} = \mathbf{1}$$

(b) Correlation

$$corr(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

where

$$cov(x, y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$sd(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$sd(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

Computing:

$$\bar{x} = \frac{1 + 1 + 1 + 1}{4} = 1$$
$$\bar{y} = \frac{2 + 2 + 2 + 2}{4} = 2$$
$$cov(x, y) = \frac{1}{4-1}[4(1-1)(2-2)] = 0$$
$$sd(x) = \sqrt{\frac{1}{4-1}[4(1-1)^2]} = 0$$
$$sd(y) = \sqrt{\frac{1}{4-1}[4(12-2)^2]} = 0$$
$$corr(x, y) = \frac{0}{0} = \textbf{undefined}$$

(c) Euclidean

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Computing

$$d(x,y) = \sqrt{4[(1-2)^2]} = \sqrt{4} = \mathbf{2}$$

==========

2. $\mathbf{x} = (0,1,0,1)$, $\mathbf{y} = (1,0,1,0)$ cosine, correlation, Euclidean, Jaccard.

(a) Cosine Similarity

$$cos(x,y) = \frac{x \cdot y}{||x||||y||}$$

where

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$

$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2}$$

$$||y|| = \sqrt{\sum_{k=1}^{n} y_k^2}$$

$$x \cdot y = 0 \times 1 + 1 \times 0 + 0 \times 1 + 1 \times 0 = 0$$
$$||x|| = \sqrt{0^2 + 1^2 + 0^2 + 1^2} = \sqrt{2}$$
$$||y|| = \sqrt{1^2 + 0^2 + 1^2 + 0^2} = \sqrt{2}$$
$$\implies cos(x,y) = \frac{0}{\sqrt{2} \times \sqrt{2}} = \mathbf{0}$$

(b) Correlation

$$corr(x,y) = \frac{cov(x,y)}{sd(x)sd(y)}$$

where

$$cov(x,y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$sd(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$sd(y) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

Computing:

$$\bar{x} = \frac{0+1+0+1}{4} = \frac{1}{2}$$
$$\bar{y} = \frac{1+0+1+0}{4} = \frac{1}{2}$$
$$cov(x,y) = \frac{1}{4-1}[4(0-\frac{1}{2})(1-\frac{1}{2})] = -\frac{1}{3}$$
$$sd(x) = \sqrt{\frac{1}{4-1}[1]} = \sqrt{\frac{1}{3}}$$
$$sd(y) = \sqrt{\frac{1}{4-1}[1]} = \sqrt{\frac{1}{3}}$$
$$corr(x,y) = \frac{-\frac{1}{3}}{\sqrt{\frac{1}{3}}\sqrt{\frac{1}{3}}} = \textbf{-1}$$

(c) Euclidean

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Computing

$$d(x,y) = \sqrt{4[(1)^2]} = \sqrt{4} = \mathbf{2}$$

(d) Jaccard Coefficient

$$J = \frac{f_{11}}{+f_{01} + f_{10} + f_{11}}$$
$$J = \frac{0}{(2+2+0)} = \mathbf{0}$$

==========

3. $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean.

(a) Cosine Similarity

$$cos(x,y) = \frac{x \cdot y}{||x||||y||}$$

where

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$
$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2}$$
$$||y|| = \sqrt{\sum_{k=1}^{n} y_k^2}$$

$$x \cdot y = 0 \times 1 + -1 \times 0 + 0 \times -1 + 1 \times 0 = 0$$
$$||x|| = \sqrt{0^2 + 1^2 + 0^2 + 1^2} = \sqrt{2}$$
$$||y|| = \sqrt{1^2 + 0^2 + 1^2 + 0^2} = \sqrt{2}$$
$$\implies cos(x,y) = \frac{0}{\sqrt{2} \times \sqrt{2}} = \mathbf{0}$$

Problem 9 continued on next page...

(b) Correlation

$$corr(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

where

$$cov(x, y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$sd(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$sd(y) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

Computing:

$$\bar{x} = \frac{0 - 1 + 0 + 1}{4} = 0$$

$$\bar{y} = \frac{1 + 0 - 1 + 0}{4} = 0$$

$$cov(x, y) = \frac{1}{4-1}[4(0)] = 0$$

$$sd(x) = \sqrt{\frac{1}{4-1}[1+1]} = \sqrt{\frac{2}{3}}$$

$$sd(y) = \sqrt{\frac{1}{4-1}[1+1]} = \sqrt{\frac{2}{3}}$$

$$corr(x, y) = \frac{0}{\sqrt{\frac{2}{3}}\sqrt{\frac{2}{3}}} = \mathbf{0}$$

(c) Euclidean

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Computing

$$d(x, y) = \sqrt{(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-1)^2} = \sqrt{4} = \mathbf{2}$$

==========

4. $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard.

(a) Cosine Similarity

$$cos(x, y) = \frac{x \cdot y}{||x||||y||}$$

where

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$

$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2}$$

$$||y|| = \sqrt{\sum_{k=1}^{n} y_k^2}$$

$$x \cdot y = 1 \times 1 + 1 \times 1 + 0 \times 1 + 1 \times 0 + 0 \times 0 + 1 \times 1 = 3$$

$$||x|| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{4} = 2$$

$$||y|| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2} = \sqrt{4} = 2$$

$$\implies cos(x, y) = \frac{3}{2 \times 2} = \mathbf{0.75}$$

(b) Correlation

$$corr(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

where

$$cov(x, y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$sd(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$sd(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

Computing:

$$\bar{x} = \frac{1+1+0+1+0+1}{6} = \frac{2}{3}$$

$$\bar{y} = \frac{1+1+1+0+0+1}{6} = \frac{2}{3}$$

$$cov(x, y) = \frac{1}{6-1}[\frac{3}{9}] = \frac{1}{15}$$

$$sd(x) = \sqrt{\frac{1}{6-1}\frac{16}{9}} = \frac{4}{3}\sqrt{\frac{1}{5}}$$

$$sd(y) = \sqrt{\frac{1}{6-1}\frac{16}{9}} = \frac{4}{3}\sqrt{\frac{1}{5}}$$

$$corr(x, y) = \frac{\frac{1}{15}}{\frac{4}{3}\sqrt{\frac{1}{5}}\frac{4}{3}\sqrt{\frac{1}{5}}} = \mathbf{0.25}$$

Problem 9 continued on next page...

(c) Jaccard Coefficient

$$J = \frac{f_{11}}{+f_{01} + f_{10} + f_{11}}$$
$$J = \frac{3}{(1+1+3)} = \frac{3}{5} = \mathbf{0.6}$$

==========

5. $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0 - 1)$ cosine, correlation.

(a) Cosine Similarity

$$cos(x, y) = \frac{x \cdot y}{||x||||y||}$$

where

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$
$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2}$$
$$||y|| = \sqrt{\sum_{k=1}^{n} y_k^2}$$

$$x \cdot y = 2 \times -1 + -1 \times 1 + 0 \times -1 + 2 \times 0 + 0 \times 0 + -3 \times -1 = 0$$
$$||x|| = \sqrt{2^2 + -1^2 + 0^2 + 2^2 + 0^2 + -3^2} = \sqrt{18}$$
$$||y|| = \sqrt{-1^2 + 1^2 + -1^2 + 0^2 + 0^2 + -1^2} = \sqrt{4} = 2$$
$$\implies cos(x, y) = \frac{0}{\sqrt{18}\sqrt{4}} = \mathbf{0}$$

(b) Correlation

$$corr(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

where

$$cov(x, y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$
$$sd(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$
$$sd(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$
$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$
$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

Computing:

$$\bar{x} = \frac{2 + -1 + 0 + 2 + 0 + -3}{6} = 0$$

$$\bar{y} = \frac{-1+1+-1+0+0+-1}{6} = 0$$

$$cov(x,y) = \frac{1}{6-1}[-2+-1+0+0+3] = 0$$

$$sd(x) = \sqrt{\frac{1}{6-1}18} = \sqrt{\frac{18}{5}}$$

$$sd(y) = \sqrt{\frac{1}{6-1}4} = \sqrt{\frac{4}{5}}$$

$$corr(x,y) = \frac{0}{\sqrt{\frac{18}{5}}\sqrt{\frac{4}{5}}} = \mathbf{0}$$