B565-Data Mining
Homework #1

Due on Tuesday, Jan 30, 2024 11:00P

Dr. Dalkilic

Nilambar Halder Tonmoy, Benson Grichinga

January 30, 2024

Contents

Problem 1

Classification is the building of an approximation of a function. To that end, determine which of the relations are functions. Provide argument (no formal proof) if it is, and an example of a violating case when it is not. $f \in \mathbb{R}^2$ such that $(x, y) \in f$

(a) $2y = x + 4$.
Solution: We can rewrite the equation as, $y = x/2 + 2$
The above equation follows the structure of a linear function $y = mx + c$, which is in terms of a valid function itself given that for an input we will get only one output for the function. So it is certain that the equation is a function.

(b) $x = |y|$.
Solution: Assuming x is dependent on y we can simplify the equation further to get two new equations $x = y$ & $x = -y$
Now, from above we can see that for any value of y, we get two results. For example, $y = 1$ will give $x = 1$ and $x = -1$ respectively. So this relation cannot be considered a function.

(c) $x^2 + y^2 = 1$.
Solution: This relation corresponds to the equation of a circle which is, $(x - k)^2 + (y - l)^2 = r^2$. By definition we know that circle is not a function, however, let's rewrite the equation as $y = \pm\sqrt{1 - x^2}$, which is evident that for a single value of x, we would get two different results as y. Hence, the given relation is not a function.

(d) $x^2 + y^2 = 1$.
Solution: This relation is the same as *Problem 1.c* so the outcome will be the same which is, the relation does not represent a function.

(e) $\sqrt{x} = \sqrt{y}$.
Solution: This relation represents the linear function $y = mx + c$, where, $m = 1$ and $c = 0$ in this relation. As explained in **Problem 1.a** this relation also supports the function theory.

(f) $y^2 = x^2$.
Solution: Let's rewrite the equation as $y = \pm x$. From this relation, it is sufficient to state that for a value of x, we would get two different values for y. Thus, it can be argued that the given relation is not a function. Also, we can further support our claim by plugging in $x = -1$ in the given relation and can see that solving the equation results in $y = \pm$.

(g)

$$xy = 8 \quad (1)$$
$$x - y + 2 = 0 \quad (2)$$

where both equations are true.
Solution: Assuming, the equation represents the approximation function of a single classifier instance, we can simplify both equations as,

$$y = 8/x \quad (3)$$
$$y = x + 2 \quad (4)$$

Now, in (3) and (4) we can set $x = 4$ and can see that for y we get 2 and 6 respectively. Hence, the equations are not function

Problem 2

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

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

Solution: The given function can be considered if it follows the four properties[5] of a metric. We provide proof that this is indeed a function below,

1. Since, $\forall I, 1 \leq i \leq |x_i - y_i| > 0$
   So, $d(x, y) = max\{|x_i - y_i|\} > 0$ as the result will be positive
2. Now, if $x = y$ then $d(x, y) = 0$
   So, here if $x = y$ then the function becomes,
   $d(x, x) = max\{|x_i - x_i|\} = 0$
   So the argument holds.
3. The commutative property states that $d(x, y) = d(x, z)$
   Let's consider two values m & n. A function using these values would be

$$d(m, n) = max\{|m - n|\} \tag{5}$$

   Now let us multiply the value with -1, to which it becomes,

$$d(m, n) = max\{|n - m|\} \tag{6}$$

   We can see that the function will still have the same result. So it is commutative.
4. From the triangle law we know, $d(x, z) \leq d(x, y) + d(y, z)$
   Now, let's assume that l is the index where the max value resides. So we can rewrite the original equation as,

$$d(x_l, y_l) \leq |x_l - y_l| \leq x_l - y_l \tag{7}$$
$$d(y_l, z_l) \leq |y_l - z_l| \leq y_l - z_l \tag{8}$$

   Now adding both [7] [8] we get,

$$d(x_l, z_l) + d(y_l, z_l) \leq x_l - y_l + y_l - z_l \tag{9}$$
$$d(x_l, z_l) + d(y_l, z_l) \leq x_l - z_l \tag{10}$$
$$d(x, z) \leq d(x_l, z_l) + d(y_l, z_l) \tag{11}$$

   So the triangle law holds and thus the function is a metric.

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

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

In class we did the proof using equivalence relations–but usually, it's done establishing a metric. This is the loss function we've studied $L(D, \hat{f})$. For this, it's easier to show for the triangle law this weaker axiom:

Weak Triangle Axiom

For distinct $a, b, c \in X$, $d(a, c) \leq d(a, b) + d(b, c)$

Proof
Suppose $a = b$. Then:

$$d(a,c) \quad = \quad d(b,c) = d(b,b) + d(b,c) \leq d(a,b) + d(b,c) \tag{12}$$

For $b = c$ the argument is the same. Now consider $a = c$.

$$d(a,c) \quad = \quad 0 \leq d(a,b) + d(b,c) \tag{13}$$

Solution: We can evaluate the given function following the triangle law to see that it is indeed a metric.
From triangle law, we know that $c(x,z) \leq c(x,y) + c(y,z)$. So, for $c(x,y)$ and $c(y,z)$ we get,

$$c(x,y) \quad = \quad 1[x \neq y] \tag{14}$$
$$c(x,y) \quad = \quad 0[x = y] \tag{15}$$
$$c(y,z) \quad = \quad 1[x \neq y] \tag{16}$$
$$c(y,z) \quad = \quad 0[x = y] \tag{17}$$

So adding the above equations based on x and y,

$$c(x,z) \quad = \quad 2[x \neq y] \tag{18}$$
$$c(x,z) \quad = \quad 0[x \neq y] \tag{19}$$

We can see that the law still holds.
■

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

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

Solution: For the given function, let's consider the properties of metric,

1. Since, n is positive incrementing value so the value of $c(x_i, y_i)$ would be positive, hence, $c(x_i, y_i) > 0$
2. For $x = y$, $c(x,x)$ can be 0 so this property also holds.
3. For $c(x,y) = c(y,x)$ the property holds as the outcome will not be affected.
4. Finally, the triangle property states that $c(x,z) \leq c(x,y) + c(y,z)$. Now since it is incremental over $1 \leq i \leq n$ the property will hold.

So we can say that this is a metric.

(d) Suppose d_0, d_1 are metrics.

i. $d_0 \times d_1$
Solution: It is not possible to state whether $d_o x d_1$ is a metric as it is intrinsic to the property of both d_0 and d_1. For example, if both are metric then it may hold for the property of $d_i(x,y) > 0$ and $d_i(x,y) = 0[x = y]$ but may fail the triangle axiom. So the given information is not sufficient to measure if the product of the metrics is a metric.

ii. $(d_0 + d_1)/(d_0 d_1)$
Solution: For the given function, the metric property mainly failed for $d(x,y) = 0$ when $x = y$. We can see that it is being divided by $d_0 x d_1$ so for $x = y$ the function becomes unstable. So, we can that it is not a metric.

iii. $max\{d_0,\ d_1\}$
Solution: As mentioned in Problem (a), since both d_0 and d_1 are metric, the max of them will also be a metric.

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

$$d(x,z) = \frac{||x \cap z||}{||x \cup z|| + 1} \tag{20}$$

$$d(x,y) = \frac{||x \cap y||}{||x \cup y|| + 1} \tag{21}$$

$$d(y,z) = \frac{||y \cap z||}{||y \cup z|| + 1} \tag{22}$$

Now, let's take an arbitrary set $x = \{1, 2, 3\}$, $y = \{7, 8, 9\}$, and $z = \{1, 2, 3, 5\}$. So for

$$d(x,z) = \frac{3}{5} \tag{23}$$

$$d(x,y) = 0 \tag{24}$$

$$d(y,z) = 0 \tag{25}$$

So, we can see that $d(x, z) \geq d(x, y) + d(y, z)$ stands and hence it is not a metric.

Problem 3

JN I
Curse of Dimensionality: Generate m-dimensional n data points from a uniform distribution with values between 0 and 1. For an arbitrary m value

$$f(m) = \log_{10} \frac{d_{max}(m) - d_{min}(m)}{d_{min}(m)}$$

where $d_{max}(m)$ and $d_{min}(m)$ are the maximum and minimum distances between any pair of points, respectively. Let m take each value from $\{1, 2, \ldots, 99, 100\}$. Repeat each experiment multiple times to get stable values by averaging the quantities over multiple runs for each m. For four different n values, e.g., $n \in \{150, 1500, 15000, 150000\}$, plot $f(m)$. Use Euclidean as your distance metric. Label and scale each axis properly and discuss your observations over different n's.

Discussion of Curse of Dimensionality

- Describe the environment (Language (version), OS, computer)
- Describe the outcome and its consequences
- Include one or two graphics that have the ordinate, abscissa labeled as well as a title. The figure(s) should have meaningful captions.

Solution: For this problem, we prepared a Jupyter notebook named *jn1.ipynb* which can be accessed in our repository.

- We used two different environments in this experiment
  - Colab : Intel Xeon (R), 2.2GHz, 12 GB RAM, and Python 3.10.12 Environment

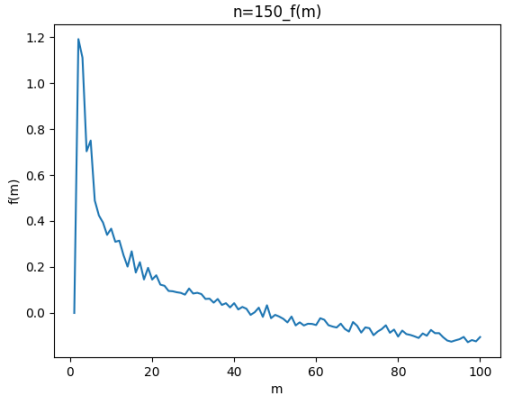

Figure 1: f(m) vs m for N=150

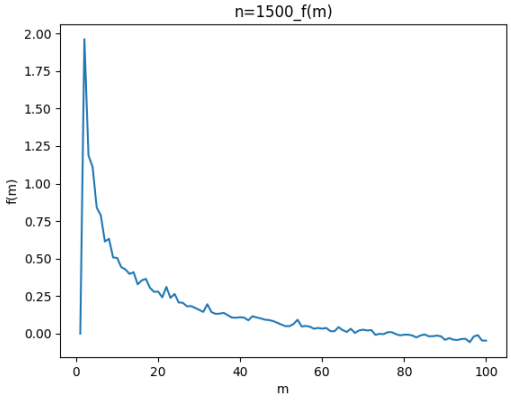

Figure 2: f(m) vs m for N=1500

  - Laptop: Apple M1 Pro, 16GB RAM, and Python 3.11 Virtual Environment

- In this experiment, we prepared a function to dynamically create a nxm dataset where $m\epsilon\{1, 2, 3 \ldots 100\}$ and $n\epsilon\{150, 1500, 15000, 150000\}$. Also, for the value of $f(m)$ for each of $m - i\epsilon m$ we went through values of n and setup a trial for 10 runs. Afterward, we projected our findings by averaging $f(m)$ over the trial runs for each n and m. However, we would like to note that for $n\epsilon\{15000, 150000\}$ each run takes a significant amount of time for each of $m_i\epsilon m$. A sample execution time-log is given in 5. Here, *d_min_max_time* represents the execution time for each n and m, simply the total calculation time to find $f(m)$. We can see that for $n = 15000$ and $m = 1$ the execution time is 481s and there's a gradual increase with the increase of m. Although the code was able to run for all the cases when $n = 15000$ and $m\epsilon 1, 2, 3 \ldots 100$, we never saw a result for $n = 150000$ value. However, based on the trend we identified, the execution time increases approximately 10 times with the increase of data size n.

- For, $n\epsilon\{150, 1500\}$ we were able to successfully project our finding of how the value $f(m)$ is affected when $n = 150$ in figure 1 and for $n = 1500$ in figure 2 when $m\epsilon\{1, 2, 3 \ldots 100\}$. The trend for execution time for both $n = 150$ & $n = 1500$ is projected in figure 3 & figure 4.

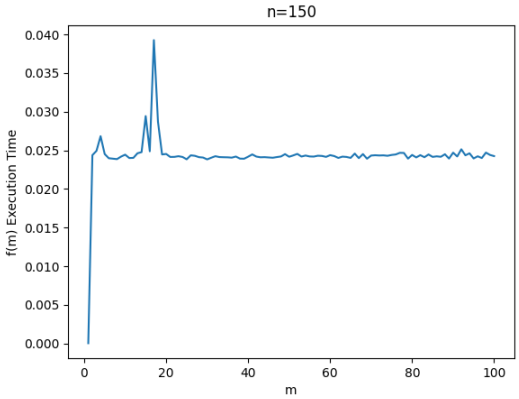

Figure 3: f(m) Execution Time vs m for N=150

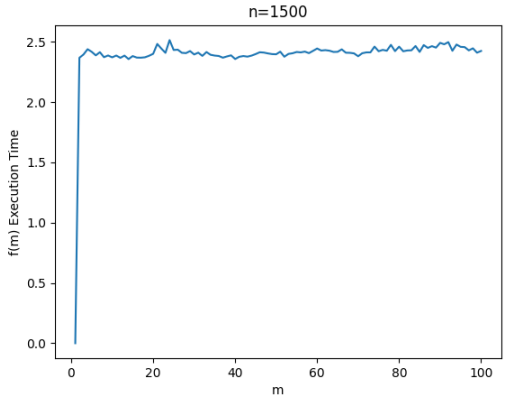

Figure 4: f(m) Execution Time vs m for N=1500

```
n=1500, f_m_time=4.8480119705200195 m=100, d_time=0.0008311271667480469 d_min=2.894419108880964, d_max=5.114163145014282, d_min_max_time=4.847172021865845 f(m)=-0.11525851889900687
n=15000, f_m_time=481.84858298301697 m=1, d_time=0.00022101402282714844 d_min=8.941062445977366e-09, d_max=0.999907077543313, d_min_max_time=481.84828996658325 f(m)=8.04857051045018
n=15000, f_m_time=473.4660620689392 m=2, d_time=0.0006780624389648438 d_min=4.273295192786656e-06, d_max=1.3962188945477105, d_min_max_time=473.46537590026855 f(m)=5.514189287393716
n=15000, f_m_time=470.95186829566956 m=3, d_time=0.0003039836883544922 d_min=0.0015470778240922892, d_max=1.6753043166447717, d_min_max_time=470.95155596733093 f(m)=3.034180307333752
n=15000, f_m_time=470.73189091682434 m=4, d_time=0.0003790855407714844 d_min=0.007332155335323237, d_max=1.8639216687345634, d_min_max_time=470.7315037250519 f(m)=2.403484235962356
n=15000, f_m_time=473.4460427761078 m=5, d_time=0.0003969669342041015 6 d_min=0.022104170736367137, d_max=2.000962654966846, d_min_max_time=473.4456367492676 f(m)=1.9519405107649692
n=15000, f_m_time=482.54797101020813 m=6, d_time=0.0004301071166992187 5 d_min=0.025070749242869422, d_max=2.1234774328224435, d_min_max_time=482.5474648475647 f(m)=1.9227223477404858
n=15000, f_m_time=480.8621780872345 m=7, d_time=0.0009009838104248047 d_min=0.05380046369215238 4, d_max=2.2720699094132275, d_min_max_time=480.8612689971924 f(m)=1.615228278552461
n=15000, f_m_time=471.67931389808655 m=8, d_time=0.0006389617919921875 d_min=0.07160624151764455, d_max=2.2901200755562856, d_min_max_time=471.67866706848145 f(m)=1.4911112622100215
n=15000, f_m_time=476.87618470191956 m=9, d_time=0.0007398128509521484 d_min=0.11245479189372 43, d_max=2.4136076265745388, d_min_max_time=476.87539410591125 f(m)=1.3109674977698207
n=15000, f_m_time=494.47951006889343 m=10, d_time=0.000942707061767578 1 d_min=0.15106878311139 79, d_max=2.491760588976673, d_min_max_time=494.4785590171814 f(m)=1.1901695038918687
n=15000, f_m_time=480.1571640968323 m=11, d_time=0.0011031627655029297 d_min=0.15981339053181 498, d_max=2.51241048858247, d_min_max_time=480.15599179267883 f(m)=1.1679343915718552
n=15000, f_m_time=478.2155718803406 m=12, d_time=0.0010700225830078125 d_min=0.25007753309855 85, d_max=2.6277923248050183, d_min_max_time=478.21446776390076 f(m)=0.9780850829490186
n=15000, f_m_time=471.09515404701233 m=13, d_time=0.0010504085382080078 d_min=0.24613429458503902, d_max=2.6979971003886805, d_min_max_time=471.094092130661 f(m)=0.9983240361476469
n=15000, f_m_time=475.55335330963135 m=14, d_time=0.0010192394256591797 d_min=0.29659859473312916, d_max=2.766509725376178, d_min_max_time=475.5523269176483 f(m)=0.9205122382463885
n=15000, f_m_time=476.3544890880585 m=15, d_time=0.0010640621185302734 d_min=0.3297299157587569, d_max=2.8814629040565363, d_min_max_time=476.35341691970825 f(m)=0.8886768765000074
n=15000, f_m_time=471.9031708240509 m=16, d_time=0.0012848377227783203 d_min=0.40048289458708203, d_max=2.8028774432749173, d_min_max_time=471.9018759727478 f(m)=0.7780603624834315
```

Figure 5: Curse of Dimensionality Notebook Execution Logs

Problem 4

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

1. A section of highway on a map. - nominal
2. The value of a stock. - ratio
3. Your grade in the class. - ordinal
4. Reviewing something, *e.g.*, movie, purchase, food. - ordinal
5. The weight of a person. - ratio
6. Visiting United Airlines (https://www.united.com) the seating is: Economy, Economy plus, and United Business. - Nominal

Problem 5

You are datamining with a column that includes a physical address in a city with only one 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?
Solution: Typically, US address follows the pattern where address line 1 contains the street address(e.g. House number, street name, etc.) and address line 2 contains the apartment number or such. The following name after that represents the city, state, and postal code. However, not necessarily this is present always and denominations after the address line 2 might not be present at all. Considering the given example we can see that the addresses presented mostly follow the address line concept mentioned before.

Let's number the addresses mentioned in this order for reference in our explanation,

1. 55 WEST CIR
2. 2131 South Creek Road
3. Apt. #1 Fountain Park
4. 1114 Rosewood Cir
5. 1114 Rosewood Ct.
6. 1114 Rosewood Drive

Problem 5 continued on next page...

Now, we can see that address numbers 1,2,4,5,6 have mostly similar patterns and 3 differ at the start. From this, we can derive a brute-force method to extract the information which is, first extract the first part of an address which is most likely the house number. For example, we can number 1 can be divided into two parts 55 & WEST CIR. We can already differentiate between area and house number. Now, for extracting information from the second part we can build a database of known street suffixes. This will help us differentiate between different streets in our addresses. Using this we can match the street address with our database remove the matched word from the street and categorize that as a street suffix. Now, the only difference here we can see is number 3 which doesn't have any street address and instead of the house number, contains the apartment number. In such cases, we can run a checker to see if an address doesn't start with a number(assuming the case is repetitive in many instances in the dataset) and have a different classifier that extracts information from this type of data. So, in general, we can project our data mining method in figure 6.

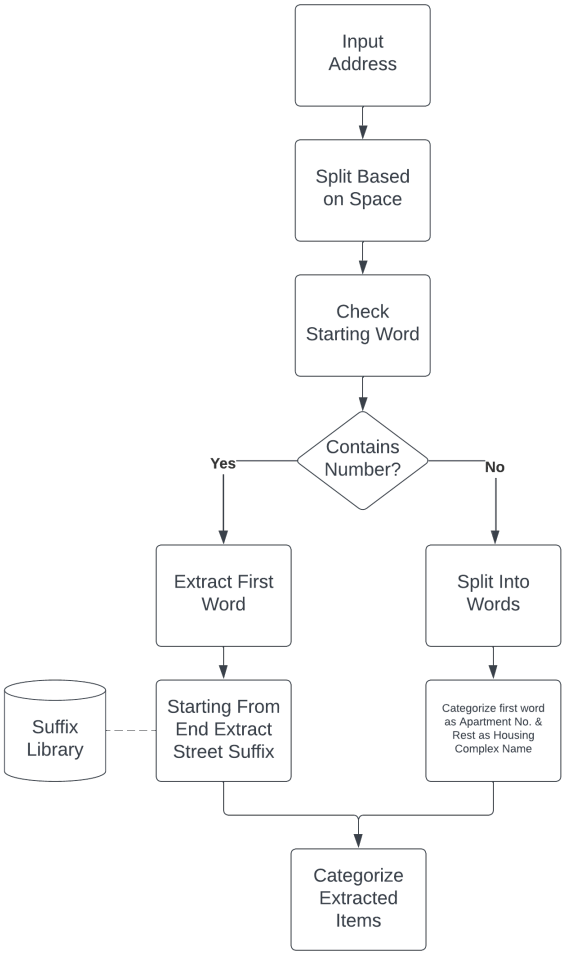

Figure 6: Address Extraction Pipeline

Problem 6

For this problem, you will be using a data set with total 81 attributes of private homes in Ames, Iowa. You'll be provided with a schema and instance.

Background

Generally this kind of data is used to determine value of a home. Given some home $h \in H$ from the set of homes H, the property taxes are $\tau \times v(h)$ where $0 \leq \tau \leq 1$ (the tax) and $v : H \rightarrow \$\mathbb{R}_{\geq 0}$ is the value. The county auditor decides v on history and similar homes. An owner can appeal v by showing $h_1, h_2, \ldots, h_6 \in H$ where $v(h_i) < v(h)$ and $h_i \approx h$ (the homes are very similar).

JN II
Data Exploration using Housing Data

- Determine the size (number of tuples, attribures (or features)).
- How many missing data exist?
- What are the three columns with the greatest number of missing data?
- What are the three columns with the largest number of values?
- What are the three collumns with the greatest variance?
- What are the three columns with the most uniform values?
- Find 10 **individual** attributes that seem to determine the class `SalePrice`. For this only use sensible plotting methods.
- Of the 10 attributes above, show the two that seem be the most linearly related.

Discussion of Home Price Data

- Discuss problems with determining value if you do *not* look at the column data.
- Make a histogram/bar plot for each of those 10 attributes that best determine `SalePrice`. Discuss the distribution of values, *e.g.*, uniform, skewed, normal of those attributes. Place images of these histograms into the document.
- Discuss the problem of determining price (numeric function). What would be the most likely kind of model to build?
- In your columns, discuss missing data and two techniques: remove the tuple and replace the missing value with the mean.

Solution:

- The dataset has a total number of 1460 entries spread into 81 columns or attributes. The first column named, "Id" is more like a key for each of the instances and cannot be considered as an attribute for data analysis. So we have considered the rest of 80 attributes in our analysis. A sample of the data can be seen in figure 20
- There are a total of 19 attributes that have missing data inside. The frequency of missing data in each attribute is projected in figure 21.

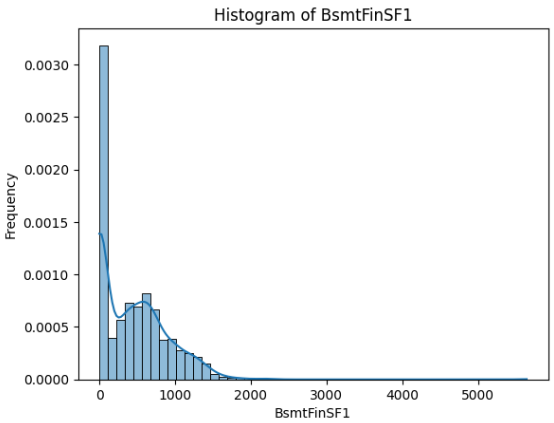

Figure 7: Data Distribution of BsmtFinSF1 Attribute

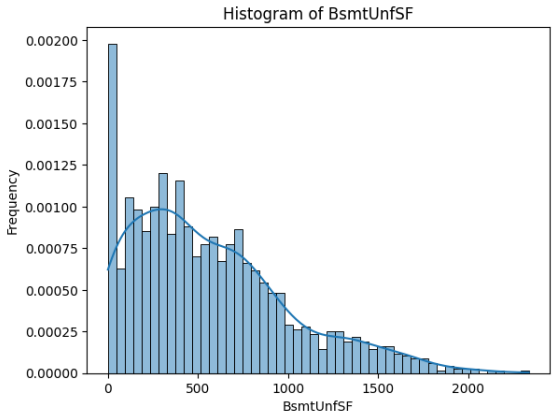

Figure 8: Data Distribution of BsmtUnfSF Attribute

- From figure 21 we can see that the top three attributes in terms of empty data are "PoolQC", "MiscFeature", and "Alley".
- The three columns that has the largest values are "SalePrice", "LotArea", and "MiscVal".
- The three columns with the greatest variances are "SalePrice", "LotArea", and "GrLivArea"
- The three columns with the most uniform values are "Neighborhood", "Exterior2nd", and "Exterior1st"
- The top 10 attributes with the most correlation with SalePrice are, "OverallQual", "OverallCond", "FullBath", "BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "1stFlrSF", "BsmtFullBath", "BsmtHalfBath", "GrLivArea".
- Now among the correlated attributes we see that only BsmtFinSF1, BsmtUnfSF, 1stFlrSF, and GrLivArea have a normal distribution and can be seen in figure 9, 9, 11, and 13 respectively. The other attributes are seen in figure 14, 15, 15, 17, 18, 19,
- To determine the SalPrice it is better to use a Regression analytic approach as we can see that the price of a house cannot be partitioned into n-classes. So, given the attributes we can run a regressor analyzer to predict the SalePrice of the houses.
- The issue with a missing value that we determined is very hard to impute for this experiment. For some attributes the missing values are high in percentage so to impute the missing values per se with the mean of the existing data will make the data highly biased. Also, for the categorical attributes it is not possible to use the mean-based imputation approach, however, we determined that we can use a semi-learning approach where we predict the missing values using existing classifiers and the rest of the available data. This would somewhat solve our issue, but, a high percentage of missing data in other attributes will bias the model into the training dataset. Another option is to understand the distribution of the data and try to impute data based-on the newly found distribution.

Problem 7

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

1. Is noise ever interesting or desirable? Outliers?

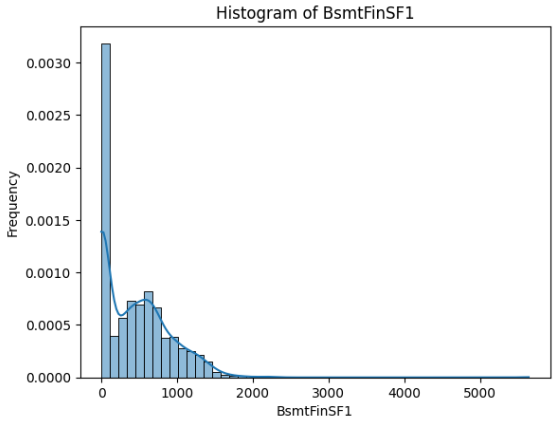

Figure 9: Data Distribution of BsmtFinSF1 Attribute

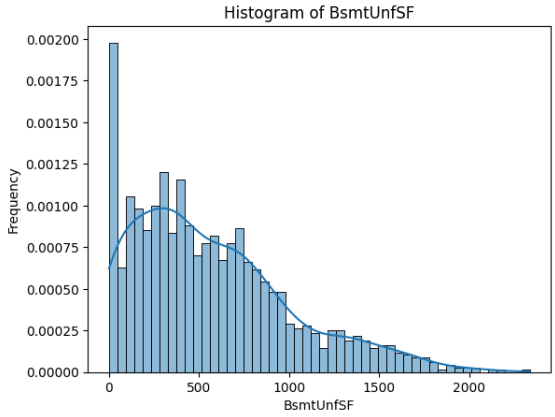

Figure 10: Data Distribution of BsmtUnfSF Attribute

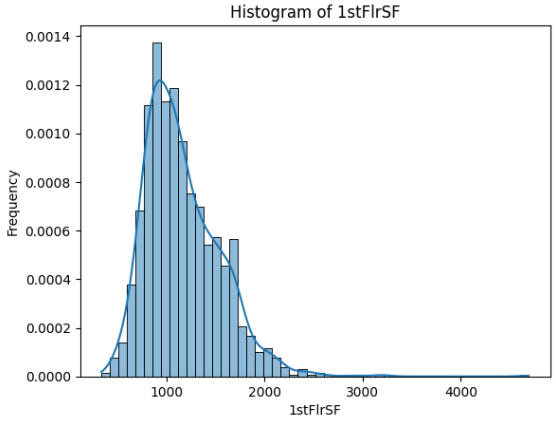

Figure 11: Data Distribution of 1stFlrSF Attribute

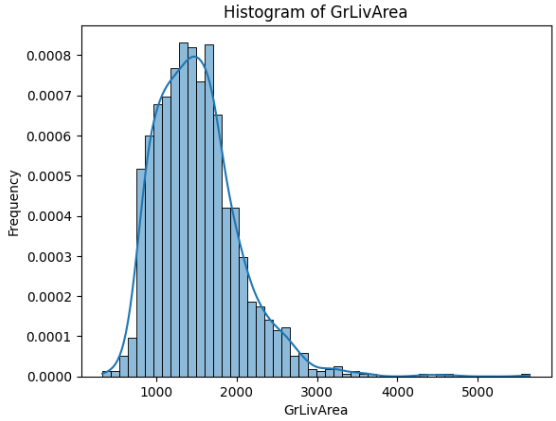

Figure 12: Data Distribution of GrLivArea Attribute

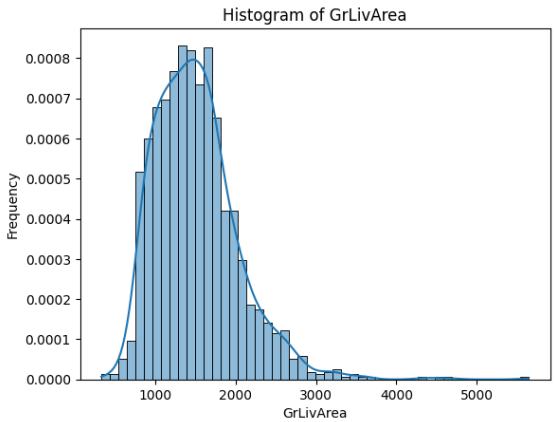

Figure 13: Data Distribution of GrLivArea Attribute

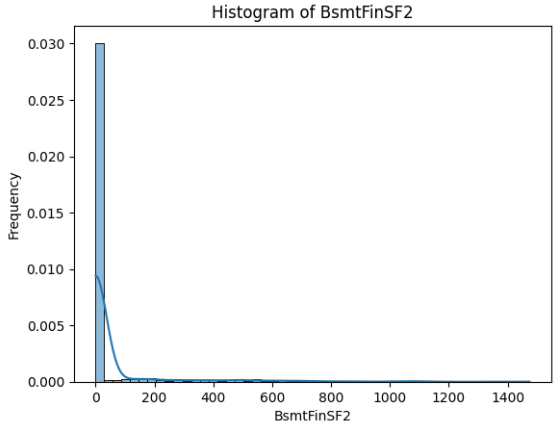

Figure 14: Data Distribution of BsmtFinSF2 Attribute

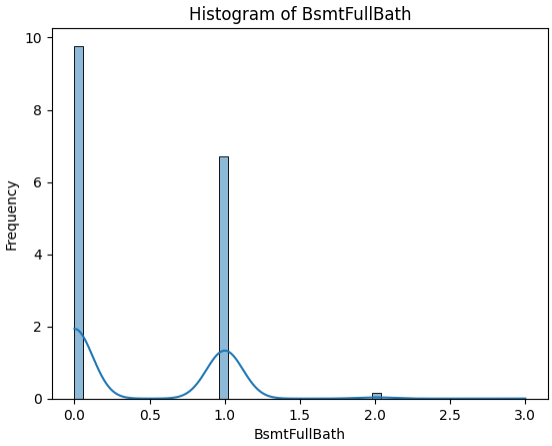

Figure 15: Data Distribution of BsmtFullBath Attribute

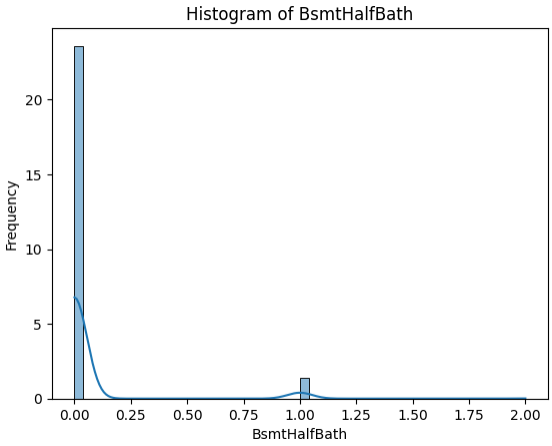

Figure 16: Data Distribution of BsmtHalfBath Attribute

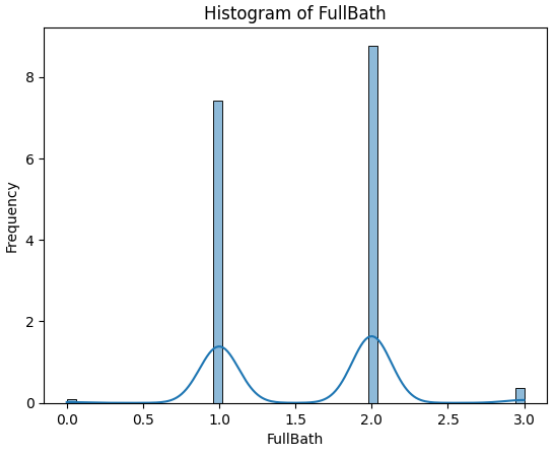

Figure 17: Data Distribution of FullBath Attribute

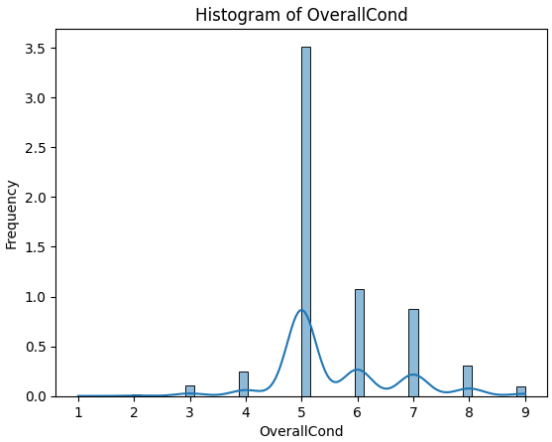

Figure 18: Data Distribution of OverallCond Attribute

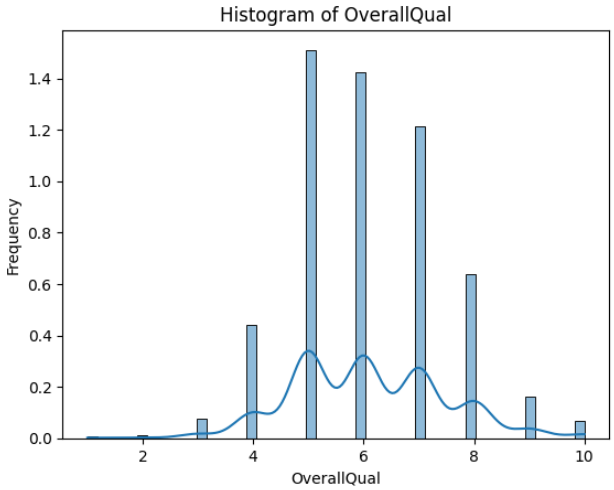

Figure 19: Data Distribution of OverallQual Attribute

```
df = pd.read_csv(dataset_path)
df.head()
```
✓ 0.0s

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000

5 rows × 81 columns

Figure 20: Sample Data From Housing Dataset

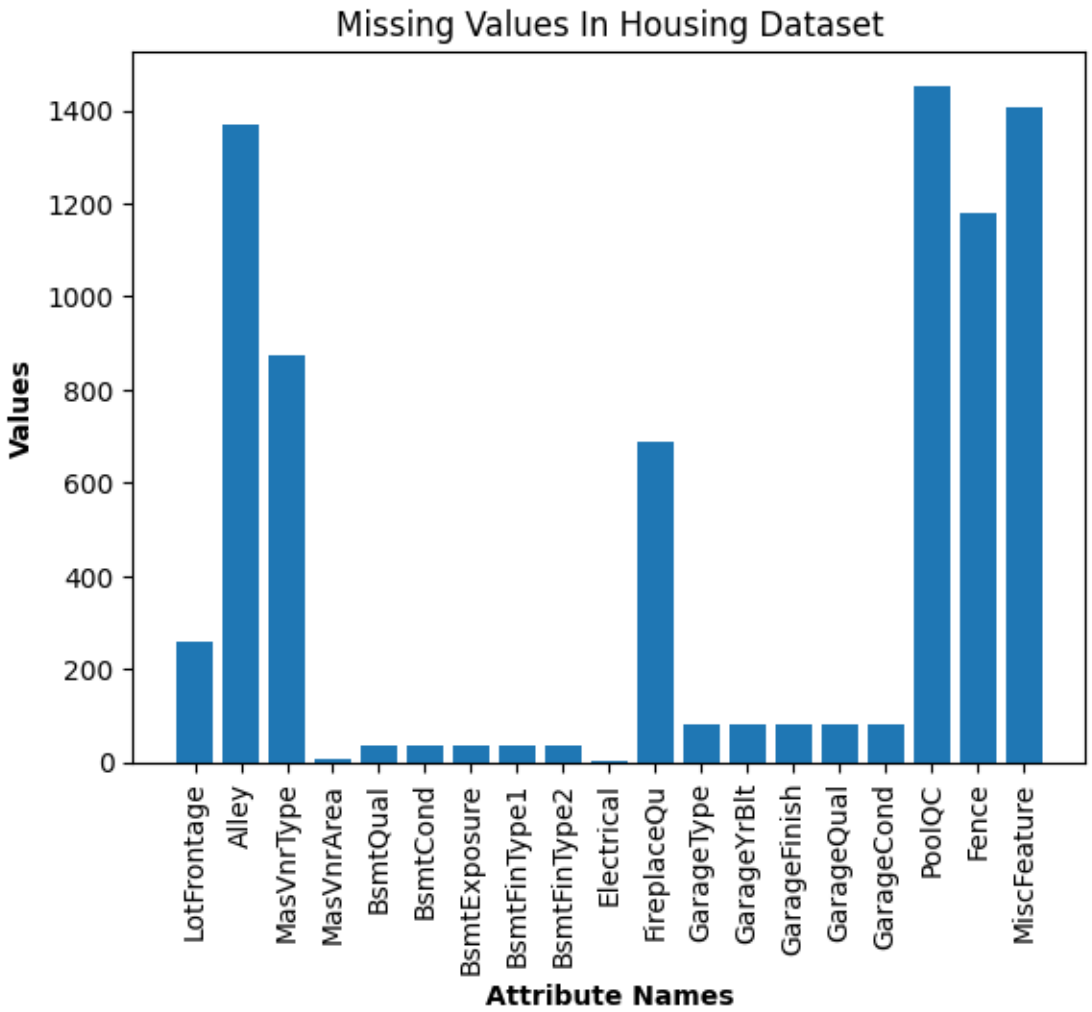

Figure 21: Number of Missing Data in the Columns of Housing Dataset

2. Can noise objects be outliers?
3. Are noise objects always outliers?
4. Are outliers always noise objects?
5. Can noise make a typical value into an unusual one, or vice versa?

Solution: According to [3], the main difference between noise and outlier is that noise is most likely an input error or mislabeling whereas, outlier is not an error but rather an underrepresented region in the dataset. However, it is very easy to misrepresent noise as an outlier as both have some similar characteristics. It is better to explain this phenomenon through an example. Suppose, we have the following age data of humans represented in years, $d_{age} = [15, 18, 23, 34, 500]$ and another data, weight of vehicles, $d_{weight} = [500, 1200, 900, 10000]$ represented in *kg's*. Now, for d_{age}, we can see that 500 is a noise as it is highly unlikely that a person's age is 500. It is possible that due to data insertion a 0 was added mistakenly. However, if we look at the data points of d_{weight} it is not entirely clear whether that is a noise or an outlier. It is mentioned the given weights are for vehicles but the type of vehicle wasn't mentioned so it is possible that although the first few data points refer to family cars, the last one might be an industrial vehicle. Another possibility is, of course, this data being a noise which was an error while input. So, depending on the context a datapoint that is not consistent with the dataset might or might not be an outlier or vice versa. However, noise can be easily identified if for example, our d_{age} contained a value such as "New York City". Here it is evident that this is a string and essentially the name of a city, hence, it is a noise. Apart from the scenario just mentioned for the noise, both noise and outlier can create significant issues while analysing the data. As it will skew the dataset remarkably and will affect pattern recognition tasks. With the increase in noise and/or outliers, the distribution of data becomes unstable. So, in short, although both of them contain similarities and dissimilarities, the presence of either can affect the understanding of the data on a huge scale.

Problem 8

Assume you have data $D = \{x_0, x_1, x_2, x_3\}$. Provide a listing of all the possible partitions.
Solution: Using bells triangle[6] we find the number of partitions to be 15 as listed below

1. Partition 1: $\{\{x_0, x_1, x_2, x_3\}\}$
2. Partition 2: $\{\{x_0\}, \{x_1, x_2, x_3\}\}$
3. Partition 3: $\{\{x_1\}, \{x_0, x_2, x_3\}\}$
4. Partition 4: $\{\{x_2\}, \{x_0, x_1, x_3\}\}$
5. Partition 5: $\{\{x_3\}, \{x_0, x_1, x_2\}\}$
6. Partition 6: $\{\{x_0, x_1\}, \{x_2, x_3\}\}$
7. Partition 7: $\{\{x_0, x_2\}, \{x_1, x_3\}\}$
8. Partition 8: $\{\{x_0, x_3\}, \{x_1, x_2\}\}$
9. Partition 9: $\{\{x_0\}, \{x_1\}, \{x_2, x_3\}\}$
10. Partition 10: $\{\{x_0\}, \{x_2\}, \{x_1, x_3\}\}$
11. Partition 11: $\{\{x_0\}, \{x_3\}, \{x_1, x_2\}\}$

Problem 8 continued on next page...

12. Partition 12: $\{\{x_1\},\{x_2\},\{x_0,x_3\}\}$
13. Partition 13: $\{\{x_1\},\{x_3\},\{x_0,x_2\}\}$
14. Partition 14: $\{\{x_2\},\{x_3\},\{x_0,x_1\}\}$
15. Partition 15: $\{\{x_0\},\{x_1\},\{x_2\},\{x_3\}\}$

Problem 9

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

$$tf'_{ij} = tf_{ij} \times \log \frac{m}{df_i},$$

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?
2. What might be the purpose of this transformation?

Solution: The inverse document frequency transformation, also known as $tf - idf$ transformation is a technique we used to find the effectness of a word in a document[11]. The given equation describes a lot about how the term can be affected in a document. For instance, we can see that given m-documents and several occurrences of a term df_i, the transformation is proportionate to its logarithmic value. Now, as a term occurs in just one document the effect of this transformation will be higher. However, as the term appears more and more in the document the effect lessens and goes to 0. This is ensured by the high term frequency and low document frequency in the entirety of the m-documents.

Problem 10

This question compares and contrasts some dissimilarity 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.

   - $\mathbf{x} = 0101010001$
   - $\mathbf{y} = 0100011000$

   Solution: Given, $x = 0101010001$ and $y = 0100011000$Now the Hamming distance of x and y is, 3 as there are 3 different elements while going through both of the arrays sequentially.
   The Jaccard similarity is defined by the formula[1],

   $$J(A,B) = \frac{A \cap B}{A \cup B} \tag{26}$$

   Where A,B denotes two vectors, $A \cup B$ denotes the total number of points between both of the vectors, and $A \cap B$ denotes the similar vector between them. So for Jaccard similarity we get, $x \cap y = 3$ and $x \cup y = 17$.
   Thus, $J(x,y) = 3/17 = 0.176$.

Problem 10 continued on next page...

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 similar, but don't let this confuse you.)
Solution: Simple Matching Coefficient(SMC), also known as Rand Similarity Coefficient generally serves as a similarity and diversity comparison method[10]. This is very similar to the Jaccard similarity coefficient as the method of finding the similarity in SMC follows the same pattern as Jaccards in finding the ratio of the total matched vector and the total number of vectors present.

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.)
Solution: Finding the most similar between two species accounts for a lot of questions. First of all, if we are comparing two species that belong to the same genre then it is likely that they share a lot of characteristics. An article[4] addresses how similar the species are because of the shared ancestors in the species tree. So, to efficiently find cross-species similarity it is better to use "Jaccard Similarity" instead of "Hamming Distance" as the former considers the shared genes as well as dissimilar genes. For example, let two species S_1 and S_2 have 1 similar gene and they contain a total of 2 genes (Assumption for the simplicity of the example). Now the "Hamming" distance would simply say that they have a distance of just 1, from which we might assume that they are not that similar. On the other hand, "Jaccard" similarity would consider the total number of genes and would result in saying that S_1 and S_2 are very similar. It can be further argued by extending the total number of genes between the species.

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.)
Solution: Sequence Alignment[9] is a popular method in bioinformatics to compare genetic similarity. We would prefer this method over the Jaccard similarity and Hamming distance because of the following reasons,

   - Dynamic Programming Solution so makes the calculation faster
   - Can easily handle different length sequences
   - Considers not only the similarities but also the substitutions of genomes
   - Can capture more semantics

Problem 11

JN III

For the following vectors, **x** and **y**, calculate the indicated similarity and distance measures. Show detailed calculations/steps. **Solution:** The result and calculations and Python implementation can be found for this problem in the file *jn3.ipynb*. The formula for Cosine, Correlation, Jaccard, and Euclidean is adapted from various sources[2] [8] [1] [7]. Kindly note that for correlations, we have used "Pearson's Correlation Coefficient" to find the correlations.

1. $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$ cosine, correlation, Euclidean.
Solution: The cosine similarity of the vectors was found to be 1. Looking at the vector we can see that each of the elements in the vector is differed by just 1. Now, given the definition of cosine similarity,

the result represents that the angle between the vectors is low. This denotes that the vectors are very similar. For correlation calculations, as the vectors don't have any variance in the Pearson's method is not able to calculate any correlations between them. Since correlations are calculated based on deviation and the deviations of the vectors are 0 a calculation is not possible. The result for Euclidean distance was found to be 2.

2. $\mathbf{x} = (0, 1, 0, 1)$, $\mathbf{y} = (1, 0, 1, 0)$ cosine, correlation, Euclidean, Jaccard.
Solution: The results for cosine, correlation, Jaccard, and Euclidean distance were found to be 0, -1, 2, 0. Looking at the vectors we can easily understand the results here. First of all, we can see that element-wise the vectors differ on every element. Thus, there is no similarity between them. Second of all, due to the difference in each unit, the correlation is -1. Last of all, the Euclidean distance we find is similar to *Problem a* but from the similarity index, we can determine that the vectors are not similar.

3. $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean.
Solution: Interestingly here, the results are 0 for cosine similarity, 0 for correlation, and 2 for distance. We can understand the result being 0 for similarity as the vectors differ by elements, however, the correlation being 0 changes the fact that we received -1 for the previous problem. Due to the nature of variance, we argue that the correlation achieved a different result. Finally, the distance seems to be unchanged given the variance has changed.

4. $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard.
Solution: For this problem, we see a slight difference in similarity for Jaccard and cosine similarity which is 0.5 and 0.75 respectively. Looking at the vectors and how Jaccard and cosine similarity is defined we argue that due to the nature of Jaccard's similarity of eliminating common elements between vectors, it generates lower similarity results than the other one. The correlations are again determined by the variance of the dataset and in this case the vectors.

5. $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0 - 1)$ cosine, correlation.
Solution: Due to the different valued elements in the vectors, we see a similarity score of 0 for this problem. For correlation, the value seems to be very low which again certifies the similarity factor.

References

[1] LearnDataSci. *Jaccard Similarity*. https://www.learndatasci.com/glossary/jaccard-similarity/#:~:text=The%20Jaccard%20similarity%20measures%20the,of%20observations%20in%20either%20set..

[2] Learndatasci. *Cosine Similarity*. https://www.learndatasci.com/glossary/cosine-similarity/#:~:text=The%20similarity%20can%20take%20values,the%20cosine%20similarity%20is%201..

[3] Cátia M. Salgado et al. "Noise Versus Outliers". In: *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, 2016, pp. 163–183. ISBN: 978-3-319-43742-2. DOI: 10.1007/978-3-319-43742-2_14. URL: https://doi.org/10.1007/978-3-319-43742-2_14.

[4] UCL. *Finding Shared Gene Species*. https://blogs.ucl.ac.uk/gee-research/2015/05/07/finding-shared-genes-species/.

[5] Davies University Of California. *Metric Properties*. https://www.math.ucdavis.edu/~hunter/m125a/intro_analysis_ch7.pdf.

[6] Wikipedia. *Bells Triangle*. https://en.wikipedia.org/wiki/Bell_triangle.

[7] Wikipedia. *Euclidean Distance*. https://en.wikipedia.org/wiki/Euclidean_distance.

[8] Wikipedia. *Pearson Correlation Coefficient*. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[9] Wikipedia. *Sequence Alignment*. https://en.wikipedia.org/wiki/Sequence_alignment.

[10] Wikipedia. *Simple Matching Coefficient*. https://en.wikipedia.org/wiki/Simple_matching_coefficient#:~:text=The%20simple%20matching%20coefficient%20(SMC,and%20diversity%20of%20sample%20sets.&text=value%201%2C%20and-,is%20the%20total%20number%20of%20attributes%20where%20A%20has,and%20B%20has%20value%200..

[11] Wikipedia. *TF-IDF Transformation*. https://en.wikipedia.org/wiki/Tf%E2%80%93idf.