**University of Central Missouri**

**Department of Computer Science & Cybersecurity**

**CS5710 Machine Learning**

**Fall 2025**

**Home Assignment 4.**

**Student name: Sunil Kumar Vuta**

**Submission Requirements:**

* Once finished your assignment push your source code to your repo (GitHub) and explain the work through the ReadMe file properly. Make sure you add your student info in the ReadMe file.
* Comment your code appropriately ***IMPORTANT.***
* Any submission after provided deadline is considered as a late submission.

**Part A: Calculation**

**Q1.** Find the cluster using the Average and MIN technique. Use Euclidean distance to build the complete distance matrix, updated the distance matrix to the final step and draw the dendrogram for each.

|  |  |  |
| --- | --- | --- |
|  | X | Y |
| P1 | 0.4 | 0.5 |
| P2 | 0.2 | 0.3 |
| P3 | 0.1 | 0.08 |
| P4 | 0.21 | 0.12 |
| P5 | 0.6 | 0.16 |
| P6 | 0.33 | 0.28 |
| P7 | 0.11 | 0.15 |

A notebook with writing on it

AI-generated content may be incorrect.

A piece of paper with writing on it

AI-generated content may be incorrect.

A notebook with writing on it

AI-generated content may be incorrect.

A piece of paper with writing on it

AI-generated content may be incorrect.

A notebook with writing on it

AI-generated content may be incorrect.

**A notebook with writing on it

AI-generated content may be incorrect.**

**Q2.** We have the following 2D data points:

Points: (2,1), (3,1), (3, 3), (4, 1), (5, 1), (6,7), (1,3), (2,5)

for K =3:

Centroid1: (2,1)

Centroid 2: (4, 1)

Centroid 3: (5, 1)

Euclidean Distance:

A math equation with black text

AI-generated content may be incorrect.

A notebook with writing on it

AI-generated content may be incorrect.

**Part B — Short-Answer**

**Q1.**

1. Describe agglomerative hierarchical clustering.

Ans:Agglomerative hierarchical clustering is a "bottom-up" approach to clustering. It starts by

treating every single data point as its own individual cluster. It then iteratively merges the two

clusters that are the most similar or nearest to each other based on a defined distance function.

This process continues until all data points are merged into a single, large cluster, forming the

root of the hierarchical tree

b) Describe divisive hierarchical clustering.

Ans: Divisive hierarchical clustering is a "top-down" approach. It begins with all data points

grouped together in one cluster . In each step, it splits the cluster into a set of child clusters. This

splitting procedure is applied recursively until only singleton clusters of individual data points

remain.

c) Which one is more commonly used and why?

Ans: Agglomerative clustering is more popular than divisive methods. The reason is generally

related to computational and conceptual simplicity. In agglomerative clustering, the primary task

is simply to find the minimum distance between existing clusters and merge them. Divisive

clustering requires a more complex method to determine the optimal way to split a cluster to

ensure the resulting sub-clusters are coherent, which can be computationally intensive

**Q2.**

1. To improve clustering quality, should inter-cluster distance be maximized or minimized?

Ans : The inter-cluster distance is the measure of separation between different clusters. By

maximizing this distance, you ensure that the clusters are as far apart as possible, meaning the

data instances in one cluster are very different ("far away") from the data instances in other

clusters. This is essential for good separation and a clear distinction between the identified

groups.

b) Same question for intra-cluster distance — explain the reasoning.

Ans : The intra-cluster distance is the measure of cohesion, or tightness, within a single cluster. It

measures how similar ("near") the data instances are to each other inside the same cluster. Byminimizing this distance, you ensure that all points belonging to a single cluster are very similar and tightly grouped, leading to high internal homogeneity.

**Q3.**

1. Define single link, complete link, and average link.

Ans :

**Single Link (MIN):** The distance between two clusters is defined by the minimum distance

between the closest pair of points, where one point is in the first cluster and the other is in the

second

**Complete Link (MAX):** The distance between two clusters is defined by the maximum distance

between the farthest pair of points, where one point is in the first cluster and the other is in the

second

**Average Link:** The distance between two clusters is defined as the average distance between all

pairs of points across the two clusters. This is typically calculated as the distance between the

cluster centroids

b) Explain one strength and one weakness of single-link clustering.

**Ans : Strength:**

Can discover non-elliptical shapes (such as elongated or curved clusters). Because it only

looks for the nearest neighbor between two clusters, it can "chain" points together to form

clusters of arbitrary shapes.

**Weakness:**

Susceptible to Noise (Chaining Effect): Single-link clustering is highly sensitive to noise

or outliers and suffers from the "chaining effect". A single point lying between two

otherwise well-separated clusters can act as a bridge, forcing the two distinct clusters to

merge prematurely at a very small distance

**Q4.**

1. What is the role of **tokenization** and give one example.

The primary role of tokenization is to break down a sequence of text (like a document,

paragraph, or sentence) into smaller, meaningful units called tokens. These tokens are the

fundamental building blocks used for further processing in natural language processing (NLP)

tasks.

Role: To transform unstructured text data into a format that can be analyzed by a machine

learning model.

Example:

Original Text: "Clustering is a great technique!"

Tokens (Word Tokenization):

"Clustering", "is", "a", "great", "technique", "!"

b) Compare stemming vs. lemmatization in terms of **speed** and **accuracy**.

|  |  |  |
| --- | --- | --- |
| FEATURE | STEMMING | LEMMATIZATION |
| SPEED | FASTER | SLOWER |
| ACCURACY | LOWER | HIGHER |
| REASONING | Uses a **crude heuristic**  **process** (set of simple rules)  to chop off word endings,  without checking if the  resulting root is a real word. | Uses a **linguistic process**  (vocabulary/dictionary and  morphological analysis) to correctly  identify the dictionary form of the word  (the lemma). |

**Q5.**

1. Explain what **word sense ambiguity** is and provide an example.

**a) Word Sense Ambiguity**

**Word sense ambiguity** occurs when a single word can have multiple distinct meanings or

interpretations, and the correct meaning can only be determined by its context within a sentence.

**Explanation**: A word is polysemous (has multiple related meanings) or homonymous (has

different, unrelated meanings but the same spelling). Without sufficient context, a computational

model cannot confidently determine which "sense" of the word is intended.

**Example**: The word **"bank"**

**Sense 1**: "I need to go to the **bank** to deposit a check." (Financial institution)

**Sense 2**: "We saw the deer drinking water near the river **bank**."

b) Explain why pronoun reference ambiguity can confuse a model.

**Pronoun reference ambiguity** (or anaphora resolution) confuses a model because the

pronoun's antecedent (the noun it refers back to) is not explicitly clear. Models struggle when a

pronoun, such as **"they"** or **"it"**, could logically refer to more than one noun in the preceding

text. For example, in the sentence, "The old city bus passed the tree because it was blocking the

street," the pronoun **"it"** could refer to either the **"bus"** or the **"tree."** Resolving this requires

the model to apply complex reasoning and world knowledge, which is a significant challenge for

purely syntactic or pattern-based models.

**Q6.**

1. Why can’t NLP tasks like POS tagging be solved by predicting each token independently?

NLP tasks like **Part-of-Speech (POS) tagging** cannot be solved by predicting each token

independently because the grammatical role of a word is often **mutually dependent** on the

words immediately preceding or following it in a sentence. A word may have multiple possible

parts of speech (e.g., "drive" can be a noun or a verb), and the correct tag depends entirely on thesurrounding context. Predicting each token in isolation would ignore these dependencies, leading

to syntactic errors and a low-accuracy analysis of the sentence structure. Models must utilize

information about adjacent tags to maintain **sequence coherence**.

b) Give one example where **decisions are mutually dependent** in a sentence.

A common example where tagging decisions are mutually dependent involves tagging sequences

that contain **homographs** (words spelled the same but having different tags).

**Sentence:** "Can you **can** the tuna?"

|  |  |  |
| --- | --- | --- |
| TOKEN | Possible Tags | Correct Tag (dependent Decision) |
| CAN | Noun/Modal Verb | Modal Verb (Followed by another verb, “can”) |
| CAN | Noun(The physical container)/Verb( To put in a container) | Verb (precede by a modal verb “can”) |
| TUNA | Noun (Object of the verb “can”) | Noun |

**Part C — Coding**

**Q1.**

Write Python code to perform the following steps:

1. Segment into tokens
2. Remove stopwords
3. Apply **lemmatization** (not stemming)
4. Keep **only verbs and nouns** (use POS tags)

Input text:  
"John enjoys playing football while Mary loves reading books in the library."

**Q2.**

Use Python and any NLP model to perform:

1. **Named Entity Recognition (NER)**
2. **Disambiguation prompt:**  
   If the text contains a pronoun ("he", "she", "they"), print:

"Warning: Possible pronoun ambiguity detected!"

Input text:  
"Chris met Alex at Apple headquarters in California. He told him about the new iPhone launch."