Lesson 8

Finding Similar Items, Collaborative Filtering, and Distance Measures for Similarity Analysis

Similarity Search

- Refers to a data mining method which helps in discovering items which have similarities in datasets using the Machine Learning algorithms
- Discovering interesting patterns
- Enables categorization and summarization of data and relationships among data

Finding Similar Items

- Finding similar excellent performance of students in Python programming
- Similar showrooms of a specific car model which show high sales per month
- Recommending books on similar topic such as in ‘Internet of Things’ by Raj Kamal from McGraw-Hill Higher Education, etc.

Nearest Neighbour Search (NNS)

- Finds that a point in a given set is most similar (closest) to a given point
- Less distant (closer) neighbour considered similar

NNS Algorithm

- Consider set $\mathbb{S}$ having points in a space M
- Consider a queried point q ∈ M, which means q is member of M.
- k-NNS algorithm finds the k-closet (1-NN) points to q in $\mathbb{S}$.

Dissimilarity function

- Having larger value means less similar
- Used to find similar items
- Greater distance means greater dissimilarity

Dissimilarity coefficient

- Relates to a distance metric in metrics space in v-dimensional space
- An algorithm computes squared Euclidean, Euclidean, Manhattan, or Minkowski distances [Refer Equations (6.20a) to (6.20d)]

Distance metric symmetry and triangular inequality

- Triangular inequality— Consider three vectors of lengths x, y, and z.
- Then, triangular inequality means $z < x + y$.

Triangular inequality Theorem

- Third side of a triangle is less than the sum of two other sides, and never equal
- Applies to v-dimensional space also

Asymmetric Dissimilarity

- Triangular inequality not true (Bergman divergence)

Jaccard Similarity of Sets

- $J(\mathcal{A}, \mathcal{B}) = (\mathcal{A} \cap \mathcal{B}) / (\mathcal{A} \cup \mathcal{B})$ ... (6.22)
- $\mathcal{A} \cap \mathcal{B}$ means the number of elements or items that are same in sets $\mathcal{A}$ and $\mathcal{B}$
- $\mathcal{A} \cup \mathcal{B}$ means the number of elements or items present in union of both the sets.

Similarity of Documents

- Compute Jaccard similarity coefficient method
- Latent Semantic Indexing method

Collaborative Filtering

- Refers to a filtering algorithm, which filters the items sets that have similarities with different items in a dataset
- Finds the sets with items having the same or close similarity coefficients

Distance Measures for Finding Similar Items or Users

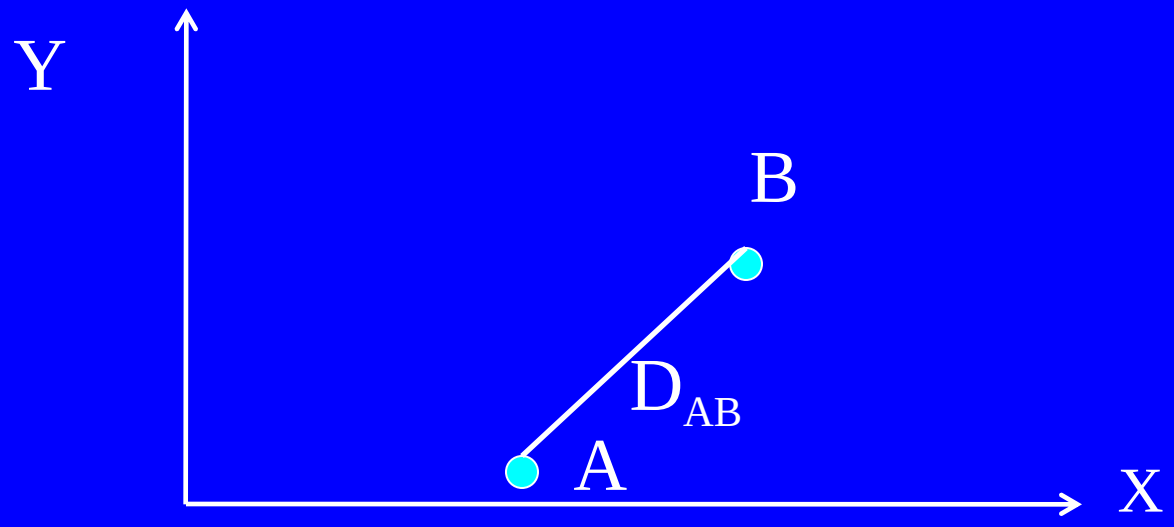
- Compute the dissimilarities
- Complement of dissimilarity gives similarity
- Distance can be defined as the reciprocal of weight in v-dimensional space.

D_{Eu}^2, D_{Eu}, D_{Ma}, D_{Mi}, and D_{Ha}

- Equations (6.20a to e)] or any other distance metric, for example, Jaccard distance D_{Ja}, cosine distance D_{Cos}, edit distance D_{Ed}.

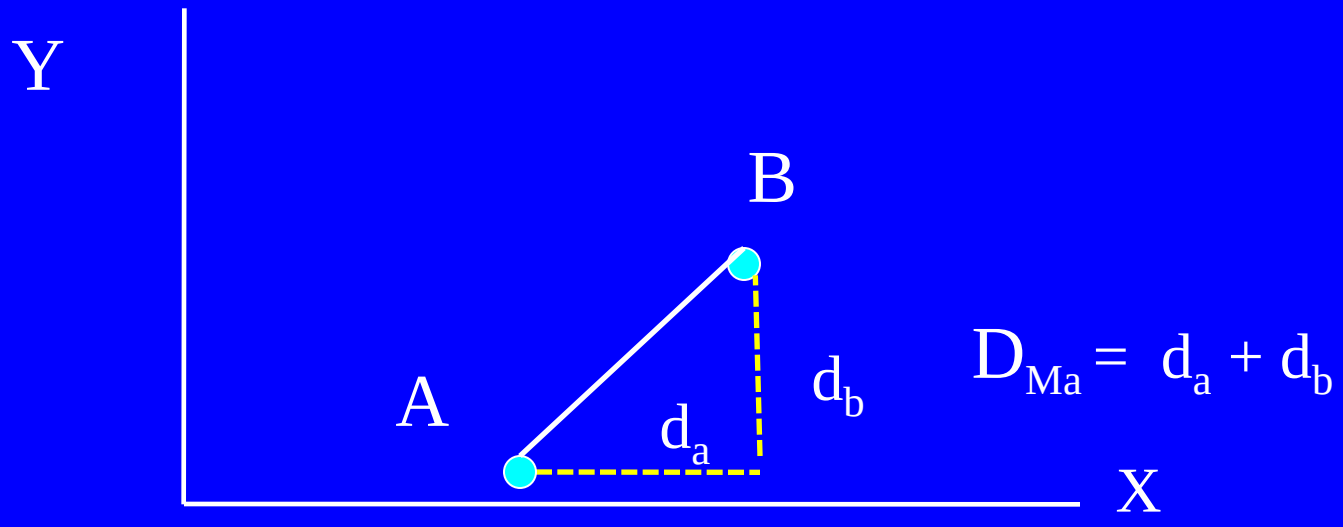
Euclidean D_{Eu}

- In terms of distance between two data-points A and B (Equations 6.20a and 20b)



Manhattan Distance D_{Ma}

- In terms of sum of axial distances between two data-points A and B (Equations 6.20a and 20b)

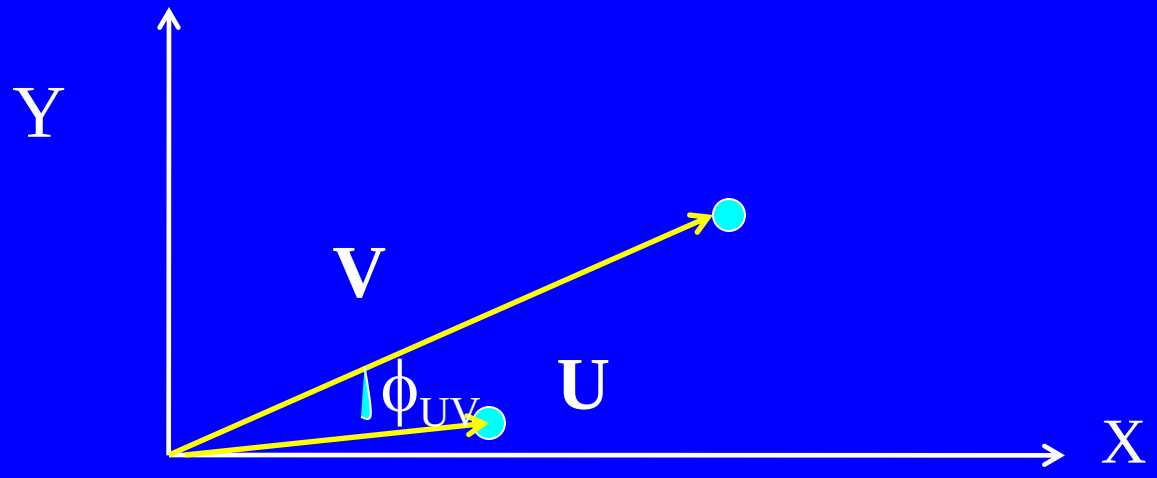


Distances D_{Ja}, D_{Cos}, *and* D_{Ed}

- Jaccard distance D_{Ja},
- $D_{Ja}(\mathcal{A}, \mathcal{B}) = 1 - J(\mathcal{A}, \mathcal{B})$ ...(6.23)
- Cosine distance D_{Cos}, [Equation (6.23a)]
- Edit distance D_{Ed}. [a distance measure for dissimilarity between two set of strings or words]

Vector Cosine-Based Similarity

- In terms of angle ϕ_{UV} between two vectors U and V (Equation 6.23b)
- $\phi_{UV} = \cos^{-1}(D_{Cos})$



Distance D_{Ed}

- Equals the minimum number of inserts and deletes of characters needed to transform one set into another
- Applications oin text analytics and natural language processing, similarities in DNA sequences [DNA sequences are strings of characters.]

Distance D_{Ha}

- If both U and V are vectors, Hamming distance D_{Ha} equals to the number of different elements between these two vectors

Summary

We learnt:

- Similarity and Dissimilarity Coefficients
- Distance measure related to Dissimilarity
- Jaccard, Squared Euclidean, Euclidean, Manhattan, Minkowski, Hamming and Edit Distances
- Cosine distance and Cosine Similarity
-

End of Lesson 8 on Finding Similar Items, Collaborative Filtering, and Distance Measures for Similarity Analysis