AI Searching Techniques

Measuring Similarity and Distance

Dr.Sherif Saad

Last-Time

- A* Search Algorithm
- Greedy Algorithm

Learning Objectives

- Understanding and building simple similarity measures.
- Study complex similarity measures and algorithms
- Applications of similarity measures in AI

Outlines

Meaning of Similarity

Types of Similarity Measures

Virtual Attributes

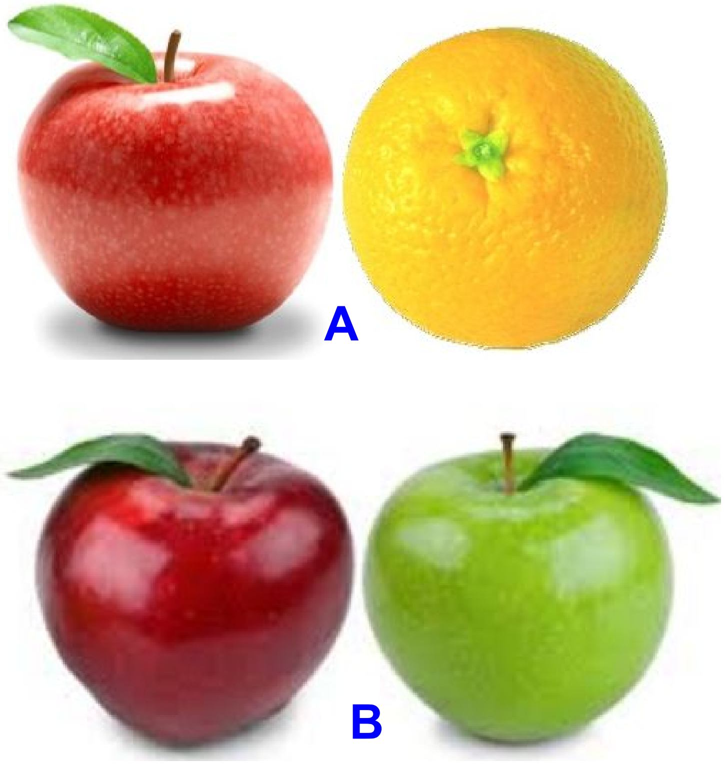
Why Should we Care about Similarity

- Similarity is the essential function in many AI branches.
- Measuring similarity or distance between two objects is fundamental to many Machine Learning algorithms such as K-Nearest-Neighbor, Clustering, etc.
- Measuring similarity or distance is important when designing heuristics.
- Measuring similarity or distance is important when building recommendation and expert systems.

Similarity is not an Exact Concept

A minimal but essential requirement for defining a similarity measure is that the data structures of both object and measure be compatible.

Simple representations need simple similarity concepts only, while complex representations require more effort



Similarity Vs Equality

- Similarity depends on the context to a much higher degree than equality
  - Similar with respect to what?
  - For instance, when are two cars similar?
- Equality is crisp, “yes or no”, while similarity is in degrees
  - Can we say that two objects are more equal than two other ones?

Meaning of Similarity

- Two objects are considered as similar if they look or sound similar
- Similarity is the measure of how much alike two objects are
- Similarity is not nearly as clearly defined as the term equal
- Similarity is a subjective concept.
- Similarity concepts are defined for the comparison of objects.
- We can think of similarity as a relation or as a function
- similarity is closely connected to the neighbour concept, the most similar object y to a given object x will be called a nearest neighbour

Similarity as a Relation

We can use binary relation to represent similarity between objects

$$\mathrm{SIM}(x, y) \Leftrightarrow \text{“}x \text{ and } y \text{ are similar”}$$

$$\mathrm{DISSIM}(x, y) \Leftrightarrow \text{“}x \text{ and } y \text{ are dissimilar”}$$

$$R(x, y, z) \Leftrightarrow \text{“}x \text{ is at least as similar to } y \text{ as } x \text{ to } z\text{”}$$

Similarity as a Function

- Expressing numerically how similar two objects are
- Assigning a degree of similarity to two objects
- more detailed and more expressive but also more difficult design and implement
- For a given object **O**, *a nearest neighbour is an Object* **O'** that has maximal similarity among the available objects
- Similarity is a value between [0,1]

$$\mathrm{sim} : U \times V \rightarrow [0, 1]$$

Types of Similarity Measures

Counting similarities: Certain occurrences in the representation are counted (with possible weights). One can, for instance, count the number of members in a family for tax reasons.

Metric similarities: They arise as variations of Euclidean metrics. This is closest to the travel view. It is justified if the metric mimics the difference between the object

Types of Similarity Measures (cont...)

Transformation similarities: Here one measures how costly it is to transform the first object into the second one

Structure-oriented similarities: The structure in which the knowledge is presented plays a role

Information-oriented similarities: The information and knowledge contained in an object plays an essential role.

Example: Buying a Used Car

Find a used car with

X = (price = 30,000, #seats = 2, max speed = 150 mph, colour = blue).

Available Cars

- Car 1: (price = 50,000, #seats = 2, max speed = 150 mph, colour = blue).
- Car 2: (price = 20,000, #seats = 4, max speed = 120 mph, colour = red).
- Car 3: (price = 20,000, #seats = 2, max speed = 120 mph, colour = red).

Importance (e.g., price: high; #seats: low; max speed: medium; colour: very low).

Counting Similarities

- Certain occurrences in the representation are counted
- Examples:
    - Hamming Measures
    - scalar (or dot) product
    - Tversky Measures
- How can we use counting similarities to solve the car example?

Counting Similarities (cont...)

Hamming Measures

$$H\big((x_1, \ldots, x_n), (y_1, \ldots, y_n)\big) := \sum (i \,|\, x_i = y_i, 1 \le i \le n).$$

- H (x, y) ∈ [0, n] and n is the maximum similarity, which means that H (x, y) is the number of agreeing attribute-value.
- H is a symmetric and reflexive
    - H (x, x) = 0 (symmetric)
    - H (x, y) = H (y, x) (reflexive)

Counting Similarities (cont...)

Scalar (or dot) Product

$$S(x, y) := \sum_{i=1}^{n} x_i \cdot y_i$$

- The scalar product is often used in pattern recognition as a simple measure. For binary attributes we distinguish two cases:
  - The values are 0 and 1: Then only values 1 contribute to the accumulated similarity.
  - The values are −1 and +1: Then, in addition, non-agreeing values give a penalty to the measure

Metric Similarities

- They are applicable to attributes with numerical values and closely related to numerical distances
- If symbolic values are present they have first to be numerically coded.
- Example:
    - Manhattan Metric $d_c(x, y) = \sum |x_i - y_i|.$
    - Euclidean Metric $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$
    - Minkowski Distance
- How can we use counting similarities to solve the car example?

Structured Similarities and Symbolic Arguments

- Symbolic arguments are often structures (e.g taxonomy)
- Symbolic arguments often are coded by numerical values (increases efficiency at runtime)
- For attributes A with unstructured symbolic values {v 1 , v 2 , . . . , v k }, there is no other way for defining measures than using tables(matrics)
- The more a structure is defined on the symbolic values, the more systematically can the similarity measures be defined.

Transformational Similarities

- Counts the number of changes needed to transform one object into another one.
- Example
  - Levenshtein (Minimum Edit) distance = (insertion, deletion, and modification)
  - converting induction ⟶ deduction
- We can assign weights for the different operation to represent the cost of these operations

Virtual Attributes

Let assume that the following table show reliability for getting a loan from a bank

Query	Case 1	Case 2	Case 3
Income 2000 Spending 1500	Income 1000 Spending 1500	Income 2000 Spending 5000	Income 6000 Spending 4500
Reliable?	No	No	Yes

Virtual Attributes (cont...)

An attribute that is not explicitly define in the data (problem)

It is usually derived or defined from other attributes

A large problem is to find and define virtual attributes.

Next Time

- The Local-Global Principle for Similarity Measures
- Graph Representations and Graph Similarities
- Taxonomic Similarities
- OOR Similarities

Questions