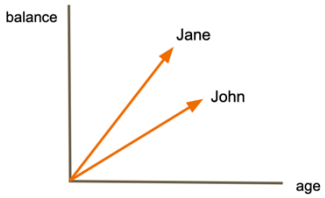
Data representation + Distance & Similarity

How to represent data?

Records:

- m-dimentional points or vectors
  e.g. (name, age, balance) -> (“John”, 20, 100)



- Graphs: nodes linked by edges
- Images: composed by pixels
- Text
- Strings: e.g. DNA sequence
- Time series: list of data at a specific intervals of time

Learning Type:

Supervised Learning (vue en CS542): learning a function that maps an input to an output based on example input-output pairs.

- Regression
- Classification

Unsupervised learning (vue en CS565): the algorithm is not provided with any pre-assigned labels or scores for the training data. As a result, unsupervised learning algorithms must first self-discover any naturally occurring patterns in that training data set.

- Clustering

Distance & Similarity

Feature Space: a **feature** is an individual measurable property; feature space refers to the n n-dimensions where your variables live

Distance:

used to compare data points

distance function:

A metric on a set X is a function (called *distance function* or simply *distance*)

$$d : X \times X \to [0, \infty),$$

where $[0, \infty)$ is the set of non-negative real numbers and for all $x, y, z \in X$, the following three axioms are satisfied:

1. $d(x, y) = 0 \Leftrightarrow x = y$ identity of indiscernibles
2. $d(x, y) = d(y, x)$ symmetry
3. $d(x, y) \leq d(x, z) + d(z, y)$ triangle inequality

- Minkowski Distance
  - p = 2, Euclidean Distance

Length of a line segment between the two points.

- p = 1, Manhattan Distance
  Sum of the absolute differences of their Cartesian coordinates

Similarity:

Cosine similarity:

s(x, y) = cos(θ) two points x, y with θ angle between them,
the larger s(x, y) is, the similar x and y are
=> dismilarity function d (x, y) = 1 / s(x, y) OR d (x, y) = k - 1 / s(x, y)
!! use cosine similarity rather than euclidean distance, WHEN DIRECTION matters more than MAGNITUDE

Jacard similarity:

used Mahattan distance
two points x, y different features x_i and $u=y_i$ counts 1, otherswise 0

$$JSim(x,y) = \frac{|x \cap y|}{|x \cup y|} \quad JDist(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$