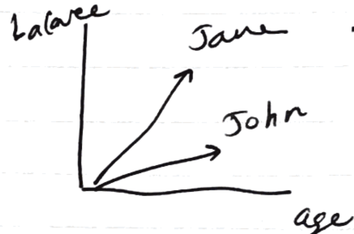
CS506 - LECTURE 4 (2/8) — HW0 due tonight
— Spark! request due tonight

- Introduction to Data Science
  - how we represent data linked to info we draw

★ $\boxed{\text{records}}$ → m-dimensional points / vectors → just m tuples

ex. name, age, balance

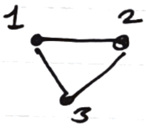


→ each person is a vector

(~~data~~. "John", 20, 100)

★ $\boxed{\text{Graphs}}$ → nodes connected by edges

ex. social networks



encode nodes & connection →

Adjacency matrix → each node represented by specific col. & row

→ symmetric = redundancy

$$\begin{array}{c|ccc} & ① & ② & ③ \\ \hline ① & 0 & 1 & 1 \\ ② & 1 & 0 & 1 \\ ③ & 1 & 1 & 0 \end{array}$$

* no connection with itself

0 → no connection between node & itself

each node is specific col / row of matrix

* generally nodes are not connected to themselves

→

Adjacency list ← more space efficient

1: {2, 3}
2: {1, 3}
3: {1, 2}

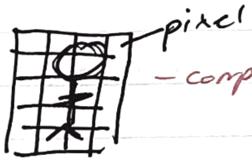
\# list the nodes it's connected to

- how info flows is directly a function of those edges / which nodes break up the network if it's disconnected?

★ Data Representation — $\boxed{\text{Images}}$

- greyscale values (0, 1)
- color rgb values

→



- computer = matrix of pixels

★ Data Representation — $\boxed{\text{Text}}$ — interested in the context

↳ split it up into words/sentences? → list of words → roots of words that capture same concept

ex. list of characters: DNA sequence

ski, skiing, skier — all same concept

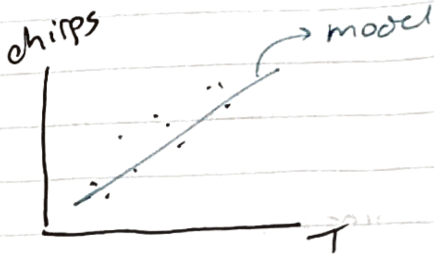
★ Data Representation — $\boxed{\text{Time Series}}$

(data @ specific intervals of time
anything of the above

· Types of learning (Supervised vs. Unsupervised)

Supervised Learning

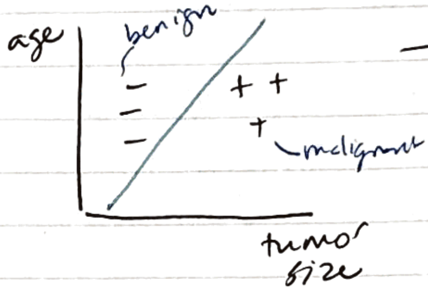
chirps	temp
10	40
5	37
17	53
55	103
40	78



goal:
estimate chirps from T
estimate ~~ch~~ T from chirps

* classification : relationship between two variables

age	size	mal.
20	12	0
23	15	1
47	20	1



→ can classify as malignant or benign

- get a new point - you can say w/ reasonable certainty what class it belongs to

Unsupervised Learning → goal: not to predict/classify

study data structure - can you identify patterns?

Ex. clustering → try to group points together → find a higher level feature

Dataset : collection of articles, are those articles covering the same topics?

---

* Distance & Similarity

Data :

$$n \text{ records} \left\{ \overbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ \vdots & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}^{m} \right.$$

- each column is a feature (m distinct attributes)

feature space: all possible values for set of features in data

· Distance — to uncover structure from data, need a way to compare data points

(dissimilarity function — takes two objects
returns LARGE value for ↑ dissimilarity (w/ respect to the function)

"d" is a distance function iff: ① $d(i,j) = 0$ iff $i = j$

② $d(i,j) = d(j,i)$ # symmetry

triangle inequality
- if you go through third point distance between i,j is necessarily smaller than i,k → k,j

③ $d(i,j) \leq d(i,k) + d(k,j)$

· Why a distance function?
it's intuitive → always want to get to a place where it's more intuitive

The extra restrictions tend to this

Minkowski Distance — For x,y points in d-dimensional space
↳ d attributes

$\underbrace{x = [x_1 \dots x_d]}_{d\text{ \# of features}}$ and $\underbrace{y = [y_1 \dots y_d]}_{d\text{ \# of features}}$

$p \geq 1$

p is a parameter (something to tune)

$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

Minkowski distance

and raise to p^{th} power

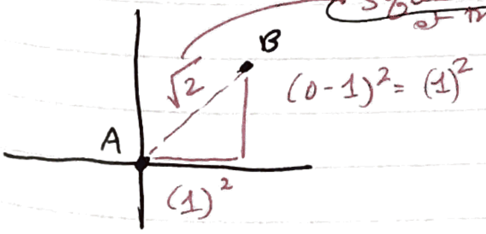
take the p^{th} root of the thing

take pairwise difference between x_i and y_i

when $p = 2$ → Euclidean Distance
when $p = 1$ → Manhattan Distance

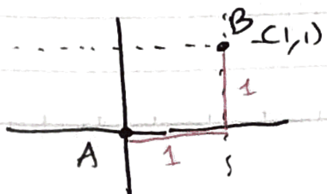
Ex. $d = 2$, $p = 2$ (Euclidean distance)

square root of the sum → p=2 (Pythagorean!!)



$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

$p = 1$

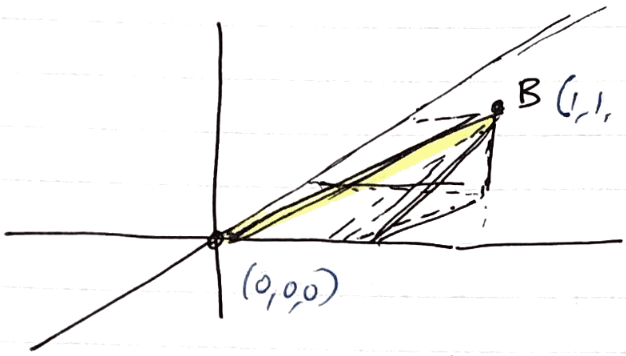


sum = 2 → distance must go through the grid (Manhattan Distance)
↳ sum along the grid

→ see slides
copy carefully

Ex. $d = 3$ → add another term to the sum

$\boxed{p = 2}$

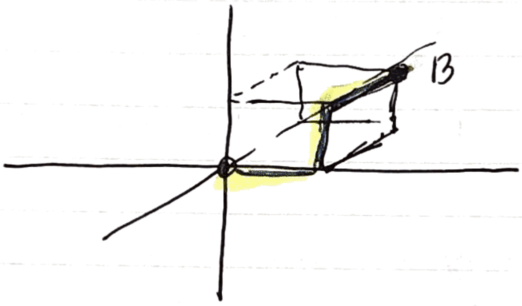


→ length?

$\sqrt{3}$

$(1^2 + 1^2 + 1^2)^{1/2} = \sqrt{3}$

$\boxed{p = 1}$

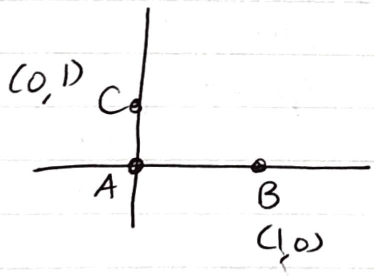


→ length?

$(1 + 1 + 1)^{1/1} = 3$

Is L_p a distance function when $0 < p < 1$?
— no because smaller distances would have greater value

recall: important axiom is triangle inequality
↳ proof by contradiction (counter-example)



$D(B, A) = D(A, C) = 1$ ✓

$D(B, C) = 2^{1/p}$

But... if $p < 1$, then $1/p > 1$

So $D(B, C) > D(B, A) + D(A, C)$
which violates the triangle inequality

- Cosine similarity — similarity function gives larger #'s for more similar objects

$s(x, y) = \cos(\theta)$

↳ θ is angle between x, y

∴ points further apart have very large angle θ (less similar)

↳ to get dissimilarity...

$d(x, y) = \frac{1}{s(x, y)}$ OR $d(x, y) = k - s(x, y)$ for some k → shift is ten here

↳ high similarity
$\cos(0°) = 1$ $\boxed{k = 1}$

$1 - 1 = 0$
↳ no dissimilarity when same point

Pull Request
↳ NOT many ASAP

→ when should you use cosine (dis)similarity over euclidean distance?

↳ when not interested in the magnitude of your vectors

Jaccard Similarity → represents documents

	w_1	w_2	...	w_d
x	1	0	...	1
y	1	1	...	0

if they have same word it's zero

one has it ↳ 1 ⟷

↳ apply distance ↳ manhattan distance

$$L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i| \quad p=1$$

↳ count # of words that the documents differ by

→ only be 1 if $x_i \neq y_i$

BUT... consider...

	w_1	w_2	w_3	w_4
x	1	1	0	1
y	1	1	1	0

only differ by last 2 words

	w_1	w_2
x	0	1
y	1	0

completely different

— manhattan distance for BOTH = 2 — need to account also for SIMILARITY (intersection)

☆☆ Jaccard Similarity → accounts for size of intersection

$$J_{sim}(x,y) = \frac{|x \cap y|}{|x \cup y|}$$

→ ratio between intersection and the union

if $x=y$, union = intersection, Jsim = 1 / if $x \neq y$, intersection = 0, Jsim = 0 — large value if similar

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

* try to prove this is a distance function — prove $d(i,j) \leq d(i,k) + d(k,j)$