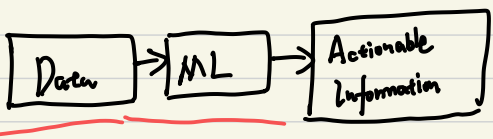
L1
CHINMAY HEGDE (professor)

Statistics / algorithm — ML — optimization / Application

What is ML:



(best used when we don't know the rule of things)

- Data representation
- Measure of goodness "Loss function"
- Method to optimize for measure "training algorithm"

Today: Data Representation

Data is a list of attributes that is collected about an object/phenomenon is interest.

Eg. Weather data

wind speed (w) · temperature (t) · pressure (p) · humidity (h) · air quality (a), etc.

[w(1), t(1), p(1), h(1), a(1)]
[w(2), t(2), p(2), h(2), a(2)]
⋮

d attributes ⟶ tuple of size d

⟶ vector in d-dimensional space

vector space:

① Collection of vectors which satisfy 2 properties:

a) addition

$x = (x_1, \dots, x_d)$

$y = (y_1, \dots, y_d)$

$x+y = (x_1+y_1, \dots, x_d+y_d)$

b) scalar multiplication 数乘

$x = (x_1, \dots, x_d)$

$\alpha x = (\alpha x_1, \dots, \alpha x_d)$

Examples:

1) weather resolution × RGB 个数值变量

2) image 2048 × 1536 × 3 (= d)

↳ R^d

3) stocks d = 500 R^{500}

② Properties of vector space

1) Dot product / inner products

$x, y \in R^d$

$\langle x, y \rangle = x_1y_1 + x_2y_2 + \dots + x_dy_d = \sum_{i=1}^{d} x_iy_i$

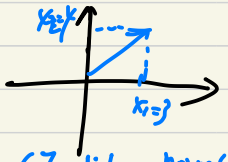
2) Cartesian products / outer products 所有的组合可能

$$x \otimes y = \begin{pmatrix} x_1y_1 & x_2y_1 & & x_dy_1 \\ x_1y_2 & x_2y_2 & & \\ & & & \vdots \\ x_1y_d & x_2y_d & & x_dy_d \end{pmatrix}$$

3) Norms

$X=(3,4)=(x_1, x_2)$

$\|X\|_2 = \sqrt{x_1^2+x_2^2} = \sqrt{3^2+4^2} = 5$



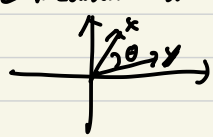
(Euclidean norms)

$\|X\|_1 = |x_1|+|x_2| = 3+4=7$ (L1 norms)

(Lp norms)

4) similarity (how we measure similarity?)

$\frac{\langle x, y\rangle}{\|x\|_2+\|y\|_2} = \cos\theta$



Application: NLP (natural language processing)

problem: Given a dataset of n documents $\{D_1, D_2, \cdots, D_n\}$ and a query document D^*, find the closest document to D^* in the dataset.

Solution:

Step 1 $d = 80,000$ (# of words in English)

$D_i \longrightarrow x_i \in R^d$ (d 种字) → 各种字数的集合

每种出现次数 $x_{i(j)}$ = # times word j appears in document i

$D \rightarrow [x_1, x_2 \cdots, x_n]$ (document)

Step 2 Define cosine similarity

$\cos\theta_i = \frac{\langle x_i, x^*\rangle}{\|x_i\|_2 \cdot \|x^*\|_2}$

cos越大，越相似

Step 3 $i^* = \arg\max_j \cos\theta_j$ $\cos\theta_i = 1$ 一模一样

["nearest neighbour search"] 夹角为0