ECE-GY 6143 INTRO TO ML

Applications of ML

- Recommendation systems
- Autonomous driving.
- Stock forecasting
- Image recognition
- Gaming.
- Voice recognition.
- Translation.

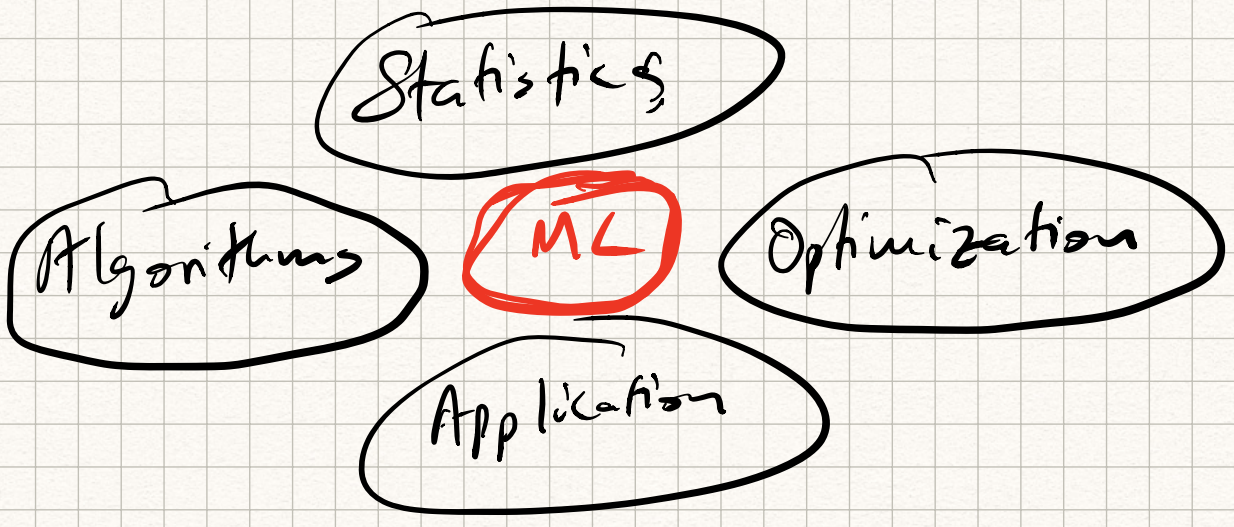
- Online ads.
- Auto complete.
- Content filtering.

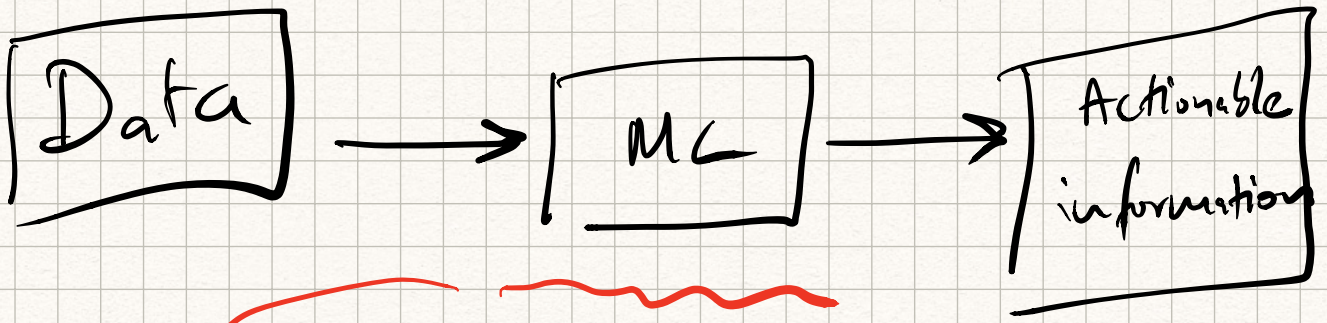
CHINMAY HEGDE

CHIMNEY HEDGE

This course: variety of ML

topics from a foundational & applied perspective.





- Data representation.
- Measure of goodness "Loss function"
- Method to optimize for this measure. "Training algorithm"

Data representations

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

E.g. Weather data

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

$$[w(1), t(1), p(1), h(1), a(1)]$$
$$[w(2), t(2), p(2), h(2), a(2)]$$
$$\vdots$$

d-attributes $\longrightarrow$ tuple of size d.

$\longrightarrow$ Vector in d-dimensional space.

Vector spaces

Collections of vectors which satisfy 2 properties

a) Addition.

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

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

$x + y = (x_1 + y_1, x_2 + y_2, \cdots, x_d + y_d)$

b) - Scalar multiplication.

$x = (x_1, \cdots, x_d)$.

$\alpha x = (\alpha x_1, \cdots, \alpha x_d)$.

Examples:

1) Weather.

2) Images. $2048 \times 1366 \times 3 \ (= d)$

↳ $\mathbb{R}^d$.

3) Stocks. $d = 500$ $\mathbb{R}^{500}$.

....

Properties of vector spaces.

1) Dot products / inner products.

$x, y \in \mathbb{R}^d$

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_d y_d$$

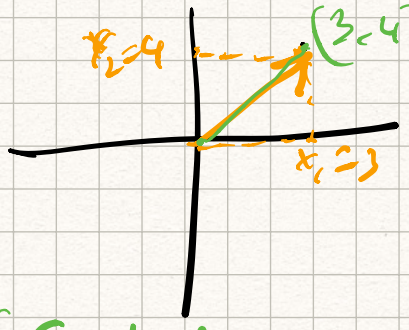
$$= \sum_{i=1}^{d} x_i y_i .$$

2) Cartesian products / outer products.

$$x \otimes y = \begin{pmatrix} x_1 y_1 & x_2 y_1 & \cdots & x_d y_1 \\ x_1 y_2 & x_2 y_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_1 y_d & x_2 y_d & & x_d y_d \end{pmatrix}$$

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 norm) or L2 norm.

$$\|x\|_1 = |x_1| + |x_2|$$

$$= 3 + 4$$

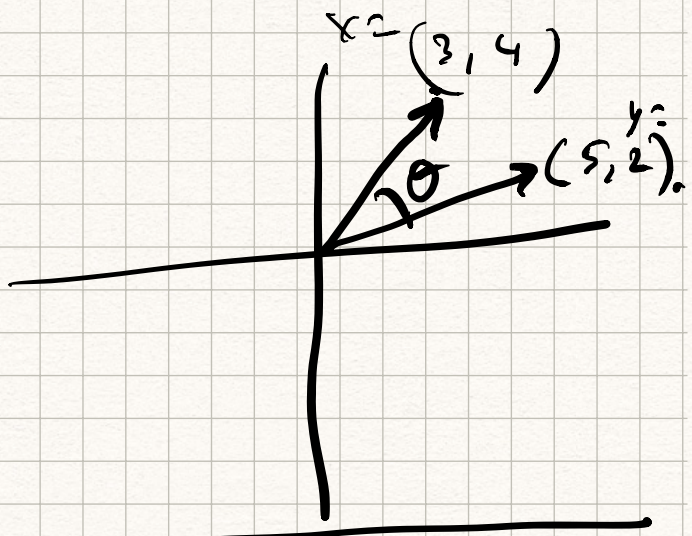
$$= 7$$

(L1 norm).

(Lp norms).

4) Similarity.

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



Application : NLP (Natural Language processing).

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

Solution

$d = 80{,}000$ (# in words $\simeq$ in English language)

$D_i \longrightarrow x_i \in \mathbb{R}^d$

Step 1 $x_i(j) =$ # times word j appears in document i.

$\mathcal{D} \longrightarrow \{x_1, x_2, \ldots, x_n\}$.

Define cosine similarity

Step 2 $\cos\theta_i = \dfrac{\langle x_i, x^* \rangle}{\|x_i\|_2 \, \|x^*\|_2}$.

Step 3 $i^* = \arg\max_i \cos\theta_i$

["Nearest neighbor search"].