PCA REVISION AND SUPERVISED ML

	Murder	Assault	Population
Wisconsin	$\frac{100-50}{5}$	10	10,000
California	$\frac{1000-50}{5}$	100	100,000
Vermont	$\frac{10-50}{5}$	5	1,000
⋮			

↓ normalize

$$\frac{x - \text{mean}}{\text{s.d.}}$$

CONCEPT

PCA is a linear model

LOADINGS (9) (10)

$$PC_1 = \phi_1 \times \text{Murder} + \phi_2 \times \text{Assault} + \phi_3 \times \text{Population}$$

(2)

Wisconsin (say without normalization)

$$PC_1 = 10 \times 100 + 9 \times 10 + 2 \times 10{,}000$$

$$= 1000 + 90 + 20{,}000$$

$$= 21{,}090$$

(2) ← LOADINGS → (3)

$PC_2 = \Phi_4 \times$ Murder $+ \Phi_5 \times$ Assault $+ \Phi_6 \times$ Population

(12)

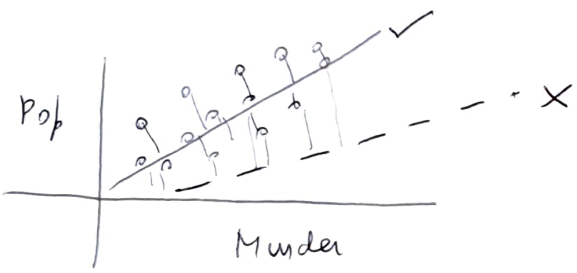
CONCEPT - SAME EQUATION FOR ALL STATES

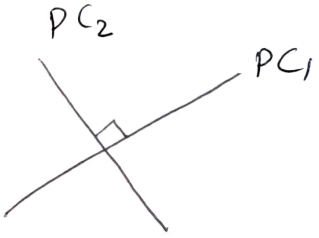
Wisconsin

$PC_2 = 2 \times 100 + 3 \times 10 + 12 \times 10,000$

$= 200 + 30 + 120,000$

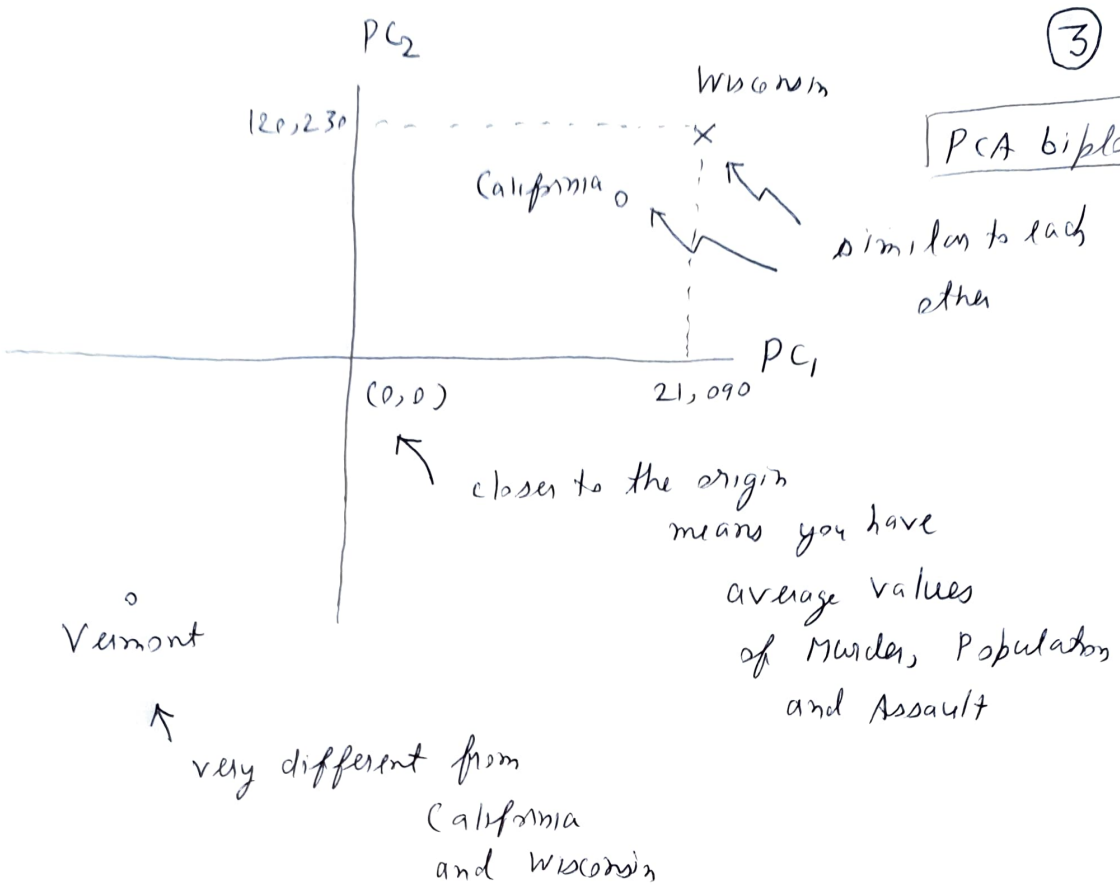
$= 120,230$





CONCEPT All principal components are at 90° to each other

orthogonal / orthogonal



PCA biplot

We can also detect outliers using PCA biplot

Vermont has negative values, because after normalizing, you can have less than average values e.g. $\frac{10-50}{5} = \frac{-40}{5} = -8$

INTERPRET

→ In order to have high values for PC_1

$$PC_1 = 10 \times M + 9 \times A + 2 \times P$$

↑ ↑

both Murder and Assault need to be high OR P

→ High "weight" on Murder, Assault
- low "weight" on Population
- PC_1 is a "crime" variable

→ Similarly PC_2 is an "urbanization" variable

PCA is not linear regression

CONCEPT

	Murder	Assault	Pop	y ← Number of people moving to that state last year
Wisconsin	1000	10	10,000	100
California	1000	100	100,000	1000
Vermont	10	50	1000	10
North Dakota	10	5	1000	(?)

→ predict this how?

$$y = a \times \text{Murder} + b \times \text{Assault} + c \times \text{Pop} + d,$$

(linear model / linear regression)

compare to

$$PC_1 = \phi_1 \times \text{Murder} + \phi_2 \times \text{Assault} + \phi_3 \times \text{Pop}$$

CONCEPT

- y is known BEFORE we do regression (label / supervision)
- Linear regression is a predictive model
- PC_1 is not known before we do PCA
  PCA is not a predictive model

Non-linear PCA?

$$PC_1 = \Phi_1 \times Murder' + \Phi_2 \times Assault' + \Phi_3 \times Population'$$

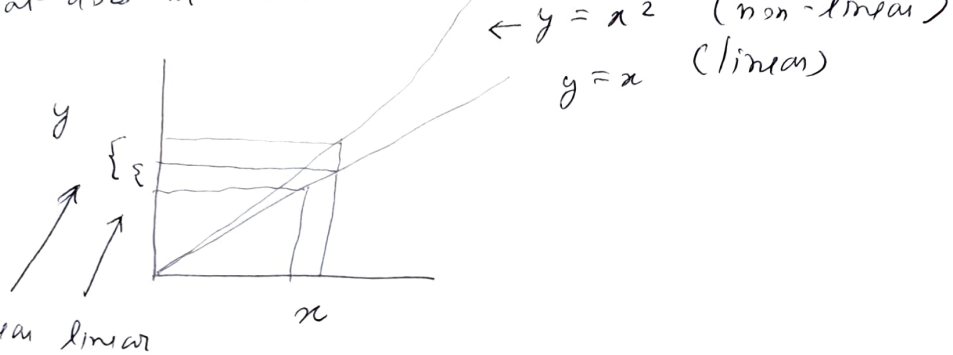
how do we make it non-linear?

log Murder

$Murder^{1.5}$

$Murder^2$

What does it mean to make it non-linear?



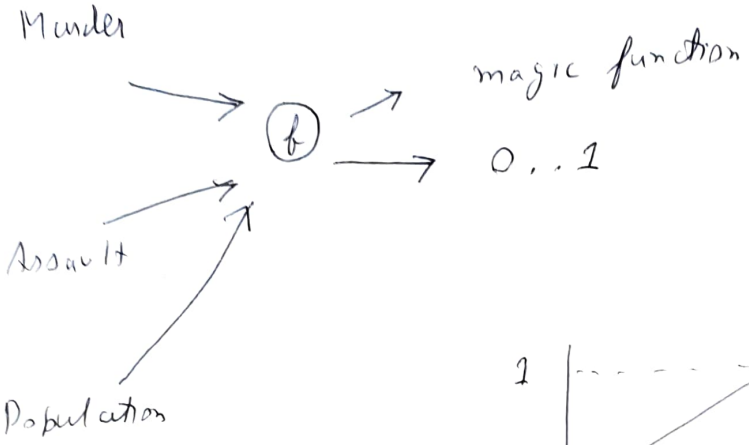
non-linear → for one unit increase in x, you get more/less than one unit increase in y.

linear → for one unit increase in x, you get one unit increase in y

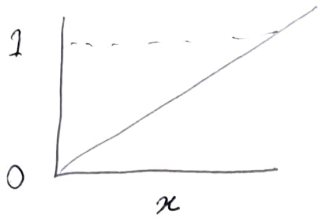
Can we make this general?

i.e., make a general, powerful non-linear function

ANN



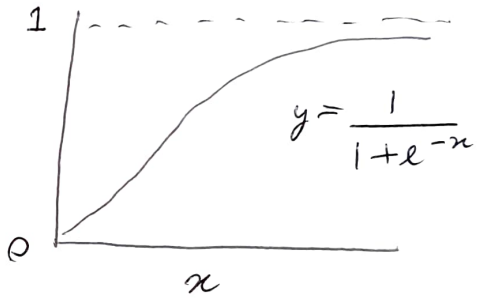
What is f ?



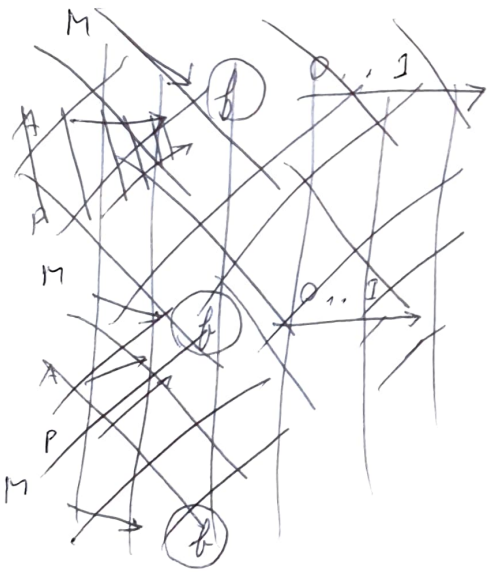
x will not do

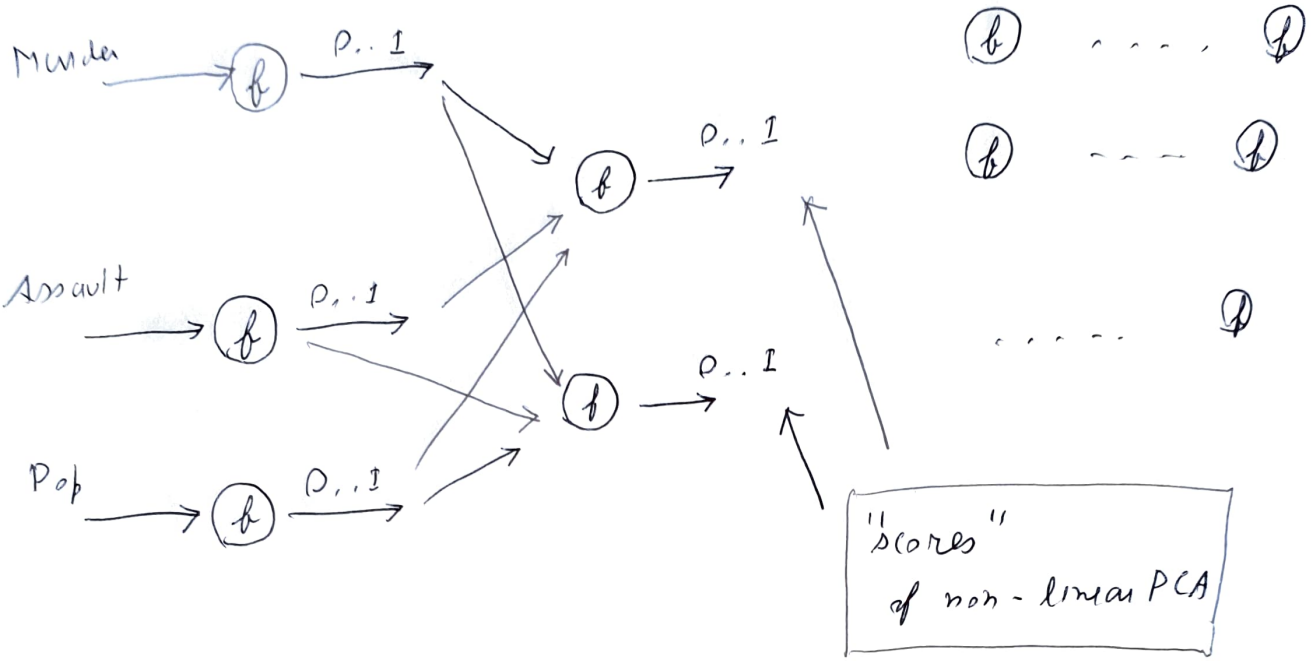
$\frac{1}{x}$ will not do

$$f = \frac{1}{1+e^{-x}}$$



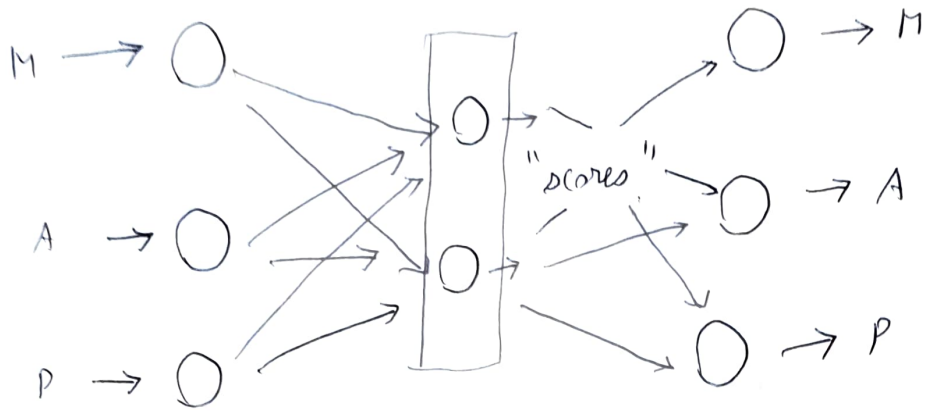
sigmoid





- repeat for California, Wisconsin, Vermont, etc.

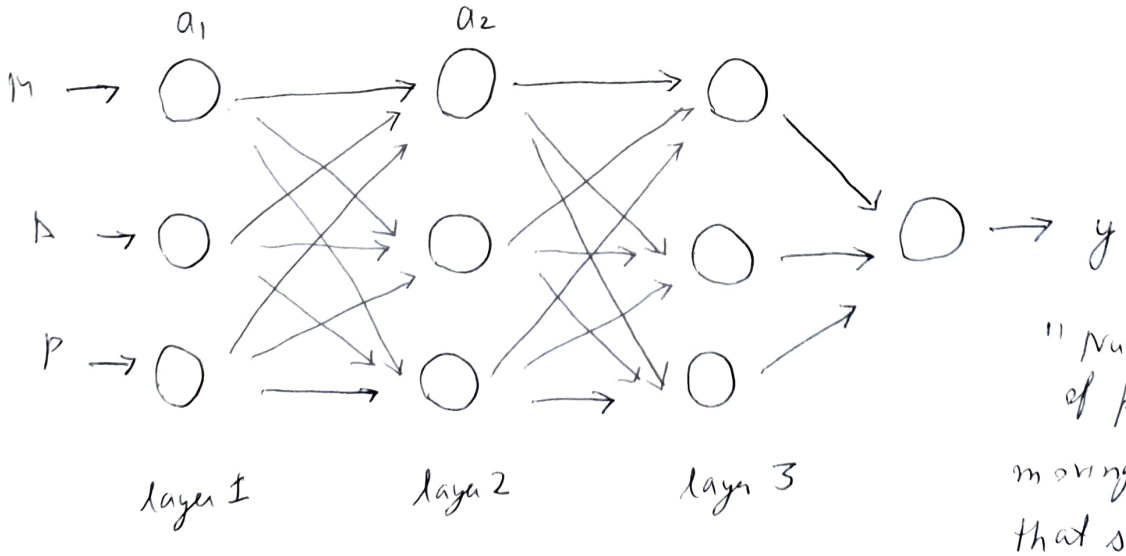
- modern deep neural networks have many layers
  "deep" learning
- 18B parameters in GPT-3.
- useful for complex, data, biomedical data
- can change distance from Euclidian to Manhattan, correlation, etc.
- "hacks" for understanding your data



"compress" information

auto encoder
Artificial Neural Network (ANN)

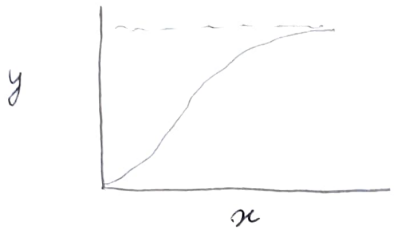
I can use ANN for predicting something
/ "supervised" learning
/ regression



- how do you estimate so many parameters? (a_1, a_2, ...) (animations show)
- change them slowly until you

logistic regression

$$y = \frac{1}{1 + e^{-(ax+b)}}$$



can predict odds, probabilities

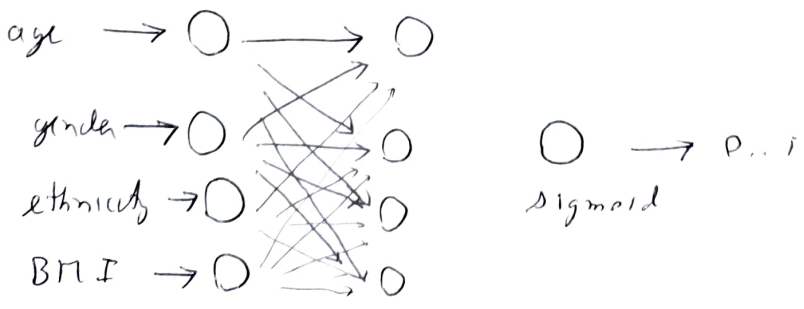
medical data

age	gender	ethnicity	BMI	Diabetes (0/1)
70	M	I	27	1
60	F	C	23	0
⋮				

supervised ML problem

- practicals in R / Python

can also use ANNs



- Black box models
- Difficult to interpret
- Maybe more accurate than simpler explainable models such as logistic regression and trees

- Ethics of black box models
  - may hide bias
  - hiring, recidivism, etc.
- case of lithium
  - not explainable but "works" in patients with bipolar

Trees

Cross-validation + Bias variance tradeoff

Sensitivity / Specificity

- CV curves

OUTLINE

NPJ ① ← – explainability

Gut ② ↗ ⓐ – trees

③ – LLMs (middle)

④ – RNNs

⑤ – how to read pictures CNNs.

– sensitivity / specificity

– ethics of black box

– Cross-validation

– bias variance

(DB)

explainability

NPJ

Gut

black box

ethics of black box

trees

Cross-validation

bias variance

sensitivity / specificity

→ LLMs middle RNN

→ pictures CNNs middle

practicals cut down

trees

LR

→ sample questions

– shopping

– image caption

– ARC

– mention training test split

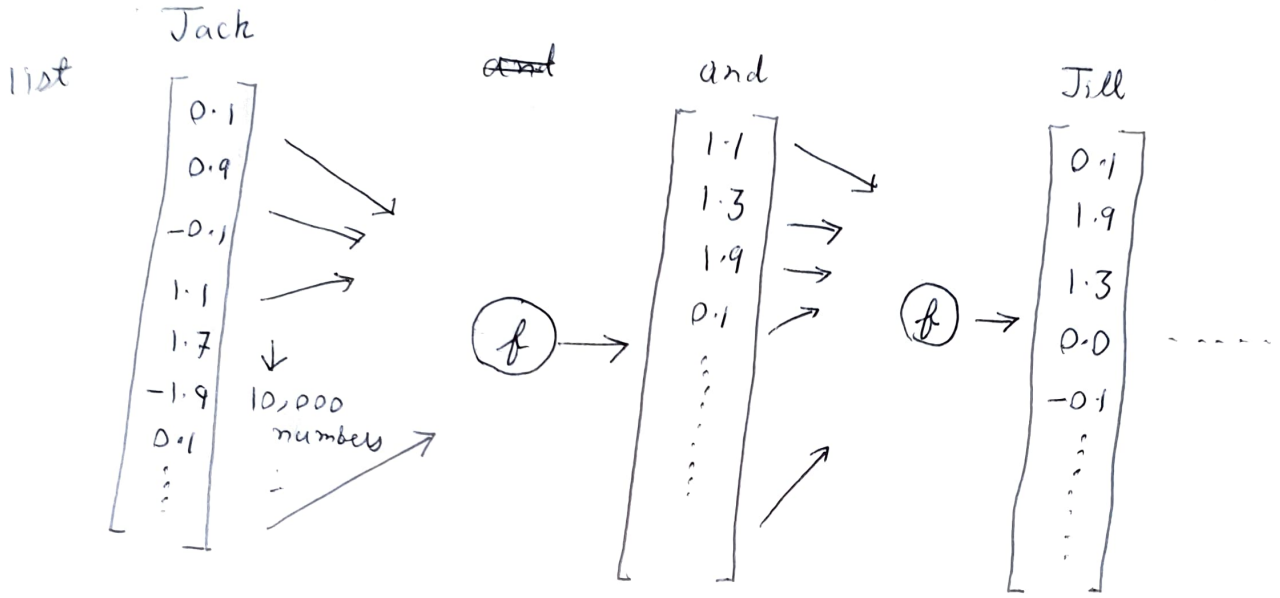
– CV

– curve

– CNN

LLMs and word embeddings

"Jack and Jill went up the ....."



- Riddle
- RNNs

- predict next word
- numbers encode content and meaning

CNNs

Riddle #2