Machine learning algorithms

Linear regression 1

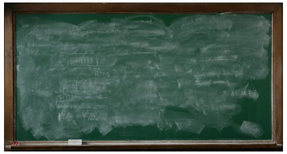
2020–09-25

CSCI 471 / 571, Fall 2020

Kameron Decker Harris

What is ML?

- Data + Optimization + Statistics → Predictions



Examples of ML applications

- Let's list some examples together
  - home assistant — learn from your behavior
    purchasing voice
  - targeted ads
  - self-driving cars — identify obstacles/objects
    — model cars around it
  - classify species iNaturalist
  - evolutionary embodied robots/simulated organism
  - optimize airflow w/ feedback
  - character recognition
  - denoising "touch-up"
  - conversation bots

Famous recent ML successes

Image classification

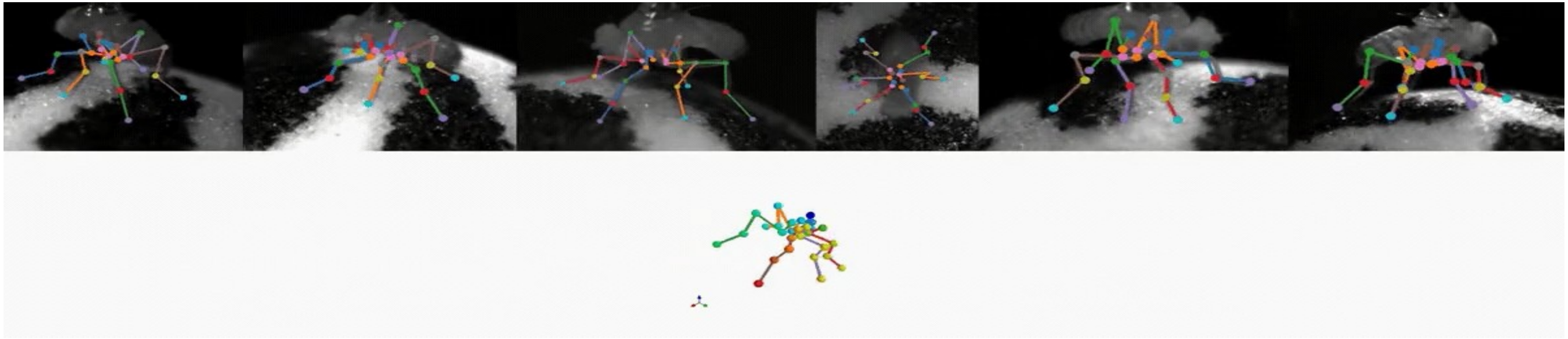
(CIFAR 100 data)

AlphaGo

Wikimedia commons Dilaudid

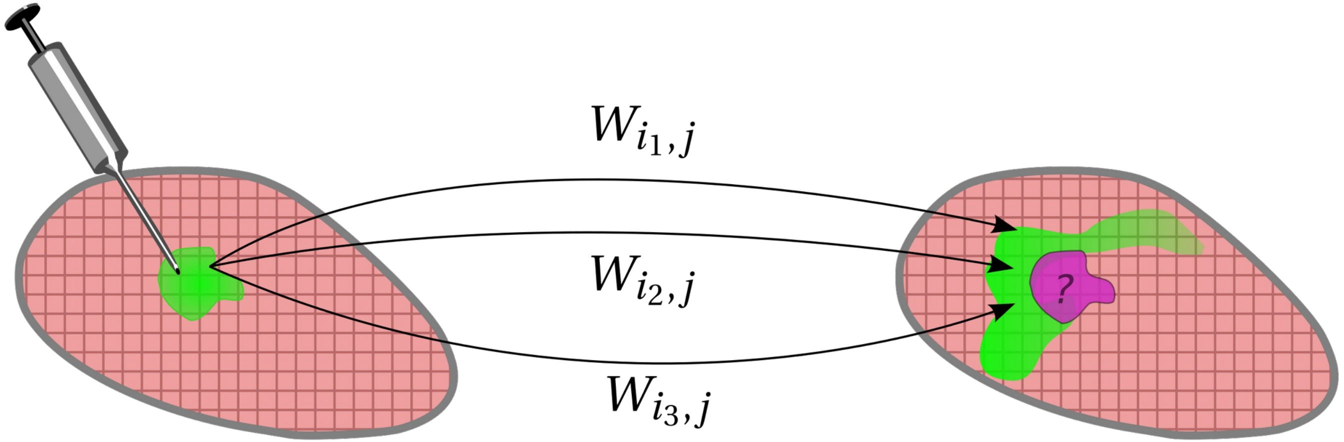
ML in data analysis

fly



Karaschuk et al., 2020

Ex: network reconstruction



x: source expression

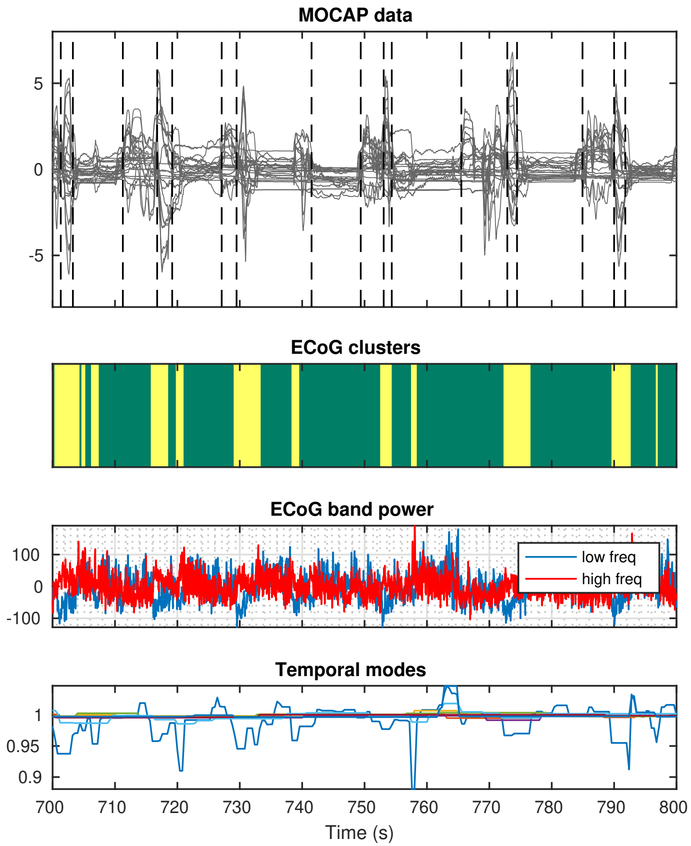
y: target expression

<u>Goal</u>

Find unknown weight matrix *W* so

$$\mathbf{y} \approx {\color{red}W}\mathbf{x}$$

ML for neuroscience



Harris et al., 2020

Hochberg et al., (2012)

Goals for the quarter

Goals for the quarter

- Understand important, existing algorithms
  - Theoretical grounding
  - Implementation in code

Goals for the quarter

- Understand important, existing algorithms
  - Theoretical grounding
  - Implementation in code
- General principles of ML
  - Tradeoffs, scalability, uncertainty
  - Building blocks of cutting-edge algorithms

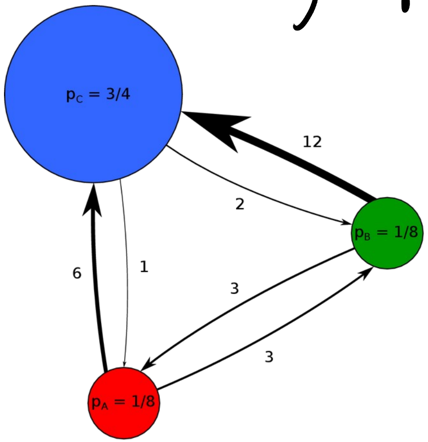
Data

census

STNAME	CTYNAME	CENSUS2000POP	ESTIMATESBASE2000	POPESTIMATE2000
Alabama	Alabama	4447100	4447382	4451849
Alabama	Autauga County	43671	43671	43872
Alabama	Baldwin County	140415	140415	141358
Alabama	Barbour County	29038	29038	29035
Alabama	Bibb County	20826	19889	19936
Alabama	Blount County	51024	51022	51181
Alabama	Bullock County	11714	11626	11604
Alabama	Butler County	21399	21399	21313
Alabama	Calhoun County	112249	112243	111342
Alabama	Chambers County	36583	36614	36593
Alabama	Cherokee County	23988	23986	24053

#'s

graph



Color
Red
Red
Yellow
Green
Yellow

Data

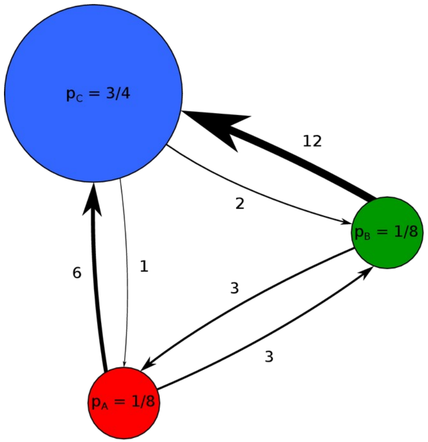
STNAME	CTYNAME	CENSUS2000POP	ESTIMATESBASE2000	POPESTIMATE2000
Alabama	Alabama	4447100	4447382	4451849
Alabama	Autauga County	43671	43671	43872
Alabama	Baldwin County	140415	140415	141358
Alabama	Barbour County	29038	29038	29035
Alabama	Bibb County	20826	19889	19936
Alabama	Blount County	51024	51022	51181
Alabama	Bullock County	11714	11626	11604
Alabama	Butler County	21399	21399	21313
Alabama	Calhoun County	112249	112243	111342
Alabama	Chambers County	36583	36614	36593
Alabama	Cherokee County	23988	23986	24053

one-hot

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

matrix



ML Taxonomy

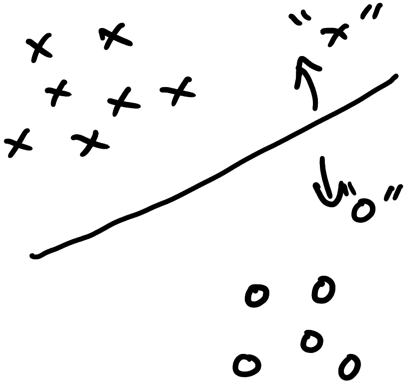
Supervised learning e.g. image classification (cat, dog, ...)

semi-

Data pts have labels (w/ noise)

Goal: given new data, w/o label, predict

- classification, categorical (true/false, colors)
- regression, just #'s (real)

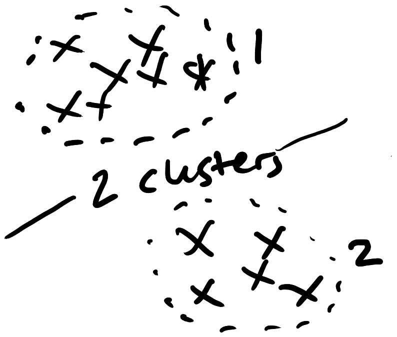


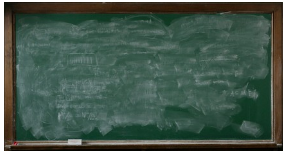
unsupervised learning

No labels

Goal: describe structure

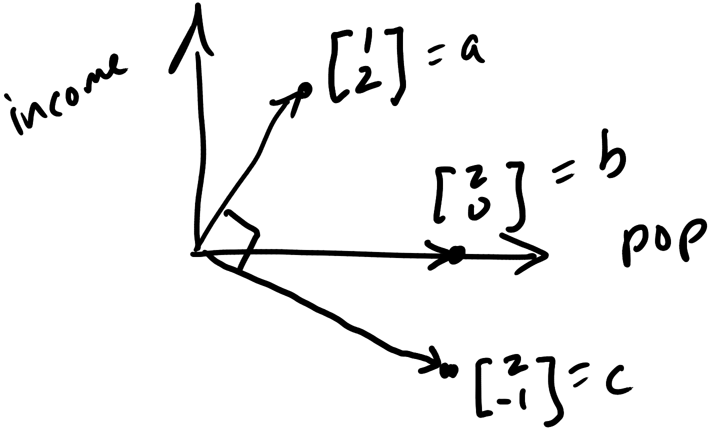
- clustering
- manifold learning
- probability distribution





Data as vectors

	Population	Income
Town 1	1	2
Town 2	2	0
Town 3	2	-1



$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$ column vector 2×1

x^T transpose

" $[x_1, x_2]$

1×2

norm

measures length

$$\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2} = \|x\|_2 \text{ "2-norm"}$$

inner product

"dot", "scalar"

$$x^T y = \sum_{i=1}^{d} x_i y_i$$

length d vectors

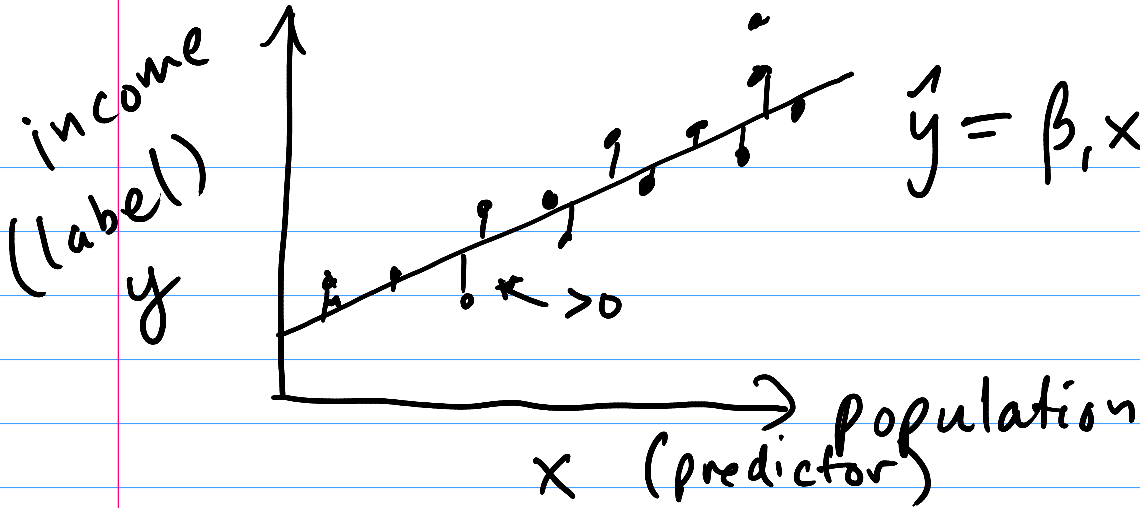
$$\|x\| = \sqrt{x^T x}$$

ex/ $a^T b = [1, 2]\begin{bmatrix} 2 \\ 0 \end{bmatrix}$

$= 1 \cdot 2 + 2 \cdot 0$

$= 2$

$a^T c = 0$ orthogonal



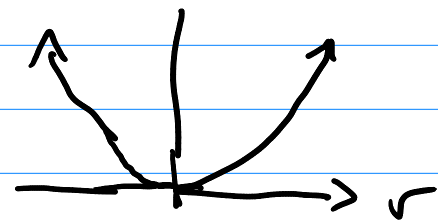
$\hat{y} = \beta_1 x + \beta_0 = f(x)$

linear fit

How to measure goodness of fit?

residual $\hat{y} - y = \beta_1 x_i + \beta_0 - y_i = \begin{cases} 0 & \text{perfect} \\ >0 & \text{overshoot} \\ <0 & \text{undershoot} \end{cases}$

Squared error $(\hat{y} - y)^2$



Pick β_0, β_1 so that

$$\sum_{i=1}^{n} \left((\beta_1 x_i + \beta_0) - y_i \right)^2 \rightarrow \min \qquad X\beta = \begin{bmatrix} \vec{x}_1^T \beta \\ \vec{x}_2^T \beta \\ \vdots \end{bmatrix}_{n \times 1}$$

in d-dimensions $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d$

$= x^T \beta$

vectors $\sum_{i=1}^{n} (\vec{x}_i^T \vec{\beta} - y_i)^2 = \|X\beta - y\|^2$

squared error $= \|X\beta - y\|^2$ "sum of squared residuals"

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_n & - \end{bmatrix}$$

$n \times d$ matrix of data

n = # data pts

d = dimension

$$X\beta = \begin{bmatrix} \vec{x}_1^T \beta \\ \vec{x}_2^T \beta \\ \vdots \end{bmatrix}$$

$n \times 1$ vector of predictions for each data pt.

$$X\beta - y = \begin{bmatrix} x_1^T \beta - y_1 \\ x_2^T \beta - y_2 \\ \vdots \end{bmatrix}$$

$n \times 1$ vector of residuals