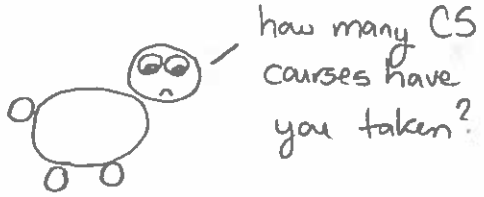
Machine Learning: A Whirlwind Guide

---

① I want to be able to predict people's ages, but unfortunately all I seem to be able to ask them is the following question:



Nevertheless, I'm pretty convinced that most people's ages are (roughly) some constant multiple of the number of CS courses they've taken, i.e.

$$y \approx ax$$

where x is the number of CS courses taken and y is the person's age

---

② But I don't know the constant... yet. However I do have a couple examples:

 a 20 year old student who's taken 5 CS courses

 a 41 year old professor who's taken 12 CS courses

③ That means I'd like:

$$20 \approx 5a$$
$$41 \approx 12a$$

Another way to express this desire is to say that I want to find a s.t.:

$$|20-5a| + |41-12a| \approx 0$$

or alternatively:

$$(20-5a)^2 + (41-12a)^2 \approx 0$$

---

④ If we let $L_1(a) = |20-5a| + |41-12a|$

$L_2(a) = (20-5a)^2 + (41-12a)^2$

then our objectives are expressible as: find a s.t.

$$L_1(a) \approx 0$$
or $$L_2(a) \approx 0$$

---

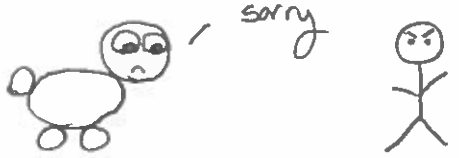
⑤ Neither $L_1(a)$ nor $L_2(a)$ can ever go negative, so another way to express our objectives is:

$$\underset{a}{\text{argmin}}\ L_1(a)$$

or $$\underset{a}{\text{argmin}}\ L_2(a)$$

⑥ Functions like $L_1(a)$ and $L_2(a)$ go by many names. Often they are called objective functions, because our objective is to find its minimal value. Another common name is loss function, because they capture the feeling of loss we experience if we use a particular value of a.

e.g. if we decide $a = 10$, then we predict the student's age is 50 and the professor's age is 120. This gives us a loss (according to L_1) of $|20-50| + |41-120| = 109$, which corresponds to 109 units of acute embarrassment



We want to find a value of a that minimizes our loss (embarrassment).

7. If we use $L_2(a)$ as our loss function, then we can compute $\operatorname{argmin}_a L_2(a)$ using standard calculus techniques:

$$L_2(a) = (20-5a)^2 + (41-12a)^2$$

$$\Rightarrow \frac{dL_2(a)}{da} = 2(20-5a)\cdot(-5) + 2(41-12a)\cdot(-12)$$

$$= -200 + 50a - 984 + 288a$$

$$= 338a - 1184$$

We can find the critical points of $L_2(a)$ by setting the derivative to zero and solving for a:

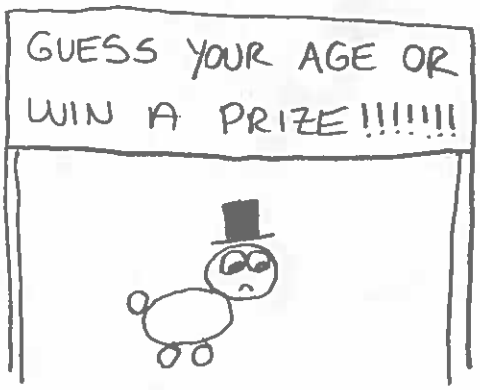
$$338a - 1184 = 0$$

$$\Rightarrow \quad a = \frac{1184}{338} \approx 3.5$$

⑧ Now we can take our prediction function

$$y = 3.5x$$

on the road! You open up a carnival booth.



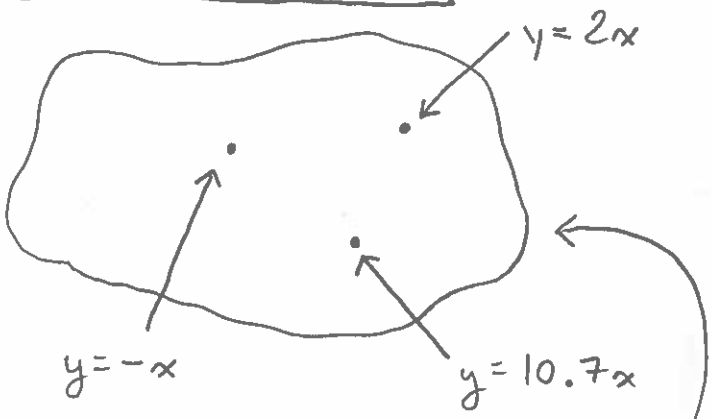
Unfortunately, most carnival attendees have never taken a CS course, and you guess their age is zero. You quickly go out of business.

⑨ This short tale highlights several high level concepts in machine learning:

training data

num courses (x)	age (y)
5	20
12	41

model (hypothesis) space



loss function

$$L_2(a) = (20 - 5a)^2 + (41 - 12a)^2$$

bias

sometimes the training data is not representative of the test data

test data

x	y
0	55
0	19
2	32
0	70

model capacity

sometimes the model space contains very few (or no) good models (like maybe number of CS courses is not enough to reliably predict age... maybe)