Machine learning

Linear regression

Joshua Loftus

It's the year 1801

There is no Google Maps

How do you travel? With the original GPS: **astronomy**

- Relatively mature science
- Millennia of observations and evolving theories
- About 2 centuries of observations *with telescopes*
- Similar time since Kepler's laws (elliptical orbit formulas)
- 20 years after discovery of Uranus
- But still, 45 years before discovery of Neptune

Discovering planets and things

Imagine being the first to observe a previously unknown celestial object in our solar system

…

and then losing it

A (pre-machine learning?) prediction contest

Yung Gauss

- Piazzi published his $n = 24$ observations in February
- An international community of scientists and mathematicians scrambled to find Ceres
- Almost a year later, it was rediscovered using the predictions of (24 year old) C. F. Gauß

Gauss became an instant celebrity

Why was this impressive?

- Ceres is small (smaller than our moon)
- Observed path only ~3 degrees of motion across the sky
- Almost a year passed, so its position was far from the initial observations
- Searching for a small dim object in a sky full of brighter stars

How did Gauss do it?

Data

Positions (two dimensional) and times for each of 24 observations

Transformations

Changing coordinates: Ceres-Earth, Earth-Sun, Ceres-Sun

Laws of motion

Gauss computed with > 80 variables in these coordinate systems to approximately solve Kepler's (non-linear) laws and determine the orbit of Ceres about the sun

Problem: Kepler's laws determine an orbit uniquely from 3 points. What to do with 24?

Gauss:

> When the number of unknown quantities is equal to the number of the observed quantities depending on them, the former may be so determined as exactly to satisfy the latter. But when the number of the former is less than that of the latter, an absolutely exact agreement cannot be determined, in so far as the observations do not enjoy absolute accuracy. In this *case care must be taken to establish the best possible agreement, or to diminish as far as practicable the differences*.

i.e. minimize errors

but why *squared* errors?

Detour: center of "mass"

Suppose (x_i, y_i) are observations we wish to model as

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

For some unknown "true" and/or "optimal" values (α, β)

For a given choice $(\hat{\alpha}, \hat{\beta})$ let $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ and $r_i = y_i - \hat{y}_i$.

Exercise: Show that $\bar{r} = 0$ if and only if the line $y = \hat{\alpha} + \hat{\beta} x$ passes through the point $(\bar{x}, \bar{y})$.

Problem: There are (uncountably) infinitely many solutions with "zero average error"

For a given $x \neq \bar{x}$, could predict *any* y with one of these lines

Detour: severe testing

If a method or theory can be used to produce *any* prediction, we are on very shaky ground scientifically

> If the new predictions are borne out, then the new theory is corroborated (and the old one falsified), and is adopted as a working hypothesis. If the predictions are not borne out, then they falsify the theory from which they are derived.

Popper, *Logic of Scientific Discovery*

Mayo, on severity:

> A conjecture passes a **severe test** only if a refutation would *probably* have occurred if it's false

Constrained methods, specific theories

- If our method only requires drawing a line passing through $(\bar{x}, \bar{y})$ then it can be made consistent with any new observation
- A theory based on this method will not be falsifiable, and can not be severely tested
- If the method is mathematically well-defined, producing a unique solution, then theories formed using that method can be severely tested

Philosophical ideas can guide our thinking about machine learning

Question: Are more flexible models easier/more difficult to falsify or severely test? Is this good/bad scientifically?

Constraints

1. Errors sum to zero
2. Minimize something else, but what?

- Around the same time, R. J. Boscovich and P-S. Laplace minimized sum of absolute errors

$$\text{minimize} \sum_i |r_i|$$

- Laplace also suggested minimizing the maximum error

$$\text{minimize} \max_i |r_i|$$

- Gauss said we can use any even power, e.g. $\sum r_i^8$

Gauss's answer

> of all these principles ours [least squares] is the **most simple**; by the others we shall be led into the most complicated calculations

If *Gauss* didn't want to do those calculations, that's really saying something...

On the other hand, he said he used least squares *thousands* of times in his years of work (without electricity!)

For more about the origin of least squares see this article.

Another answer: nice geometry

At a minimum of

$$\ell(\hat{\alpha}, \hat{\beta}) = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

we have

$$0 = \frac{\partial \ell}{\partial \alpha} = -2 \sum_i r_i$$

i.e. the first constraint is satisfied, and

$$0 = \frac{\partial \ell}{\partial \beta} = -2 \sum_i x_i r_i$$

i.e. **orthogonality**.

Orthogonality, uncorrelatedness, bias

- Since $\bar{r} = 0$ and $\sum x_i r_i = 0$, we also have

$$\text{cor}(x, r) = 0$$

- Correlation measures *linear dependence*
- If we minimized a different loss function and the resulting residuals were correlated with x, this would mean there is some remaining (linear) signal
- A (linear) pattern in residuals, i.e. bias

Lessons for ML from the re-discovery of Ceres

- **Severity (or novelty)**: lots of mathematicians used methods to fit the initial observations, what distinguished Gauss was predicting a *new* data point
- The *right amount* of **complexity**: some predictions assumed a circular orbit instead of elliptical, this simplified calculations but missed Ceres
- **Theory *and* observation**: without the heliocentric model of the solar system this search would have been a lost cause. That model itself evolved from previous iterations of theories and observations

An absurdly abbreviated history of optics

- Ptolemy (100-170) measured refraction of light passing from air to water, altering measurements so they would fit to a quadratic curve (ancient "machine learning" or curve fitting)
- Ibn Sahl (940-1000) described the correct law, which English speakers refer to as Snell's law
- Fermat (1607-1665) proposed the more general "principle of least time" which can be used to derive Snell's law *and* solve other optical problems -- predict values in different settings

Put the "science" back in "data science"!

Feynman, Lectures on Physics, Vol. 1:

> Now in the further development of science, we want more than just a formula. First we have an observation, then we have numbers that we measure, then we have a law which summarizes all the numbers. *But the real glory of science is that we can find a way of thinking such that the law is evident*.

- Ceres: elliptical orbits (later understood using gravitation)
- Optics: least time
- A typical modern application of machinee learning: ...

By contrast, regression done badly

Convenience of calculation enables a lot of bad science

- A. Quetelet in 1835, "social physics," correlates basically any social data together, tries to predict "crime," poverty, alcohol consumption, etc

- F. Nightingale (1820-1910) believes correlations observable this way demonstrate "God's will"

- F. Galton (1822-1911) founds the field of eugenics...

And much of modern ML is similarly fitting curves to model relationships in any available data *because we can* -- not because there is any scientific or theoretical reason to do so

Regression began with an exemplary application, the re-discovery of Ceres

Scientifically questionable applications have exploded since then

Computers speed up the process, which perhaps decreases quality

The era of "surveillance capitalism" means scientifically (and ethically) questionable data is multiplying faster than ever