Machine Learning Basics

Neal Parikh

Learn 2 Quant Conference, New York
November 16, 2018

Finding spy planes

US Federal Agents Flew A Secret Spy Plane To Hunt Drug Cartel Leaders In Mexico

Neither the US Marshals Service nor the Mexican government wants to talk about their joint efforts to hunt drug kingpins. But BuzzFeed News spotted a Marshals spy plane circling around the time of a prominent capture in Sinaloa.

Posted on August 3, 2017, at 8:00 a.m.

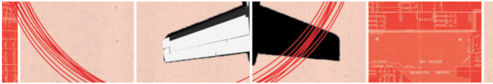
Peter Aldhous
BuzzFeed News Reporter

Karla Zabludovsky
BuzzFeed News Reporter

in August 2017, Buzzfeed News publishes articles finding

- military contractors flying over SF Bay Area
- secret US Marshals plane hunting drug cartel kingpins in Mexico
- Air Force special operations planes flying over US
- ...

Finding spy planes

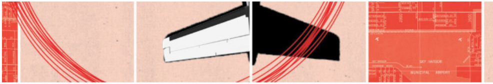


BuzzFeed News Trained A Computer To Search For Hidden Spy Planes. This Is What We Found.

From planes tracking drug traffickers to those testing new spying technology, US airspace is buzzing with surveillance aircraft operated for law enforcement and the military.

1. pull a publicly available dataset (not intended for this purpose)
2. train a simple machine learning model
3. validate (here, ‘do journalism’)

Finding spy planes



BuzzFeed News Trained A Computer To Search For Hidden Spy Planes. This Is What We Found.

From planes tracking drug traffickers to those testing new spying technology, US airspace is buzzing with surveillance aircraft operated for law enforcement and the military.

1. pull 4 months of flight-tracking data from website Flightradar24
2. extract 'features': turning rates, speeds, altitudes, manufacturers
3. train a binary classifier to distinguish between previously identified FBI/DHS planes and not
4. validate

Examples

- Adobe (font recognition using phone camera)
- Amazon (speculative shipping, Kindle browser prefetching)
- American Express (fraud detection, individual credit limits)
- Cheesecake Factory (predict food ingredient demand)
- C-SPAN (automatically name politicians on screen)
- HireVue (video analysis of job interviews for hiring/screening)
- Nest Thermostat (embedded control of smart thermostat)
- Target (market research, individualized product catalogues)
- USPS (handwriting recognition)
- Walmart (inventory, product placement)

Automated sepsis detection for hospital operations

- sepsis is #3 leading cause of death in US, but hospitals often miss early signs and don't catch it until it's too late
- university hospitals (Duke, Johns Hopkins) deploying ML systems, some this month (Sepsis Watch), for automated sepsis detection
- *e.g.*, Duke system trained on 50K patient records, over 32M data points, with many variables (vital signs, lab tests, medical history)
- pulls patient data every 5 min to evaluate conditions, then alerts nurses
- nurses make decisions about alert, and if approved, are guided through checklist of actions

What is machine learning?

- no precise technical definition
- usage evolved over time
- ‘classical’ usage is as a sub-discipline of AI research

What is machine learning?

- intersection of computer science and statistics
- computationally tractable algorithms that learn from data
- the mathematical foundation of modern AI, but now also used in a huge variety of other domains

What is machine learning?

- modern usage: how to build *learning procedures*, *i.e.*, how to use historical data to build a *prediction rule*
- prediction rule: algorithm mapping observable inputs to prediction of unknown quantity (the *response*)
- focus is on making predictions, and doing well on data you *haven't yet seen* (how to select the right prediction rule among several)

What is machine learning?

- modern usage: how to build *learning procedures*, *i.e.*, how to use historical data to build a *prediction rule*
- prediction rule: algorithm mapping observable inputs to prediction of unknown quantity (the *response*)
- focus is on making predictions, and doing well on data you *haven't yet seen* (how to select the right prediction rule among several)
- informally, is mostly interchangeable with the terms 'AI' and 'modern statistical prediction' (*e.g.*, Sepsis Watch can be called 'an AI')

Machine learning and AI

1950s	Dartmouth conferences; chess & checkers; LISP; perceptron
1960s	early foundational & philosophical work; formal logic
1970s	neural networks; AI winter
1980s	expert systems; AI winter

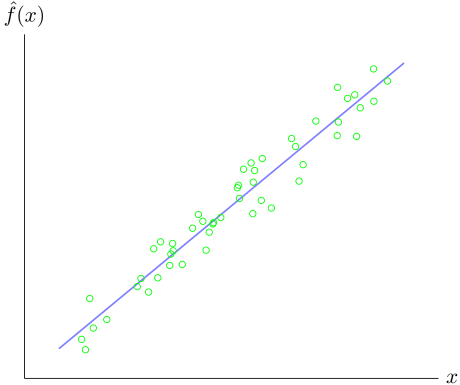
1990s	probabilistic revolution; graphical models; kernel methods
2000s	convex optimization; continuing development from 90s
2010s	deep learning; large-scale & widespread applications

Machine learning and statistics

(Wasserman; Tibshirani)

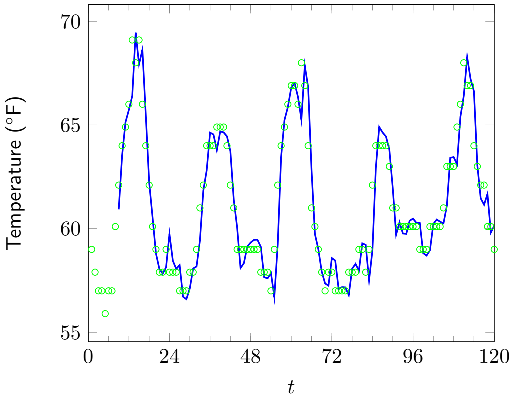
statistics	computer science
estimation/fitting	learning
regression/classification	supervised learning
clustering/density estimation	unsupervised learning
data	training sample
covariates	features, inputs
response	outputs
test set performance	generalization ability

Linear regression



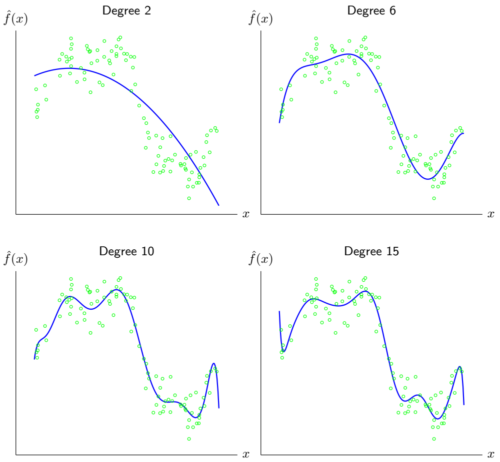
Straight line fit to 50 points in a plane.

Autoregressive time series



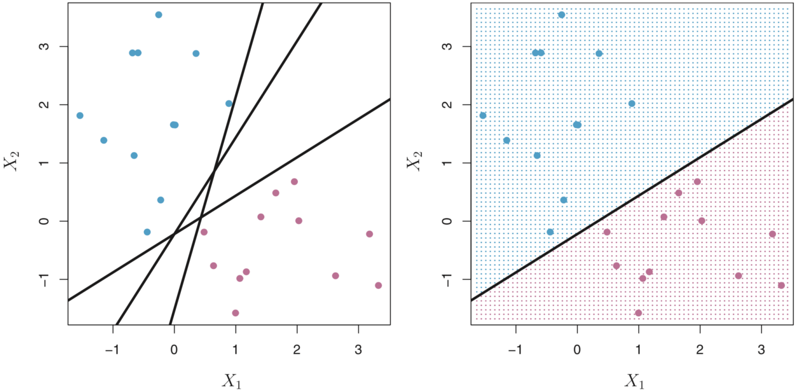
Hourly temperature at LAX.

Polynomial regression

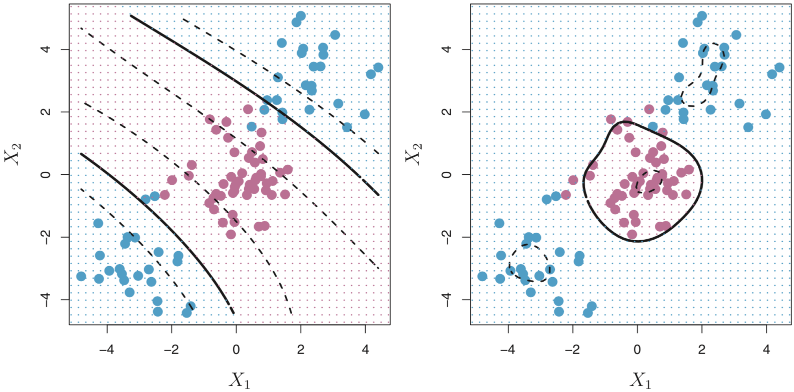


Least squares fits of degree 2, 6, 10, and 15 to 100 points.

Support vector machine

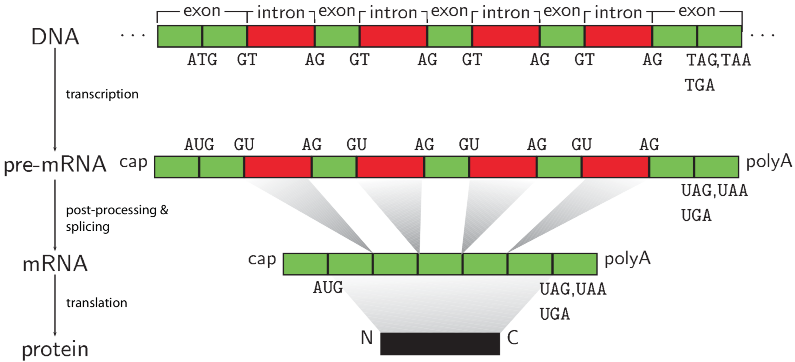


Support vector machine



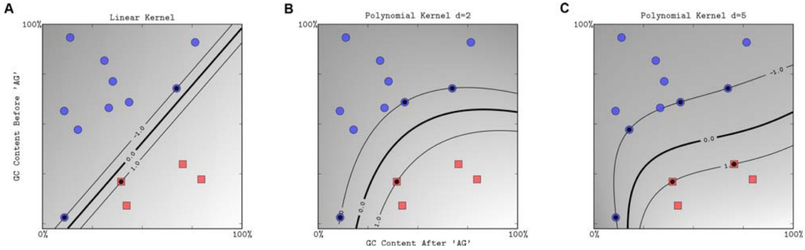
Splice site recognition

(Ben-Hur et al., *PLoS Computational Biology*, 2008)



Splice site recognition

(Ben-Hur et al., *PLoS Computational Biology*, 2008)



Uses and pitfalls

uses:

- explore new, richer, unused datasets (text, image, . . .)
- internal operations (anomaly detection, data processing, . . .)
- actual trading signals, portfolio construction, . . .

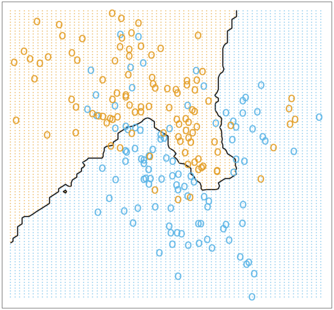
pitfalls:

- need appropriate team and workflows (*e.g.*, model diagnostics)
- ‘bias’ and ethics
- (wrongly) anthropomorphizing models

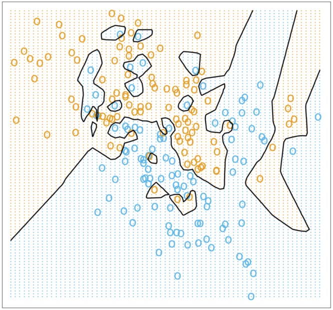
Thanks

Questions?

k-nearest neighbors



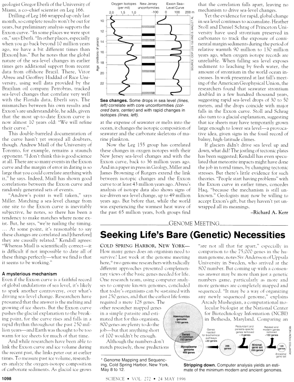
$k = 15$



$k = 1$

Source: Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*

Topic models



geologist Gregor Eberli of the University of Miami, a co–chief scientist on Leg 166.

Drilling of Leg 166 wrapped up only last month, so complete results won't be out for years, but preliminary analysis supports the Exxon curve. "In some places we were spot on," says Eberli. "In other places, especially when you go back beyond 10 million years ago, we have a bit different times than [Exxon] has." But he notes that the global nature of the sea-level changes in earlier times gets additional support from recent data from offshore Brazil. There, Vitor Abreu and Geoffrey Haddad of Rice University, using well data provided by the Brazilian oil company Petrobras, tracked sea-level changes that correlate very well with the Florida data, Eberli says. The mismatches between his own results and Exxon's are understandable, he adds, given that the most up-to-date Exxon curve is now almost 10 years old: "We will refine their curve."

This double-barreled documentation of the curve hasn't yet swayed all doubters, though. Andrew Miall of the University of Toronto, for example, remains a staunch opponent. "I don't think this is good science at all. There are so many events in the Exxon curve and the margin of error in dating is so large that you could correlate anything with it," he says. Indeed, Miall has shown good correlations between the Exxon curve and randomly generated sets of events.

"Andrew's point is well taken," says Miller. Matching a sea-level change from one site to the Exxon curve is inevitably subjective, he notes, so there has been a tendency to make matches where none exist. But, he says, "we're nailing the timing. … At some point, it's reasonable to say these changes are correlated and [therefore] they are causally related." Kendall agrees: "Whereas Miall is scientifically correct—it is difficult if not impossible to date all of these things perfectly—what we find is that it seems to be working."

A mysterious mechanism

Even if the Exxon curve is a faithful record of global undulations of sea level, it's likely to spark another controversy, over what's driving sea-level change. Researchers have presumed that the answer is the melting and growing of ice sheets. But the Exxon curve pushes the glacial explanation to the breaking point, for the curve rises and falls in a rapid rhythm throughout the past 250 million years—and Earth was thought to be too warm for ice sheets for much of that time.

And while researchers have been able to link the Exxon curve and ice volume during the recent past, the links peter out at earlier times. To measure past ice volume, researchers analyze the oxygen-isotope composition of carbonate sediments. As glacial ice grows

Sea changes. Some drops in sea level *(lines, left)* correlate with core unconformities *(colored bars, center)* and with rapid changes in isotopes *(lines, left)*.

at the expense of seawater or melts into the ocean, it changes the isotopic composition of seawater and the carbonate skeletons of marine plankton.

Now the Leg 155 group has correlated these changes in oxygen isotopes with their New Jersey sea-level changes and with the Exxon curve, back to 36 million years ago. And in a paper in press in *Geology*, Miller and James Browning of Rutgers extend the link between isotopic changes and the Exxon curve to at least 43 million years ago. Abreu's analysis of isotope data also shows signs of ice-driven sea-level change, up to 49 million years ago. But before that, while the world was experiencing the warmest heat wave of the past 65 million years, both groups find that the correlation falls apart, leaving no mechanism to drive sea-level changes.

Yet the evidence for rapid, global change in sea level continues to accumulate. Heather Stoll and Daniel Schrag of Princeton University have used strontium preserved in carbonates to track the exposure of continental margin sediments during the period of relative warmth 90 million to 130 million years ago, when oxygen isotope records are unreliable. When falling sea level exposes sediment to leaching by fresh water, the amount of strontium in the world ocean increases. In work presented at last fall's meeting of the American Geophysical Union, the researchers found that seawater strontium doubled in a few hundred thousand years, suggesting rapid sea-level drops of 30 to 50 meters, and the drops coincide with major falls in the Exxon curve. Stoll and Schrag also turn to a glacial explanation, suggesting that ice sheets may have temporarily grown large enough to lower sea level—a provocative idea, given signs in the fossil record of balmy, high-latitude climes.

If glaciers didn't drive sea level up and down, what did? The jostling of tectonic plates has been suggested; Kendall has even speculated that meteorite impacts might have done the job in torrid times, by changing tectonic stresses. But there's little evidence for such theories. "People start having problems" with the Exxon curve in earlier times, concedes Haq, "because the mechanism is still unknown." Geologists may now be willing to accept Exxon's gift, but they haven't yet unwrapped all its meanings.

–Richard A. Kerr

GENOME MEETING

Seeking Life's Bare (Genetic) Necessities

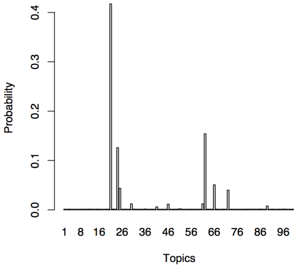
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

1098 SCIENCE • VOL. 272 • 24 MAY 1996



Topic models

sound	quantum	brain	computer	ice
speech	laser	memory	data	climate
acoustic	light	human	information	ocean
language	optical	visual	problem	sea
sounds	electron	cognitive	computers	temperature
stars	research	materials	fossil	volcanic
universe	national	organic	species	years
galaxies	science	molecules	evolution	fig
astronomers	new	molecular	birds	deposits
star	funding	polymer	evolutionary	rocks

Topic models