Statistical foundations of machine learning

INFO-F-422

Gianluca Bontempi

Machine Learning Group
Computer Science Department
mlg.ulb.ac.be

ML in the news

GADGETS / GAMING

NEWS

Mind Reading to Predict the Success of Online Games

Engineers devise a way to predict an online game's success by gamers' initial emotional response

By YU-TZU CHIU / TUE, FEBRUARY 05, 2013

ML in the news

CNNMoney

BEST JOBS 2012

Best new jobs in America

IT Data Scientist

10-year job growth: 18.7%
Median pay: $98,600

What they do all day? Tech firms like LinkedIn, Facebook and Twitter are at the heart of the big data movement. Their users are generating loads of information by the second. Turning those heaps of data into business value falls to data scientists, who apply various tools and methods to find meaningful patterns and insights in large data sets.

How to get the job? An affinity for numbers is key, as well as a command of computing, statistics, math and analytics. One can't underestimate the importance of soft skills either. Data scientists work closely with management and need to express themselves clearly.

What makes it great? This is a cutting-edge field. The information explosion is spurring types of analysis that have never been performed before. The skill set is unique, and employers are willing to pay for qualified candidates. Six-figure paydays aren't uncommon.

What's the catch? It's an intense job. After 20 years of crunching info, it's easy to become susceptible to burnout, says Vincent Granville, a data scientist who left a corporate job to launch Analyticbridge, a social network for data science professionals.

NEXT: Wind Turbine Mechanical Engineer

Source: PayScale.com, CNNMoney research
Note: Median pay is for an experienced worker (at least five to seven years in field). Job growth is estimated for 2010-20, and based on people working in broader 'job family' from the Bureau of

ML in the news

GIGAOM

TAGS: analytics / big data / sports

Can machine learning make sense of the NFL's big data?

by Derrick Harris

NOV. 25, 2012 - 3:25 PM PST

SUMMARY: *According to the New York Times, National Football League teams are often years behind their peers in other professional sports when it comes to data analysis. Machine learning could help teams make more sense of the myriad variables that currently keep them relying on human intuition.*

When it comes to using data to determine how to build a team or manage a game, the National Football League appears years behind its professional sports brethren such as Major League Baseball and the National Basketball Association. But perhaps the increasing popularity of machine learning can change that by helping NFL teams make more sense of their very complex datasets.

Delving deep into the world of computer science might sound like overkill, but professional football is big

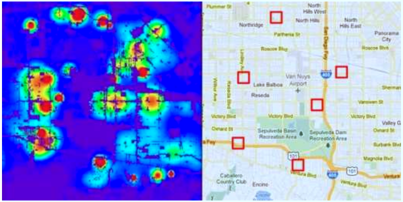
ML in the news

MIT Technology Review

L.A. Cops Embrace Crime-Predicting Algorithm

Burglary reports dropped after officers began taking patrol orders from computers.

By David Talbot on July 2, 2012

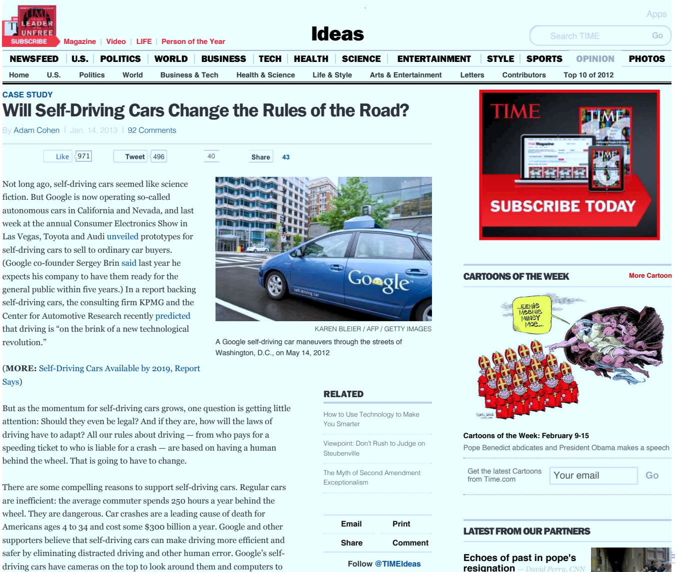


On patrol: A computer-generated "heat map," left, shows predicted crime activity. This is translated into patrol instructions in the form of the red boxes on the map, right.

A recent study suggests that computers could be better than seasoned police analysts at predicting when and where crime will strike next in a busy city.

Software tested in Los Angeles was twice as good as human analysts at predicting where burglaries and

ML in the news



Magazine | Video | LIFE | Person of the Year

Ideas

Apps

Search TIME Go

NEWSFEED | U.S. | POLITICS | WORLD | BUSINESS | TECH | HEALTH | SCIENCE | ENTERTAINMENT | STYLE | SPORTS | OPINION | PHOTOS

Home | U.S. | Politics | World | Business & Tech | Health & Science | Life & Style | Arts & Entertainment | Letters | Contributors | Top 10 of 2012

CASE STUDY

Will Self-Driving Cars Change the Rules of the Road?

By Adam Cohen | Jan. 14, 2013 | 92 Comments

Like 971 Tweet 496 40 Share 43

Not long ago, self-driving cars seemed like science fiction. But Google is now operating so-called autonomous cars in California and Nevada, and last week at the annual Consumer Electronics Show in Las Vegas, Toyota and Audi unveiled prototypes for self-driving cars to sell to ordinary car buyers. (Google co-founder Sergey Brin said last year he expects his company to have them ready for the general public within five years.) In a report backing self-driving cars, the consulting firm KPMG and the Center for Automotive Research recently predicted that driving is "on the brink of a new technological revolution."

KAREN BLEIER / AFP / GETTY IMAGES

A Google self-driving car maneuvers through the streets of Washington, D.C., on May 14, 2012

(MORE: Self-Driving Cars Available by 2019, Report Says)

But as the momentum for self-driving cars grows, one question is getting little attention: Should they even be legal? And if they are, how will the laws of driving have to adapt? All our rules about driving — from who pays for a speeding ticket to who is liable for a crash — are based on having a human behind the wheel. That is going to have to change.

There are some compelling reasons to support self-driving cars. Regular cars are inefficient: the average commuter spends 250 hours a year behind the wheel. They are dangerous. Car crashes are a leading cause of death for Americans ages 4 to 34 and cost some $300 billion a year. Google and other supporters believe that self-driving cars can make driving more efficient and safer by eliminating distracted driving and other human error. Google's self-driving cars have cameras on the top to look around them and computers to

RELATED

How to Use Technology to Make You Smarter

Viewpoint: Don't Rush to Judge on Steubenville

The Myth of Second Amendment Exceptionalism

Email | Print

Share | Comment

Follow @TIMEIdeas

CARTOONS OF THE WEEK | More Cartoon

Cartoons of the Week: February 9-15

Pope Benedict abdicates and President Obama makes a speech

Get the latest Cartoons from Time.com | Your email | Go

LATEST FROM OUR PARTNERS

Echoes of past in pope's resignation — David Perry, CNN

ML in the news

IMPETUS
Innovation Architected.
Blogs

Search

Impetus Corporate

Big Data

Home » Big Data » Machine Learning

Sentiment analysis using statistical machine learning

Posted by Sridhar S | Jul 04, 2012 | (0) | Add a Comment | SHARE

Sentiment analysis is a task of identifying whether the opinion expressed in a **text source is positive or negative** about a given topic, we can **classify it further more to other nine identified human sentiments** like love, laughter, compassion, anger, valor, fear, disgust, wonder and peace. Sentiment analysis in context of product reviews involves automatically classifying whether users have expressed positive or negative opinions regarding a product. However, it is usually early-on in a product life cycle that a company wants to quickly assess popular sentiment towards a product.

Under such circumstances, the only option available is to manually label a large number of product reviews to generate training data; a costly endeavor. We have explored some solutions to this problem. In particular we analyze the extent to which classification models trained on one set of products can be used for analysis of reviews of a different product. In particular we explore appropriate strategies for combining multiple classification models trained on different products.

So to **answer the problem of predicting the sentiment polarity of a new sentiment** in form of a product review using training data available for a different set of products, we are **focusing on a statistical machine learning technique of Support Vector Machine classifiers**. This approach of SVM classifiers amalgamation with vocabulary inter-sectional heuristic will help build a outperforming SVM based sentiment analysis tool.

We **define the problem of product sentiment analysis as follows**.

Given a significant number of labeled reviews for products P1 to Pn we want to learn a classification model for a new product Pn+1 for which only a small number of labeled reviews are provided. Let Di represent a set of labeled product reviews for product i. Then each element of Di is a two tuple (rij,lij), where rj is the jth review for product Pi and lj is its sentiment label (either positive:1 or negative:0).

Our goal is to use the available data in sets D1…Dn to achieve a high prediction accuracy on the test set of reviews Dn+1 for product Pn+1.

Blogs Category

- Management Thoughts
- Big Data
  - NoSQL
  - Hadoop Ecosystem
  - Erlang Universe
  - Technologies
  - Machine Learning
- Mobility
- Social Media
- Test/Performance Eng.
- Agile/Scrum
- Product Engineering
- Other Technologies

Recent comments

Three Generations of Tools for Realizing Machine Learning Algorithms
Feb 05, 2013
Dr. Vijay Srinivas Agneeswaran

Using Diagnostic tools to identify Software Performance Bottlenecks
Dec 03, 2012
Mustafa B

ML in the news

eWEEK® Enterprise IT Technology News, Opinion and Reviews

IBM Genomic Analytics Platform Offers Personalized Medicine

By **Brian T. Horowitz** | Posted 2012-03-15

IBM is moving on from "Jeopardy" to finding treatments using genetic data. The company announced it has developed a data-analytics platform called Clinical Genomics that uses algorithms and analytics similar to that of Big Blue's Watson supercomputer to find treatments for conditions based on a patient's genetic profile.

Doctors can use Clinical Analytics to analyze patients' similarities, predict outcomes, evaluate risk benefits and view treatment options. In addition to genetic data, the software also takes into account a patient's age, gender, symptoms and past diagnoses.

The new platform combines knowledge of patient histories with automatic statistical analysis and machine learning, said Dr. Haim Nelken, manager of IBM Research's health care and privacy solutions division in Haifa, Israel.

"This is a standard-based generic framework that can be used in the context of various diseases and for a range of decision-support tasks, such as patients' similarity assessment, prospective outcome prediction, risk-benefits analysis, and presentation of the treatment options most likely to succeed," Nelken wrote in an email to *eWEEK.*

Researchers can access data through a standard Web browser on tablets or PCs.

IBM's Clinical Genomics fits into the growing trend in health care of using big data to develop personalized medicine, which is the ability to use a patient's personal genetic characteristics to prescribe medical treatment for conditions, such as cancer, hypertension and AIDS.

Clinical Genomics enables doctors to aggregate medical statistics and develop recommendations and weighted predictions, Nelken wrote. The software analyzes medical guidelines and the knowledge clinicians provide, and correlates it with patient

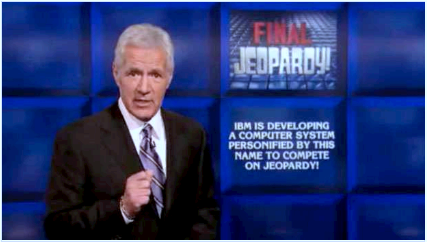
ML in the news

Back to All Icons

A Computer Called Watson

Overview
Transforming the World
Technical Breakthroughs
The Team
In Their Words

To successfully compete on *Jeopardy!* Watson had to analyze each question to determine exactly what was being asked, analyze the available content to extract precise answers, and quickly compute its level of confidence in the answer based on supporting and refuting information found.



IBM "Watson" System Challenged Humans at Jeopardy!

In its process of analyzing a question and determining the best answer,

What has all this in common?

- Large and increasing availability of big data (samples, variables) made possible by ubiquitous measurement systems (web, communications, gps, bank, sensors).
- Desire of extracting value from data
- Growing computing power.
- Awareness of the existence of which information patterns hidden somewhere within complex masses of data.
- New adaptive, automated, intelligent, smart, applications.
- Prediction.
- Solving hard problems in a new manner.
- Decision support.

About the course

- Why is it a course for computer scientists?
  - Information.
  - Automatic improvement of computer capabilities.
  - Models.
  - Algorithms.
  - Simulations, programs.
- Requirements: Preliminary course on statistics and probability.
- Exam: oral questions and project.
- TP:
  - introduction to the R language.
  - Hands-on
  - Real case studies
- Web page: `http://www.ulb.ac.be/di/map/gbonte/InfoF422.html`
- Syllabus in english (on the web page)

Outline of the course

- Foundations of statistical inference.
- Estimation
- Hypothesis testing
- Nonparametric methods
- Statistical machine learning
- Linear models
    - Regression
    - Classification
- Nonlinear models
    - Regression
    - Classification

Machine Learning: a definition

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Machine learning and statistics

Reductionist attitude: *ML is a modern buzzword which equates to statistics plus marketing*

Positive attitude: ML paved the way to the treatment of real problems related to data analysis, sometimes overlooked by statisticians (nonlinearity, classification, pattern recognition, missing variables, adaptivity, optimization, massive datasets, data management, causality, representation of knowledge, non stationarity, high dimensionality, parallelisation)

Interdisciplinary attitude: ML *should* have its roots on statistics and complements it by focusing on: algorithmic issues, computational efficiency, data engineering.

Who is a data scientist ?

Someone who knows statistics better than a computer scientists and programs better than a statistician...

What is statistics?

- Early definition: ".. describes tabulated numerical facts relating to the state".
- ".. refers to the methodology for the collection, presentation and analysis of data, and for the uses of such data"
- "use of data to make intelligent, rigorous, statements about a much larger phenomenon from which the data were selected"
- "...aid the interpretation of data that are subject to appreciable haphazard variability"
- "... builds on the use of probability to describe variation".
- " ... treats data that were obtained by some repetitive operation... probability as an idealization of the proportion of times that a certain result will occur in repeated trials of an experiment"

What is statistics?

- "The object of statistics is information. The objective of statistics is the understanding of information contained in data"
- "...is the study of how degrees of belief are altered by data".
- "Statistics concerns the *optimal* methods of treating, analyzing data generated from some chance mechanism".
- "Statistics is fundamentally concerned with the understanding of structure in data".
- "... presents management with quantitative facts which may assist in making decisions".
- "...permits the decision maker to evaluate the magnitude of the risk in the light of possible gains to be achieved".

A little history

- Statistics, as an organized and formal scientific discipline has a relative short history in comparison with the traditional deterministic sciences of chemistry, physics or mathematics.
- Colbert(1619-1683), Vauban (1633-1707): census in Canada and France.
- Petty (1623-1687): Political Arithmetic school (data modeling, forecasting laws in the economic and demographic context).
- Books of Cardano (1663), Galilei (1656) on games of chance.
- Pascal (1623-1662) and Fermat (1601-1665) work on probability.
- Gauss and Laplace apply the probability models to astronomy.
- Galton introduces the notions of correlation and regression in biometry.

Advances in the XX century

- 1890-1920: mathematical statistics (Pearson, Yule and Gosset (UK); Borel, Fréchet and Poincaré (France); Chuprov and Markov (Russia)).
- 1921-1932: estimation theory (Fisher).
- 1940-1945: hypothesis testing (Neyman and Pearson), sampling theory (Neyman), experimental design (Fisher).
- 1958: control theory, system identification (Kalman).
- 1958-: neural networks (Rosenblatt,Widrow, Hoff).
- See *A History of Mathematical Statistics from 1750 to 1930* by Anders Hald or *Against the Gods: the remarkable history of risk* by P. L. Bernstein.

An integrated definition

We will adopt the integrated definition proposed by Vic Barnett in the book "Comparative statistical inference" (1999)

Definition

Statistics is the study of how information should be employed to reflect on, and give guidance for action, in a practical situation involving uncertainty.

This definition requires some clarification, specifically

- What is meant by uncertainty?
- What is meant by *situation involving uncertainty*?
- What is meant by *information*?
- What is the difference between the *reflection* and the *guidance* function of statistics?

Some examples of uncertain situations

- A student tackling an exam.
- A doctor prescribing a drug.
- An oil company deciding where to drill.
- A football trainer deciding who will shoot the decisive penalty.
- A financial investor in the NASDAQ trade market.
- An employee in front of an offer of a new job.

What is typical to uncertain situations?

- There is more than one possible outcome (e.g. success or failure).
- The actual outcome is unknown to us in advance: it is indeterminate and variable.
- We could be interested in knowing what that outcome will be.
- We could have to take a decision anyway.

Why are we interested to uncertainty?

- Uncertainty is pervasive in our world.
- We would like to know what the outcome will be (e.g. will the drug be successful in curing the patient?)
- We want to decide on a course of action relevant to, and affected by, that outcome (e.g. where is the oil company going to drill?, how much should I study to pass the exam?).

Why stochastic modeling?

- Any attempt to construct a theory to guide behavior in a situation involving uncertainty must depend on the construction of a formal model of such situations.
- This requires a formal notion of uncertainty.
- In this we will recur to the formalism of probability in order to represent uncertainty.
- In very general terms, a stochastic model is made of
  1. a statement of the set of possible outcomes of a phenomenon and
  2. a specification of the probabilistic mechanism governing the pattern of outcomes that might arise.
- Notions like independence, randomness, etc., has to be defined for distinguishing and characterizing the different situations in terms of their degree of uncertainty.

Models and reality

- A model is a formal (mathematical, logical, probabilistic ...) description of a real phenomenon.
- A model is an idealization of a real situation. No mathematical model is perfect.
- A model makes assumptions.
- The adequacy of a model depends on how valid and appropriate are the assumptions on which it is based.
- A model can be used to make deductions and take decisions.
- The biggest concern is how to define an adequate model, either as a description of the real situation or to suggest a reasonable course of action relevant to that situation.
- Different aspects of the same phenomenon may be described by different models (e.g. physical, chemical, biological...)

Two configurations involving a model

Consider

1. a real phenomenon P , e.g. the car traffic in a Brussels boulevard leading to Place Montgomery.
2. a probabilistic model M describing the number of cars per hour in a boulevard leading to Place Montgomery.

Deductive or probabilistic configuration. We use the model M to predict the number of cars passing through the square in the time interval $[t, t + \Delta t]$.

Inductive or statistic configuration. We collect measures in order to estimate the model M starting from the real observations.

Statistics and machine learning look backward in time (e.g. what model generated the data), probability is useful for deriving statements about the behavior of a phenomenon described by a probabilistic model.

Deduction and induction

- **Under the assumption that a probabilistic model is correct**, logical deduction through the ideas of mathematical probability leads to a description of the properties of data that might arise from the real situation.
- The theory of statistics is designed is designed to reverse the deductive process. It takes data that have arisen from a practical situation and uses the data to
  - estimate the parameters of a parametric model,
  - suggest a model,
  - validate a guessed model.
- Deduction is a form of reasoning that works from the general to the specific (top-down), drawing **necessary conclusions** from the premises. Induction is a form of reasoning that works from the specific to the general (bottom-up), drawing **probable conclusions** from the premises.
- Note that the inductive process is possible only because the "language" of probability is available to form the deductive link.

Rules of deduction and induction

An example of deductive rule is

> *All university professors are smart. I am listening to an university professor.*
> *So this professor is smart*

There is no possible way in which the premise can be true without the corresponding conclusion to be true. This rule is always valid since it never leads from true premises to false conclusions.

Rules of deduction and induction

An induction rule looks like

> *Until now, all university professors I met, were smart. I am listening to an university professor.*
> *So this professor is smart*

Note that in this case the premise could be true even though the conclusion is false. How to measure the reliability of this rule?

Black swan example: note that, *before discovery of Australia, people in the Old World were convinced that all swans were white.*

About inductive reasoning

- Though inductive reasoning is not logically valid, there is no doubt that we continuously carry out this kind of reasoning, and that our practical life would be impossible without it.
- The most obvious class of inductive claims are the ones concerning the future.
- Every human act relies on considerations about the future or estimations, though future is by definition not observable.
- Induction postulates what Hume calls the *uniformity of nature*: for the most part, if a regularity holds in my experience, then it holds in nature generally, or at least in the next instance.
- The truth of this principle cannot be logically demonstrated but the success of the human beings in using it is an empirical proof of its validity.
- *Could you logically prove that the sun will raise tomorrow?*

Interpretations of probability

Probability is the language of stochastic modeling and statistical machine learning. However, a variety of philosophical interpretations of the probability concept can exist.

- Frequentist: statistical analysis must be based on the use of sample data evaluated through a frequency concept of probability. Information comes only from repeated observations.
- Bayesian: wider concept of information than that of just sample data. Earlier experience or prior information must be also taken into consideration. This approach addresses the issue of combining prior information with sample data.
- Decision theory: it provides rules for action in situations of uncertainty. Uncertainty as strictly related to the notion of action. The assessment of consequences of alternative actions and their formal quantification through the concept of utility is central to this approach.

Considerations

- The three approaches address three general forms of information that may, depending on circumstances, be relevant to a statistical study.
- Behind the frequentist approach there is the intention to produce a theory which should be universal, free of subjective assessments and based on quantifiable elements.
- The three forms of information refer to three different time horizons: the prior information (Bayesian) accumulated from *past* experiences, the sample data (frequentist) arising from the *current* situation, and assessment of consequences referring to potential *future* action (decision theory).
- Specific statistical tools have been developed to model, incorporate, use these types of information.

Inference and decision making

- Any statistical procedure that utilizes information to obtain a description of the practical situation in probabilistic terms is called a *statistical inference procedure*.
- Any statistical procedure with the aim of suggesting action to be taken is a *statistical decision making procedure*.
- Decision-making extends the descriptive aims of inference by incorporating assessments of consequences.

The doctor example

- Consider a doctor prescribing a drug for a patient.
- Let us consider two possible outcomes (drug works, drug does not work) and suppose that uncertainty exists about the effect of this drug.
- How to model the uncertain effect of the drug?
- We can use a simple probabilistic model that assigns a probability p to the success and a probability $1 - p$ to the failure.
- How do we interpret the probability value p?
- In the frequentist interpretation p is the probability of success that the same drug had on "similar" patients in the past. The value p is then related to the proportion of successes, or potential successes, in a large population.

The doctor example (II)

- In the Bayesian interpretation, the probability p is regarded as a measure of the personal doctor's *degree of belief* in the success of the treatment.
- In the decision theory framework, the probability p has to be combined with a certain measure of utility u (or cost) to take the best action.

	p	$1-p$
	drug works	drug doesn't work
drug administrated	0	c_1
no drug	c_2	0