Random forests and decision trees: machine learning, empirical statistics and the challenge of interpretability

Matthew L. Jones
mjones@columbia.edu
@nescioquid

@nescioquid

NSA-GCHQ data mining

UK TOP SECRET STRAP1 COMINT
AUS/CAN/NZ/UK/US EYES ONLY

Reference: OPC-M/TECH.A/455 (v1.0, r206)
Date: 20 September 2011
Copy no:

HIMR Data Mining Research Problem Book

[redacted]

OPC-MCR, GCHQ

Summary
In this problem book we set out areas for long-term data mining research at the Heilbronn Institute for Mathematical Research starting in October 2011 and continuing for at least three years. The four areas are beyond supervised learning, information flow in graphs, streaming exploratory data analysis and streaming expiring graphs.

Friends of Random Forests

- "The NSA were very early adopters of Random Forests through direct contact with [redacted] via the NSA Statistical Advisory Group (NSASAG) [W31]."
  - steganography detection (Random Forest) [I74]
  - website classification (decision tree) [I36]
  - protocol classification (Random Forest and neural network) [W1]
  - spam detection (Random Forest) [I44]
  - payphone detection (Random Forest)

- [redacted] OPC-MCR-GCHQ. "HIMR Data Mining Research Problem Book," September 20, 2011.

Interpretability

- A problem with the use of Random Forests is that their decisions can not be simply and intuitively explained to an analyst. This black box nature can lower analyst trust in a prediction. [Redacted] (NSA R1) has been leading an effort to make Random Forests more interpretable. [I18].

  - [redacted] OPC-MCR-GCHQ. "HIMR Data Mining Research Problem Book," September 20, 2011.

Decision Trees

Table 2: A small training set of credit card applications.

Number	Attributes				Class
	account	balance	employed	monthly expense	
1	bank	700	yes	200	accept
2	bank	300	yes	600	reject
3	none	0	yes	400	reject
4	other inst	1200	yes	600	accept
5	other inst	800	yes	600	reject
6	other inst	1600	yes	200	accept
7	bank	3000	no	300	accept
8	none	0	no	200	reject

"Training set"

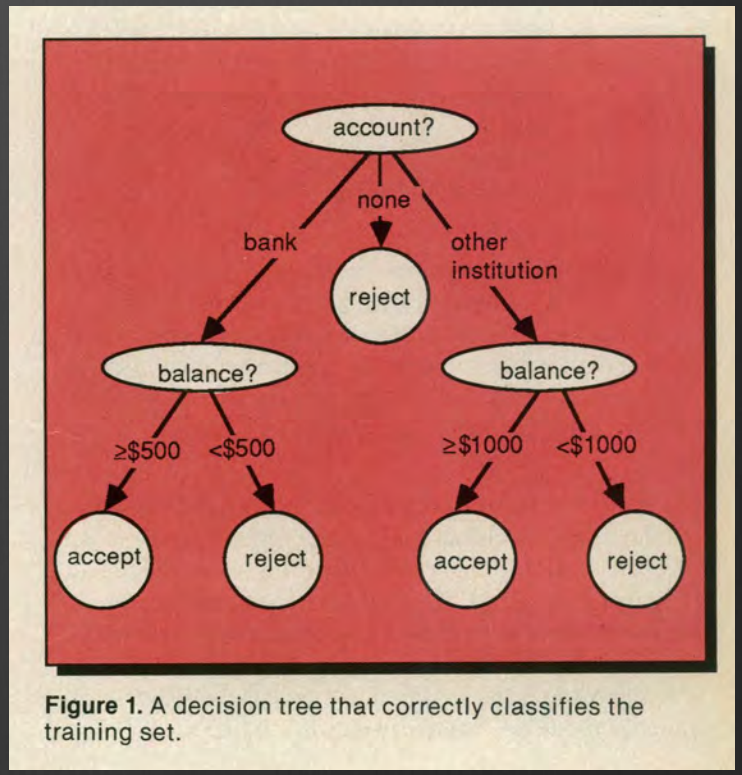


Figure 1. A decision tree that correctly classifies the training set.

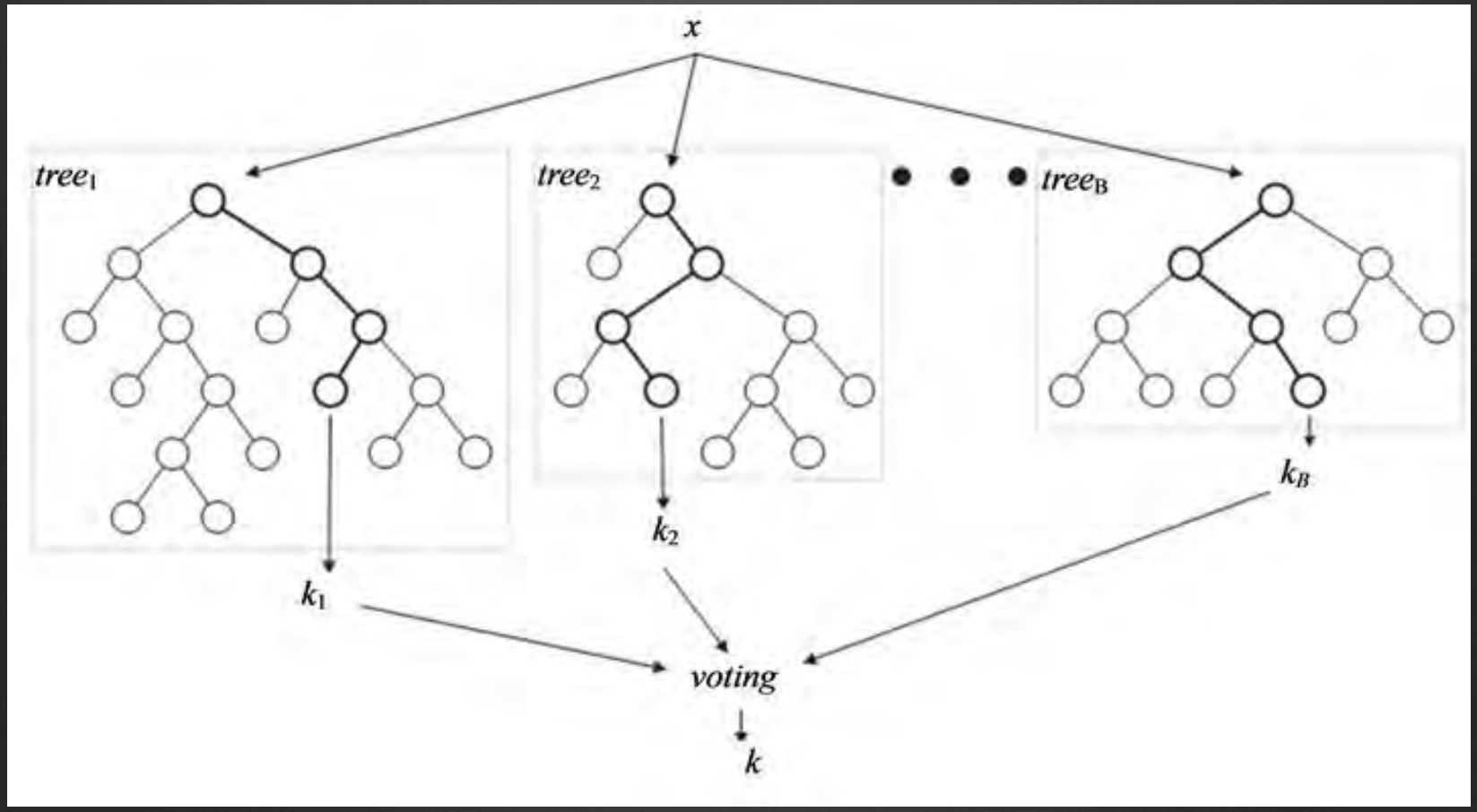
Carter, Chris, and Jason Catlett. "Assessing Credit Card Applications Using Machine Learning." *IEEE Expert* 2, no. 3 (September 1987): 71–79.

@nescioquid

Decision Trees

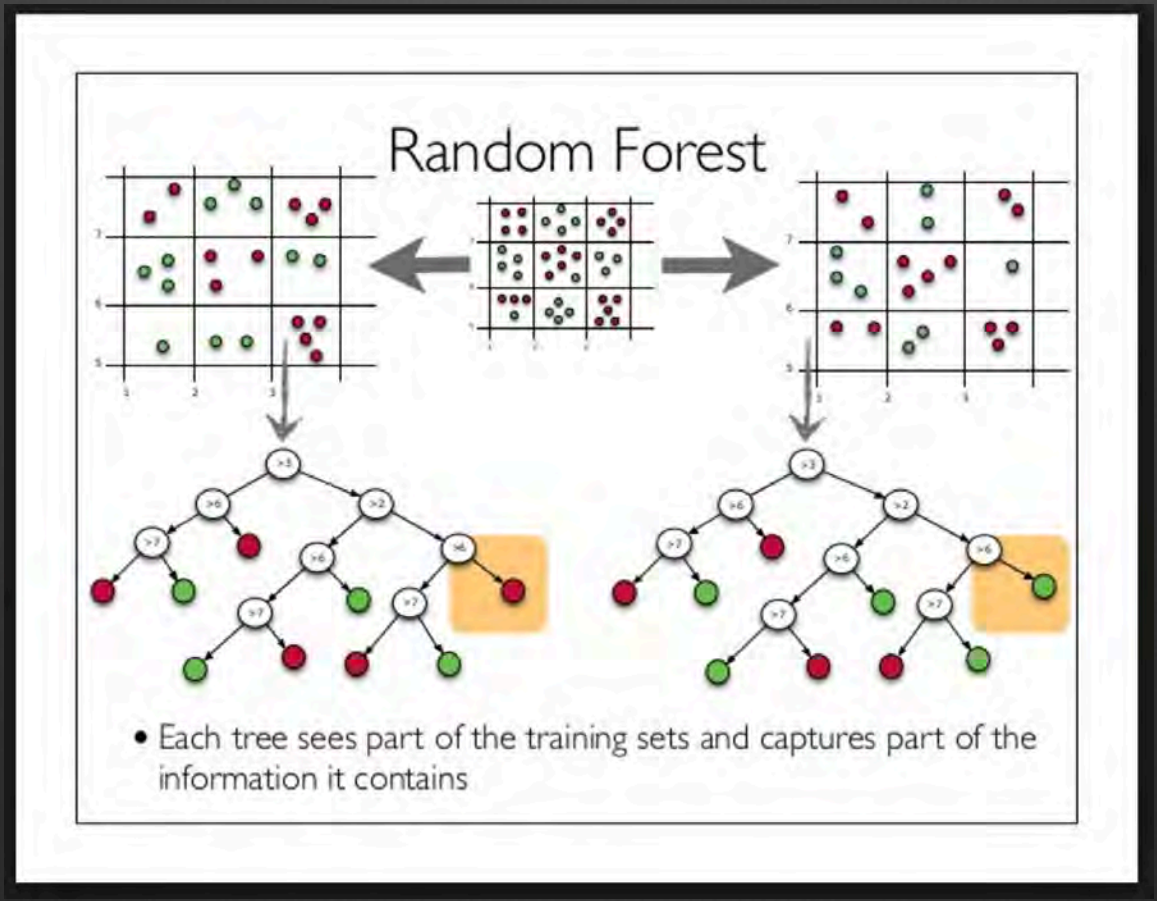
- Intuitive and highly interpretable
  - Crucial for numerous industries that must be able to justify their decision-making: credit, insurance, loans
- Moderately predictive
- Computable in polynomial time
- Extremely prone to overfit by becoming too "bushy"
  - i.e. take noise in "training" data for signal

Random forest



http://file.scirp.org/Html/6-9101686_31887.htm

Random forest



@nescioquid http://www.slideshare.net/0xdata/jan-vitek-distributedrandomforest522013

@nescioquid

Data mining

dba

Knowledge discovery in databases (1989-2005?)

Predictive analytics (2005+)

Data science (2010-?)?

&c. &c. &c.

Inadequate stats

. . . scientists can reformulate and rerun their experiments should they find that the initial design was inadequate. Database managers rarely have the luxury of redesigning their data fields and recollecting the data.

Frawley, Piatetsky-Shapiro and Matheus 1991, p. 8.

Critique of artificial reason

Not to alter our criteria for the analysis of vast data is

- to deny ourselves what we potentially could know
- to indulge in a luxury we don't have

- Challenge to epistemic virtues of
  - Machine learning (AI)
  - Academic statistics
  - Database practitioners
  - Spooky spooks

Volumetric determinism

- Volume of high dimensional data requires OR CAUSES
  - Development of new algorithms
  - Loosening of traditional account of statistical modeling
  - Creation of new epistemic virtues
  - Creation of new experts
  - Creation of new expectations of privacy

  - That is,
  - A “new paradigm”

Promise and Threat

Analysis of massive amounts of high-dimensional data

NOT just generalizations about aggregates

BUT better understanding of individuals

Amazon Recommendations

Political Campaigns

Telephony metadata

Wellsprings

- “Data Mining, or Knowledge Discovery in Databases as it is also called, is claimed as an offspring by three disciplines: databases, statistics, and the machine learning subfield of artificial intelligence.”

  --J. Ross Quinlan, 1999

Inadequate stats

. . . scientists can reformulate and rerun their experiments should they find that the initial design was inadequate. Database managers rarely have the luxury of redesigning their data fields and recollecting the data.

Frawley, Piatetsky-Shapiro and Matheus 1991, p. 8.

Inadequate AI

- "knowledge acquisition bottleneck"
  - "Human experts find it difficult to express their knowledge, . . . in terms of concise situation-action rules. If pressed to do so, they typically produce rules that are incorrect . . . The articulation of specific intuitive knowledge into deterministic rules is a difficult, sometimes unrealistic, problem for human experts."
  - Fayyad, U. M., K. B. Irani, J. Cheng, and Z Quin. "Machine Learning of Expert System Rules: Applications to Semiconductor Manufacturing." In *Collected Notes on the Workshop for Pattern Discovery in Large Databases (NASA Ames, January 14-15, 1991)*, 17–29. NASA Ames Research Center, 1991.

Inadequate databases

If I were to draw on a historical analogy of where we stand today with regards to digital information manipulation, navigation, and exploitation, I find myself thinking of Ancient Egypt. … A large data store today, in practice, is not very far from being a grand, write-only, data tomb.

> Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery," (1998) 48.

@nescioquid

Decision Trees

Table 2: A small training set of credit card applications.

Number	Attributes				Class
	account	balance	employed	monthly expense	
1	bank	700	yes	200	accept
2	bank	300	yes	600	reject
3	none	0	yes	400	reject
4	other inst	1200	yes	600	accept
5	other inst	800	yes	600	reject
6	other inst	1600	yes	200	accept
7	bank	3000	no	300	accept
8	none	0	no	200	reject

"Training set"

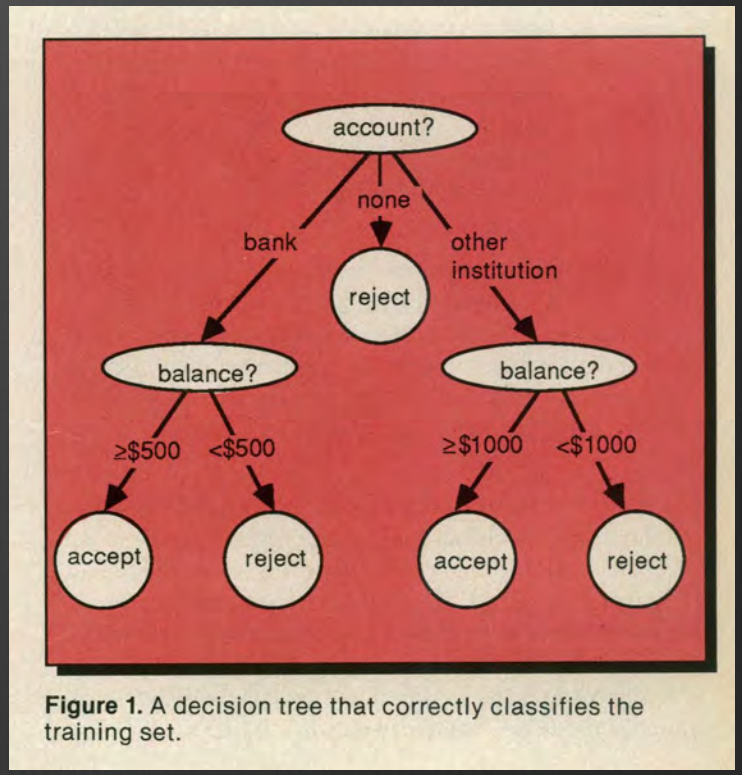


Figure 1. A decision tree that correctly classifies the training set.

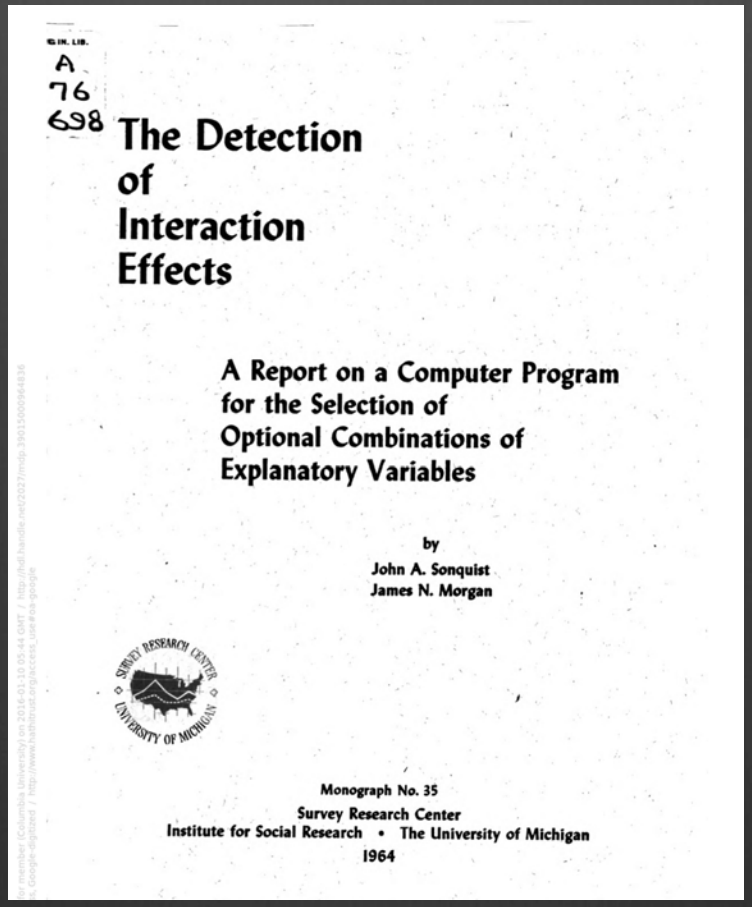
Carter, Chris, and Jason Catlett. "Assessing Credit Card Applications Using Machine Learning." *IEEE Expert* 2, no. 3 (September 1987): 71–79.

@nescioquid

Decision Trees

- Intuitive and highly interpretable
  - Crucial for numerous industries that must be able to justify their decision-making: credit, insurance, loans
- Moderately predictive
- Computable in polynomial time
- Extremely prone to overfit by becoming too "bushy"
  - i.e. take noise in "training" data for signal

Institute for Social Research



G.IN. LIB.
A
76
698

The Detection of Interaction Effects

A Report on a Computer Program for the Selection of Optional Combinations of Explanatory Variables

by
John A. Sonquist
James N. Morgan

SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

Monograph No. 35
Survey Research Center
Institute for Social Research • The University of Michigan
1964

@nescioquid

“Automatic Interaction Detection” (AID)

The motivation for the development of the computer program described in this report comes from two sources. First, is the belief that the multivariate statistical techniques in common usage are often inadequate for the analysis of the rich body of data from a cross section sample survey, and second is the conviction that a large-scale digital computer can be used for more than just a high-speed adding machine.

Statistical Restraint

We have tried to break away from the habit of asking the question, "What is the effect of x on y when everything else is held constant?" This has been replaced with, "What do I need to know most in order to reduce predictive error a maximum amount?"

This is the type of question that might be asked by a research scientist working in a substantive area in which theory is not yet very precise. Once he receives an answer, he may well ask, "Now that I know this, what additional information would help to reduce predictive error still further?" and so on. He would certainly ask other questions as

Ubiquitous now, not then

- **1963-70s: laughed out of water**
- Decision trees "were called 'a recipe for learning something wrong'" says Dan Steinberg. "This was a death sentence, like a restaurant with *E. Coli.* Trees were finished."

Data mining as pejorative

- “. . . proceeding via a ‘dustbowl’ empiricism is dangerous at worst and foolish at best . . . . The purely empirical approach is particularly dangerous in an age when computers and packaged programs are readily available, since there is temptation to substitute immediate empirical analysis for more analytic thought and theory building.”
  - Einhorn, “Alchemy in the Behavioral Sciences,” 1972

@nescioquid

Revivifying decision trees

Two fold

Critiques of artificial reason

Empirical machine learning

Critique of Axiomatic, model-oriented Statistics

Stats, not Maths

Decision trees: Stats

- Leo Breiman
  - Consultant for DOD contractor after leaving academia
  - Technology Services Corporation
- Breiman: “We were working on prediction problems like next day ozone in the Los Angeles basin, carbon monoxide levels on freeways, but also things such as could we recognize the sender of handset Morse code—this was something we were doing for the spook agencies—or could we recognize from sonar returns whether the other submarine was Russian or American?” (in Ohlsen, 188).

Classifiers

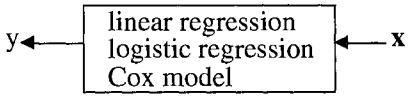
- "Classifiers are not constructed whimsically. They are based on past experience. … Los Angelenos know that one hot, high pollution day is likely to be followed by another." (CART,4)

Decision trees: CART

- Data-focused academic 1963 then late 1980s-1990s
- Breiman: "*Alice in Wonderland*. That is, I knew what was going on out in industry and government in terms of uses of statistics, but what was going on in academic research seemed light years away. It was proceeding as though it were some branch of abstract mathematics." (in Ohlsen, 196)
- Epistemic virtue: Predictive accuracy >> causal model

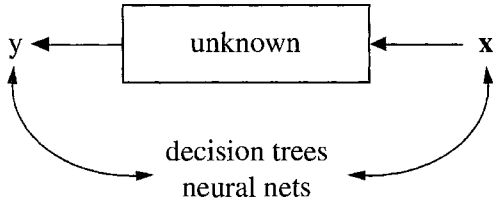
Two statistical cultures

"Data modeling culture"



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

"Algorithmic modeling culture"

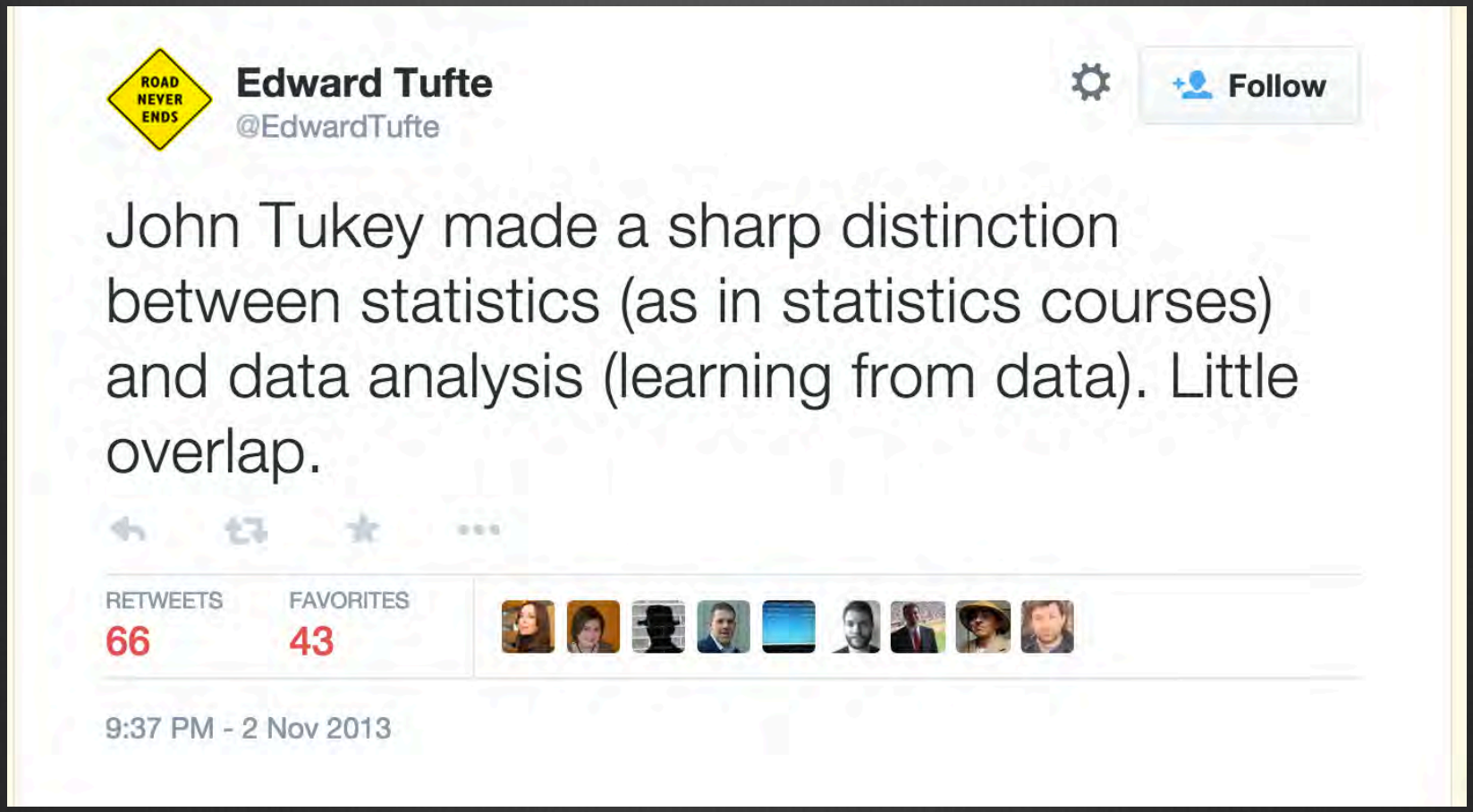


Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Approaching problems by looking for a data model imposes an a priori straightjacket that restricts the ability of statisticians to deal with a wide range of statistical problems.

Breiman, 2001, p. 204.

Tufte


ROAD NEVER ENDS
Edward Tufte
@EdwardTufte
Follow
John Tukey made a sharp distinction between statistics (as in statistics courses) and data analysis (learning from data). Little overlap.
RETWEETS
66
FAVORITES
43
9:37 PM - 2 Nov 2013

Tukey: hero of big data

Data analysis, […], must then take on the characteristics of a science rather than those of mathematics, specifically:

(1) Data analysis must seek for scope and usefulness rather than security.

(2) Data analysis must be willing to err moderately often […].

(3) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proofs or stamps of validity.

Tukey, "Future of Data Analysis," 1961, III:401

Tukey: judgment, not rules

"If data analysis is to be well done, much of it must be a matter of judgment, and 'theory', whether statistical or non-statistical, will have to guide, not to command."

Tukey, "Future of Data Analysis," 1961, III:401

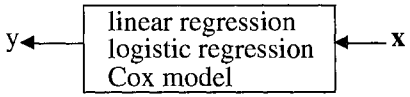
War time data work

Thanks to “war problems,” “it was natural to regard statistics as something that had the purpose of being used on data—maybe not directly, but at most at some remove. Now, I can’t believe that other people who had practical experience failed to have this view, but they certainly—I would say—failed to advertise it.”

Tukey interview, 85.

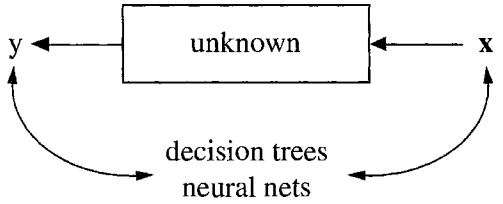
Two statistical cultures

"Data modeling culture"



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

"Algorithmic modeling culture"



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Approaching problems by looking for a data model imposes an a priori straightjacket that restricts the ability of statisticians to deal with a wide range of statistical problems.

Breiman, 2001, p. 204.

Beyond Prediction

Classifications from data have two purposes:

(1) “to predict the response variable corresponding to future measurement vectors as accurately as possible;

(2) to understand the structural relationships between the response and the measured variables.” (CART, ?)

@nescioquid

Revivifying decision trees

Two fold

Critiques of artificial reason

Empirical machine learning

Critique of Axiomatic, model-oriented Statistics

Empirical Machine Learning

Machine intelligence chez Donald Michie

Mastery is not acquired by reading books — it's acquired by trial-and-error and teacher-supplied examples. That is how humans acquire skill. People are very reluctant to accept this. Their reluctance tells us something about the philosophical self-image that we, as thinking beings, prefer. It tells us nothing about what actually happens when a teacher or a master trains somebody. That somebody has to regenerate rules from example to make them an intimate part of his intuitive skill.

Humans who already have a skill value how-to-do-it manuals, but they keep them on the shelf as reference texts. It may help in the more scholarly side of training others. There are a number of good examples where a machine-generated how-to-do-it text has been adopted by professionals, but none where it constituted the means whereby they became professional practitioners.

Induction from databases

Induction Over Large Data Bases

J. R. Quinlan
Basser Department of Computer Science
University of Sydney

Abstract: Techniques for discovering rules by induction from large collections of instances are developed. These are based on an iterative scheme for dividing the instances into two sets, only one of which needs to be randomly accessible, These techniques have made it possible to discover complex rules from data bases containing many thousands of instances. Results of several experiments using them are reported,

Keywords: Induction, Inference, pattern recognition, rule formation, concept learning, decision trees.

Decision trees from ML

- Ross Quinlan
- Avoiding the "knowledge acquisition bottleneck"
  - "Human experts find it difficult to express their knowledge, or explain their actions, in terms of concise situation-action rules. If pressed to do so, they typically produce rules that are incorrect, or that have many exceptions. The articulation of specific intuitive knowledge into deterministic rules is a difficult, sometimes unrealistic, problem for human experts."
  - Fayyad, U. M., K. B. Irani, J. Cheng, and Z Quin. "Machine Learning of Expert System Rules: Applications to Semiconductor Manufacturing." In *Collected Notes on the Workshop for Pattern Discovery in Large Databases (NASA Ames, January 14-15, 1991)*, 17–29. NASA Ames Research Center, 1991.

Decision trees from ML

- Empirical Machine Learning (Quinlan)
  - Introduction of *entropy* as splitting criterion for making trees
- Examples: Chess endgames
- Symbolic RULES for action
  - Comprehensibility CRUCIAL epistemic value

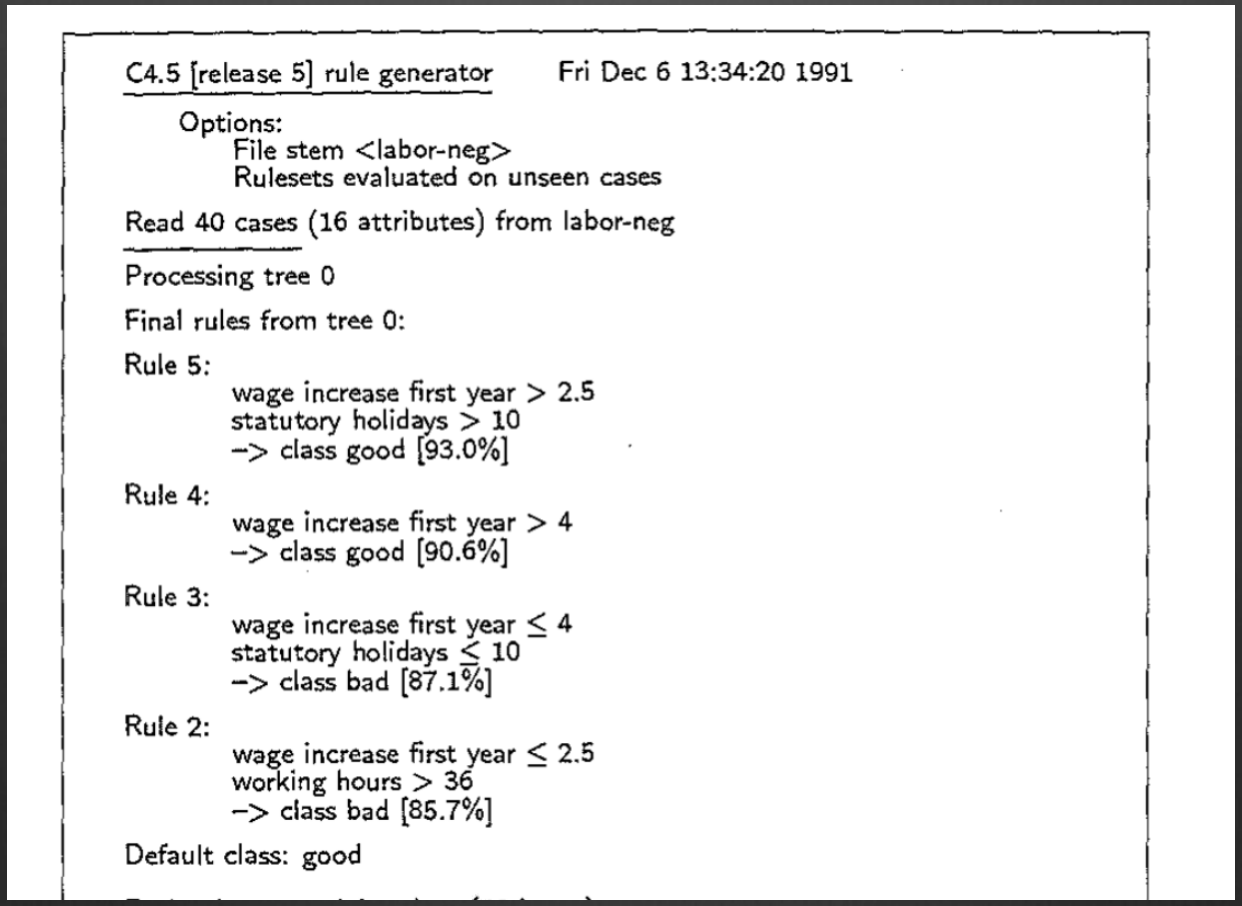
ML: interpretability

- “As important as a good fit to the data, is a property that can be termed “mental fit”. As statisticians, Breiman and colleagues (1984) see data-derived classifications as serving “two purposes: (1) to predict the response variable corresponding to future measurement vectors as accurately as possible; (2) to understand the structural relationships between the response and the measured variables.” (Feng and Michie, Machine Learning of Rules and Trees, 51)

ML: interpretability

- [continues] “The soybean rules were sufficiently meaningful to the plant pathologist associated with the project that he eventually adopted them in place of his own previous reference set. *ML requires that classifiers should not only classify but should also constitute explicit concepts, that is, expressions in symbolic form meaningful to humans and evaluable in the head*.” (Feng and Michie, Machine Learning of Rules and Trees, 51, my italcis)

ML rule sets



C4.5 [release 5] rule generator Fri Dec 6 13:34:20 1991

Options:
File stem <labor-neg>
Rulesets evaluated on unseen cases

Read 40 cases (16 attributes) from labor-neg

Processing tree 0

Final rules from tree 0:

Rule 5:
wage increase first year > 2.5
statutory holidays > 10
-> class good [93.0%]

Rule 4:
wage increase first year > 4
-> class good [90.6%]

Rule 3:
wage increase first year ≤ 4
statutory holidays ≤ 10
-> class bad [87.1%]

Rule 2:
wage increase first year ≤ 2.5
working hours > 36
-> class bad [85.7%]

Default class: good

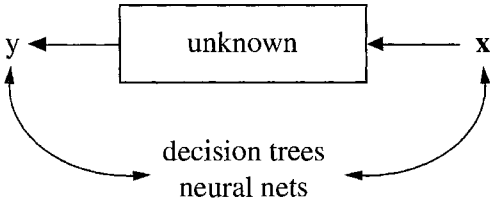
@nescioquid

Two Algorithmic Cultures

Machine learning: comprehensibility

- Conversion to RULES: "C4.5 also contains a mechanism to re-express decision trees as ordered lists of if-then rules. […] There are substantially fewer final rules than there are leaves, and yet the accuracy of the tree and the derived rules is similar. Rules have the added advantage of being more easily understood by people." (Kohavi and Quinlan, 1999, 12)

Statisticians: prediction



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Ramification of Trees

- 1990s
  - trees among the major candidate algorithms for KDD
  - Simple to understand and track record
- Late 1990s
  - tree mining among multiple databases
- 2000s
  - Random Forests
  - Ensemble modeling

Ramification of Trees: Database values

- Major effort to make trees scale in late 1990s
- SPRINT- "Scalable PaRallelizable Induction of Decision Trees" (1996)
  - IBM Almaden lab
  - "With the recent emergence of the field of data mining, there is a great need for algorithms for building classifiers that can handle very large databases. . . . By eschewing the need for any centralized memory-resident data structures, SPRINT efficiently allows classification of virtually any sized dataset." (554)

Map Reduce Random Forest

- COMET (Cloud Of Massive Ensemble Trees)
- "a single-pass MapReduce algorithm for learning on large-scale data. It builds multiple random forest ensembles on distributed blocks of data and merges them into a mega-ensemble. This approach is appropriate when learning from massive-scale data that is too large to fit on a single machine."
  - arXiv:1103.2068v2, Sandia Lab group, cited as an "internal publication" in [redacted] OPC-MCR-GCHQ. "HIMR Data Mining Research Problem Book," September 20, 2011.

Interpretability

- A problem with the use of Random Forests is that their decisions can not be simply and intuitively explained to an analyst. This black box nature can lower analyst trust in a prediction. [Redacted] (NSA R1) has been leading an effort to make Random Forests more interpretable. [I18].

  - [redacted] OPC-MCR-GCHQ. "HIMR Data Mining Research Problem Book," September 20, 2011.