Relative importance of steps

Jeffrey Leek, Assistant Professor of Biostatistics

May 18, 2016

Relative order of importance

question $>$ data $>$ features $>$ algorithms

An important point

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

John Tukey

Garbage in = Garbage out

question -> input data -> features -> algorithm -> parameters -> evaluation

1. May be easy (movie ratings -> new movie ratings)
2. May be harder (gene expression data -> disease)
3. Depends on what is a “good prediction”.
4. Often more data > better models
5. The most important step!

Features matter!

question -> input data -> features -> algorithm -> parameters -> evaluation

Properties of good features

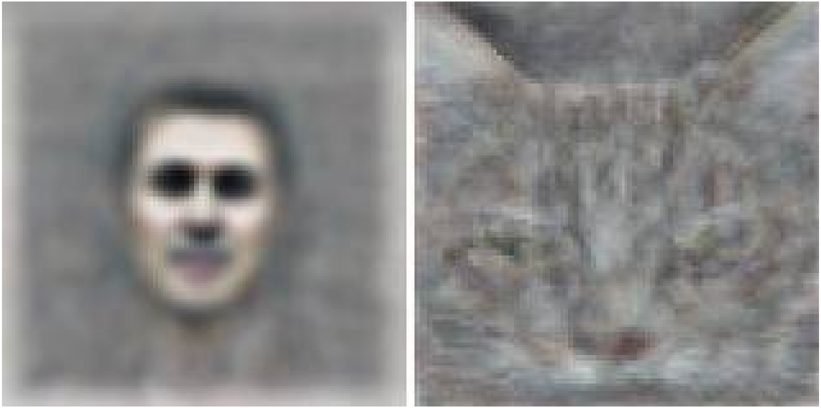
- Lead to data compression
- Retain relevant information
- Are created based on expert application knowledge

Common mistakes

- Trying to automate feature selection
- Not paying attention to data-specific quirks
- Throwing away information unnecessarily

May be automated with care

question -> input data -> features -> algorithm -> parameters -> evaluation



http://arxiv.org/pdf/1112.6209v5.pdf

Algorithms matter less than you'd think

question -> input data -> features -> algorithm -> parameters -> evaluation

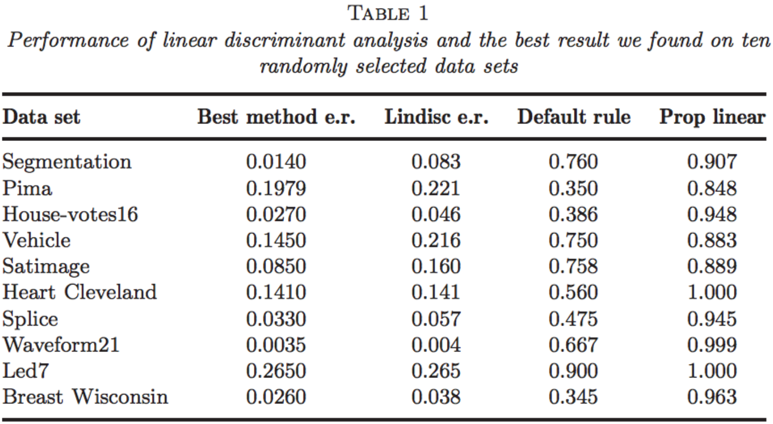


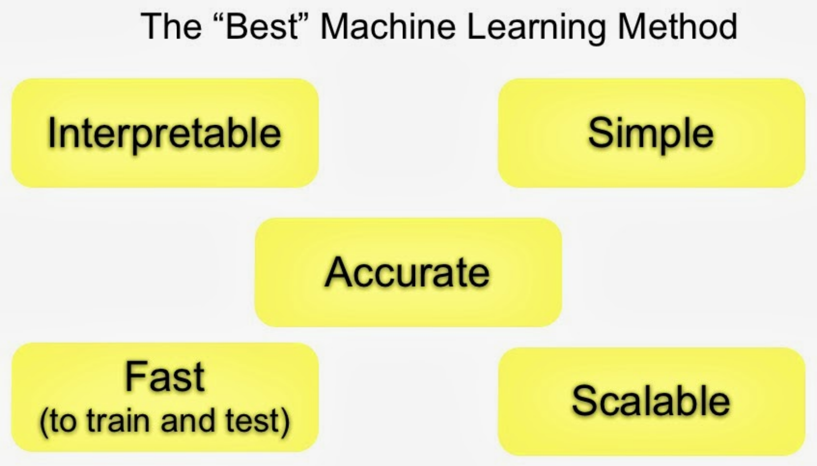
TABLE 1

Performance of linear discriminant analysis and the best result we found on ten randomly selected data sets

Data set	Best method e.r.	Lindisc e.r.	Default rule	Prop linear
Segmentation	0.0140	0.083	0.760	0.907
Pima	0.1979	0.221	0.350	0.848
House-votes16	0.0270	0.046	0.386	0.948
Vehicle	0.1450	0.216	0.750	0.883
Satimage	0.0850	0.160	0.758	0.889
Heart Cleveland	0.1410	0.141	0.560	1.000
Splice	0.0330	0.057	0.475	0.945
Waveform21	0.0035	0.004	0.667	0.999
Led7	0.2650	0.265	0.900	1.000
Breast Wisconsin	0.0260	0.038	0.345	0.963

`http://arxiv.org/pdf/math/0606441.pdf`

Issues to consider



The "Best" Machine Learning Method

Interpretable

Simple

Accurate

Fast
(to train and test)

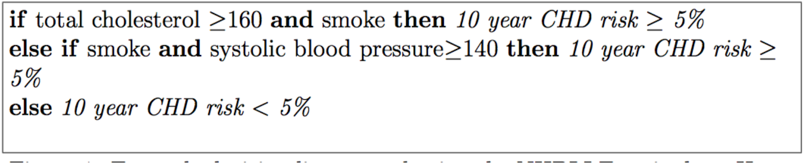
Scalable

http://strata.oreilly.com/2013/09/gaining-access-to-the-best-machine-learning-methods.html

Prediction is about accuracy tradeoffs

- Interpretability versus accuracy
- Speed versus accuracy
- Simplicity versus accuracy
- Scalability versus accuracy

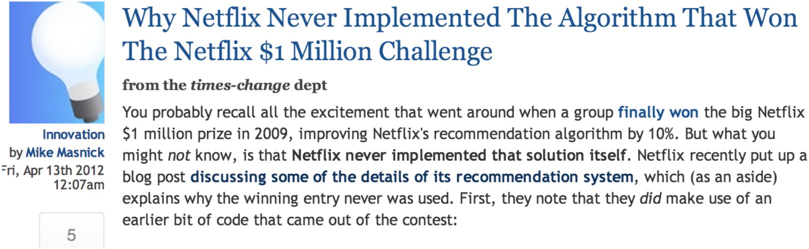
Interpretability matters



if total cholesterol ≥160 **and** smoke **then** *10 year CHD risk* $\geq$ *5%*
else if smoke **and** systolic blood pressure≥140 **then** *10 year CHD risk* $\geq$ *5%*
else *10 year CHD risk* $<$ *5%*

`http://www.cs.cornell.edu/~chenhao/pub/mldg-0815.pdf`

Scalability matters



Why Netflix Never Implemented The Algorithm That Won The Netflix $1 Million Challenge

from the *times-change* dept

Innovation
by Mike Masnick
Fri, Apr 13th 2012
12:07am

5

You probably recall all the excitement that went around when a group **finally won** the big Netflix $1 million prize in 2009, improving Netflix's recommendation algorithm by 10%. But what you might *not* know, is that **Netflix never implemented that solution itself.** Netflix recently put up a blog post **discussing some of the details of its recommendation system**, which (as an aside) explains why the winning entry never was used. First, they note that they *did* make use of an earlier bit of code that came out of the contest:

`http://www.techdirt.com/blog/innovation/articles/20120409/03412518422/`

`http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html`