Machine learning

Brian Caffo, Jeff Leek, Roger Peng

@bcaffo

www.bcaffo.com

Machine learning is a set of algorithms that can take a set of inputs (data) and return a prediction

Non-exhaustive list of ML activities:
Unsupervised learning
Supervised learning

Unsupervised learning: trying to uncover unobserved factors; clustering, mixture models, principal components

WIKIPEDIA
The Free Encyclopedia

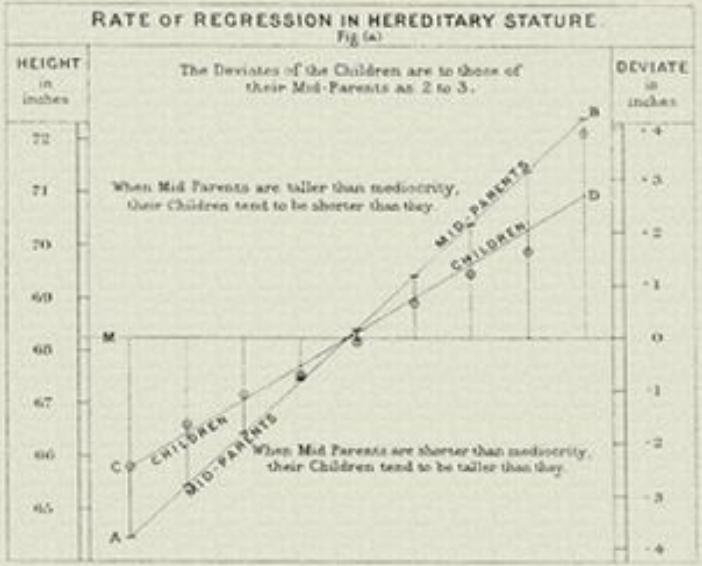
Article Talk

g factor (psychometrics)

From Wikipedia, the free encyclopedia

Supervised learning: using a collection of predictors and some observed outcomes to build an algorithm to predict the outcome when it is not observed; random forests, boosting, SVMs

Regression Models for Data Science In R



Brian Caffo

Contrasting machine learning and traditional statistical analyses

Machine learning

- Emphasize predictions
- Evaluates results via prediction performance
- Concern for overfitting but not model complexity per se
- Emphasis on performance
- Generalizability is obtained through performance on novel datasets
- Usually no superpopulation model specified
- Concern over performance and robustness

Traditional statistical analyses

- Emphasizes superpopulation inference
- Focuses on a-priori hypotheses
- Simpler models preferred over complex ones (parsimony), even if the more complex models perform slightly better
- Emphasis on parameter interpretability
- Statistical modeling or sampling assumptions connects data to a population of interest
- Concern over assumptions and robustness

Machine learning

- build an automated movie recommender system
- success - anything that produces reliable recommendations

Statistical analysis

- build a parsimonious and interpretable model to better understand why people choose the movies that they do
- success - anything true learned about movie choices

Examples

Improve Healthcare, Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)

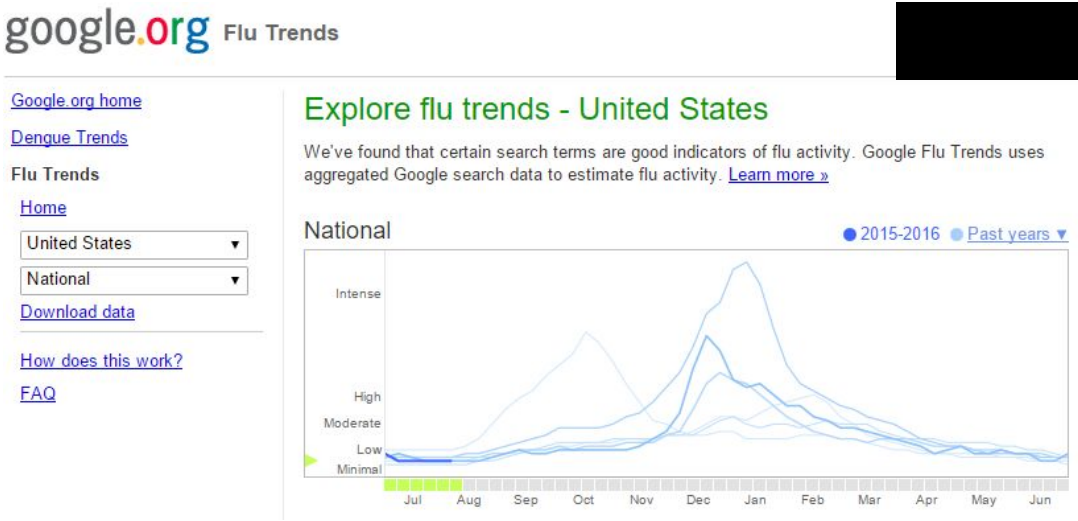
Machine learning

- build an automated system for predicting hospital stays from previous claims
- success - anything that produces reliable predictions

Statistical analysis

- build a parsimonious and interpretable model to better understand why people stay in the hospital longer
- success - anything true learned about hospital stays

Google flu trends



Machine learning

- build an automated system for predicting flu outbreaks
- success - anything that produces reliable predictions

Statistical analysis

- build a parsimonious and interpretable model to better understand flu outbreaks
- success - anything true learned about the flu

Lessons learned

- Both approaches are extremely valuable and have their place
- Amount of tolerable model/algorithm complexity changes dramatically between the approaches
- Goals of the approaches are very different
    - Caveat - there's a fair amount of work on making machine learning algorithms more interpretable and work on producing better predictions from more traditional approaches
    - (Meeting in the middle?)

Normal Deviate

Thoughts on Statistics and Machine Learning

RISE OF THE MACHINES
Larry Wasserman

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

“In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;
- Kept statisticians from using more suitable algorithmic models;
- Prevented statisticians from working on exciting new problems;”

Comment

D. R. Cox

Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question or a scientific hypothesis, although I would be surprised if this were a real source of disagreement. These issues may evolve, or even change

Statistical Science
2006, Vol. 21, No. 1, 1–14
DOI 10.1214/088342306000000060

Classifier Technology and the Illusion of Progress

David J. Hand

“so that the apparent superiority of more sophisticated methods may be something of an illusion. In particular, simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.”