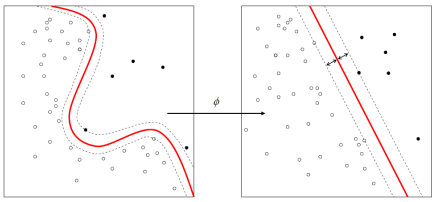


Introduction to Machine Learning

Two-day workshop
Duke University
Adam Chekroud

McCarthy Lab!
(2014)

Background

- BA Experimental Psych, MSc Neuroscience (Oxford)
- PhD Psychology (Neuroscience)

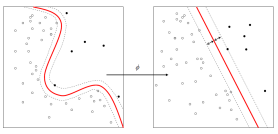
Interests

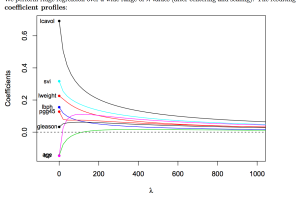
- Computational approaches to psychiatric illness
  - Optimising treatment selection in depression

Objectives

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(|$
Algorithm:

- Initialize model with a constant value:
  - $F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$.
- For m = 1 to M:
  - Compute so-called ''pseudo-residuals'':
    - $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, \ldots$
  - Fit a base learner $h_m(x)$ to pseudo-residuals, i.e. train
  - Compute multiplier γ_m by solving the following one-d
    - $\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$.
  - Update the model:
    - $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.
- Output $F_M(x)$.





Objectives

Thursday

- Introduction to ML
- Pipeline overview
- Feature selection (theory, code, practice)
- Algorithms (theory, code)
- Performance & **~Validation~**
- Review

Friday

- Practical session and project brainstorming
- Research talk (@ DIBS)

What is machine learning?

- *"The goal of machine learning is to program computers to use example data or past experience to solve a given problem"*
- *Machine learning is a subfield of computer science that … explores the study and construction of algorithms that … can learn from and make predictions on data"*
- *"… algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions"*
    - Developing computer algorithms for doing — usually predicting — stuff.

What is machine learning?

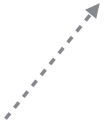
Doing [stuff]:

- Systems that analyze past sales data to predict customer behavior
- Optimize robot behavior so that a task can be completed using minimum resources
- **Extract knowledge from bioinformatics data**
    - Diverse range of methods and approaches
        - Plays a central (often hidden) role in our life

Many faces of machine learning

- Classification
- Clustering
- Dimensionality reduction
- Graphical models
- Genetic algorithms
- Deep learning
- Also many others that haven't yet filtered through to biological sciences: recommendation engines, bayesian networks, association learning, similarity matching

  - *ML is not just classification!*



Reinforcement Learning

A machine learns to interact with its (dynamic) environment (e.g. playing a computer game, driving a car)

Machine learning

"Supervised" Learning

"Unsupervised" Learning

Machine learning

<u>"Supervised" Learning</u>

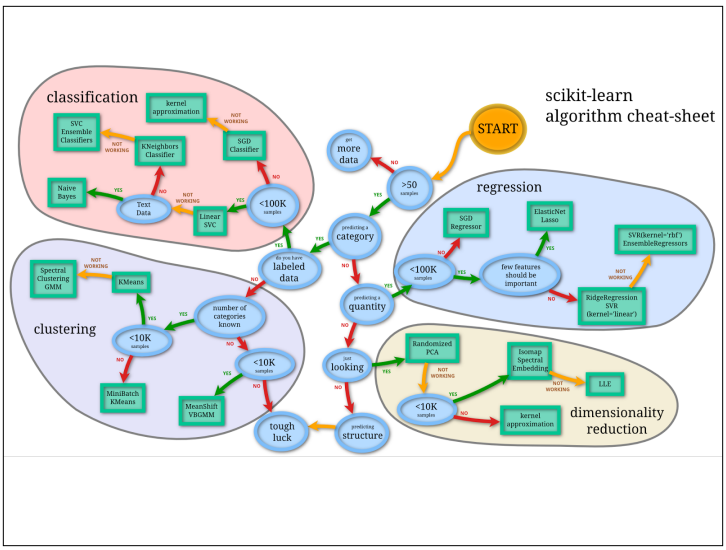
(a.k.a prediction, classification)

- We have labelled training data
- Use data to train algorithm
- Use algorithm to make predictions

<u>"Unsupervised" Learning</u>

(inc. clustering, PCA)

- No labels on our data
- Use machines to *organise* or *find hidden structure* in our data
- Use this organisation to help our understanding



The Perfect Milk Machine: How Big Data Transformed the Dairy Industry

ALEXIS C. MADRIGAL | MAY 1, 2012 | TECHNOLOGY

Dairy scientists are the Gregor Mendels of the genomics age, developing new methods for understanding the link between genes and living things, all while quadrupling the average cow's milk production since your parents were born.

REVIEWS

Machine learning applications in genetics and genomics

Maxwell W. Libbrecht[1] and William Stafford Noble[1,2]

Abstract | The field of machine learning, which aims to develop computer algorithms that improve with experience, holds promise to enable computers to assist humans in the analysis of large, complex data sets. Here, we provide an overview of machine learning applications for the analysis of genome sequencing data sets, including the annotation of sequence elements and epigenetic, proteomic or metabolomic data. We

Libbrecht & Noble, 2015, *Nat. Rev. Genetics*

REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case.

Early example of ML in biology
- Discover new cancer classes
- Assign tumors to classes
  - ~11k citations!

Golub et al. 1999, *Science*

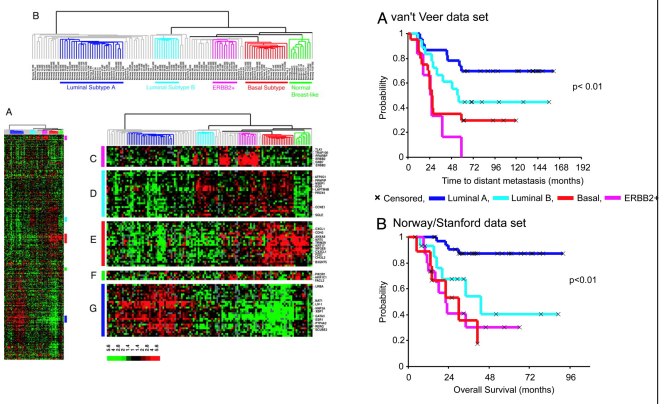
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[i], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

Repeated observation of breast tumor subtypes in independent gene expression data sets

Therese Sørlie*, Robert Tibshirani†, Joel Parker‡, Trevor Hastie§, J. S. Marron¶, Andrew Nobel¶, Shibing Deng‖, Hilde Johnsen**, Robert Pesich*, Stephanie Geisler††, Janos Demeter*, Charles M. Perou‡,‡‡, Per E. Lønning††, Patrick O. Brown§§, Anne-Lise Børresen-Dale**, and David Botstein*¶‖

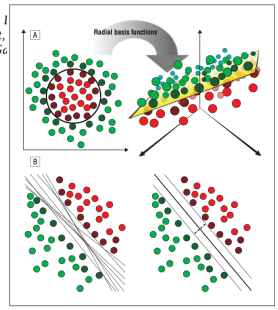
Sorlie et al. 2001; 2003, *PNAS*



ORIGINAL ARTICLE

Use of Neuroanatomical Pattern Classification to Identify Subjects in At-Risk Mental States of Psychosis and Predict Disease Transition

Nikolaos Koutsouleris, MD; Eva M. Meisenzahl, MD; Christos l
Thomas Frodl, MD; Johanna Scheuerecker, BA; Gisela Schmitt,
Maximilian Reiser, MD; Hans-Jürgen Möller, MD; Christian Ga



- Can we use a brain scan to predict conversion to psychosis?
  - Support vector machine classifier (86% accuracy)

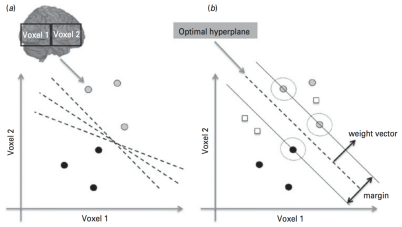
Koutsouleris et al, 2009. *Archives*

Psychological Medicine (2012), **42**, 1037–1047. © Cambridge University Press 2011
doi:10.1017/S0033291711002005

ORIGINAL ARTICLE

Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study

J. Mourao-Miranda[1,2], A. A. T. S. Reinders[3,4], V. Rocha-Rego[1], J. Lappin[3], J. Rondina[1], C. Morgan[3], K. D. Morgan[3], P. Fearon[3], P. B. Jones[5], G. A. Doody[6], R. M. Murray[3], S. Kapur[3] and P. Dazzan[3,7]*



doi:10.1093/brain/aws084 Brain 2012: 135; 1508–1521 | 1508

BRAIN
A JOURNAL OF NEUROLOGY

Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder

Benson Mwangi,[1] Klaus P. Ebmeier,[2] Keith Matthews[1] and J. Douglas Steele[1]

Psychological Medicine (2015), **45**, 2805–2812. © Cambridge University Press 2015
doi:10.1017/S0033291715000768

ORIGINAL ARTICLE

Identifying neuroanatomical signatures of anorexia nervosa: a multivariate machine learning approach

L. Lavagnino[1]*, F. Amianto[2], B. Mwangi[1], F. D'Agata[2], A. Spalatro[2], G. B. Zunta-Soares[1], G. Abbate Daga[2], P. Mortara[2], S. Fassino[2] and J. C. Soares[1]

[1] *UT Center of Excellence on Mood Disorders, Department of Psychiatry and Behavioral Sciences, UT Houston Medical School, Houston, TX, USA*
[2] *Department of Neuroscience, AOU San Giovanni Battista, Turin, Italy*

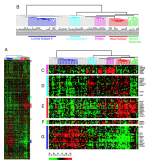
Koutsouleris et al, 2009. *Archives*

Clinical Questions

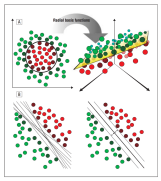
1. Prognosis:
   A. Can we predict who will develop a disease based on a baseline scan (e.g. fMRI, sMRI)?
2. Treatment Response:
   A. Can we predict treatment response based on things that we have right now?
      i) Even better if they are cheap/available! (patient report, blood assay, biomedical test/examination, behavioural task, also genetics, brain scans)
3. Interpretation:
   A. How to interpret classifiers/regression weights?
   B. Can we find the most relevant brain regions for diagnosis/prognosis?
   C. How likely is it that this model is actually real?

4. (Diagnosis:
   A. Can we classify groups of subjects (e.g. patients vs. controls) using structural sMRI/fMRI scans?
   B. Can we combine information from different imaging modalities and/or clinical information?
   C. Are patients outliers with respect to a "normal population"?)

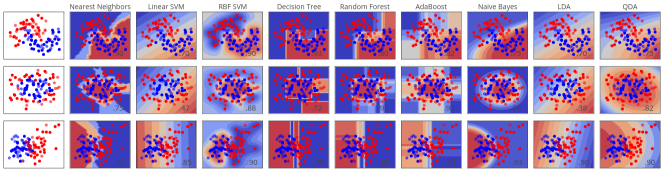
Interim summary



- Machine learning is a subfield of computer science concerned with getting computers to do useful stuff for us
- Helpful to consider supervised, unsupervised, and reinforcement learning as the three divisions
  - ML is not just classification! - Computational statistics
- Popular analysis tool in psychiatry

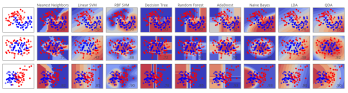


What is all the fuss about?



- Mechanistic reasons
- Pragmatic reasons

What is all the fuss about?



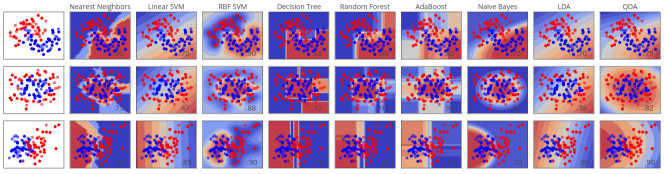
Mechanistic reasons

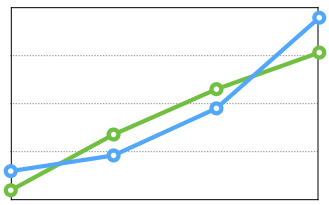
- Develop systems for learning *complicated* relationships in data
  - Driven by: larger data sets, more computational resources
- Allows us to discover new knowledge

Pragmatic reasons

- Make better predictive tools
  - Often, without domain-specific knowledge (c.f. kaggle!)
- Performance usually outperforms traditional statistical procedures
- ML can replace boring tasks
  - Frees up humans for human things (like treating patients)

What is all the fuss about?





Data volume

Computational resources

Summary

- Machine learning is a subfield of computer science concerned with getting computers to do useful stuff for us
  - ML is not just classification! - Computational statistics
- Popular analysis tool in psychiatry
  - Find complex relationships—more data, more resources
  - Build better tools to improve diagnosis and treatment