

Machine Learning I: basics

Andre F. Marquand

a.marquand@donders.ru.nl



- Pattern Recognition (PR) is a subfield of machine learning, which relates to the automatic discovery of patterns of statistical regularity in data
- Aim to learn from empirical data rather than following fixed rules
- **Learn by example**
- Increasingly used in computational psychiatry for:
 - ① Predicting clinical variables (diagnosis or treatment response)
 - ② Stratifying psychiatric disorders
 - ③ Learning mappings between behaviour and brain systems

Outline



- 1 Introduction to Machine Learning
- 2 Basics of Pattern Recognition Analyses
- 3 Applications in Psychiatry
- 4 Conclusions

Outline



1 Introduction to Machine Learning

2 Basics of Pattern Recognition Analyses

3 Applications in Psychiatry

4 Conclusions

What is pattern recognition used for?



Historically, has been applied in many application domains:

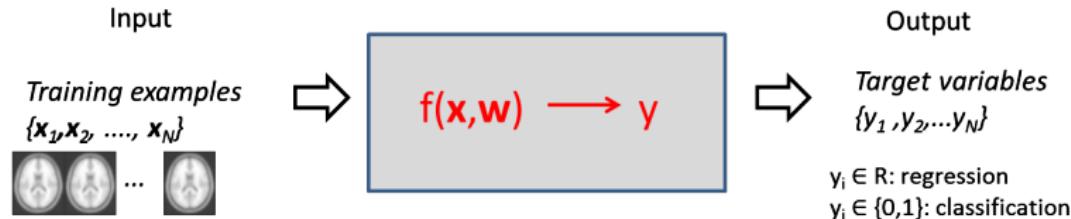
Example Applications

- Speech Recognition
- Automatic Character recognition / handwriting recognition
- Document classification (e.g. spam filters)
- Analysis of genetic microarray data
- Self-driving cars
- Recommender systems / online shopping
- ...

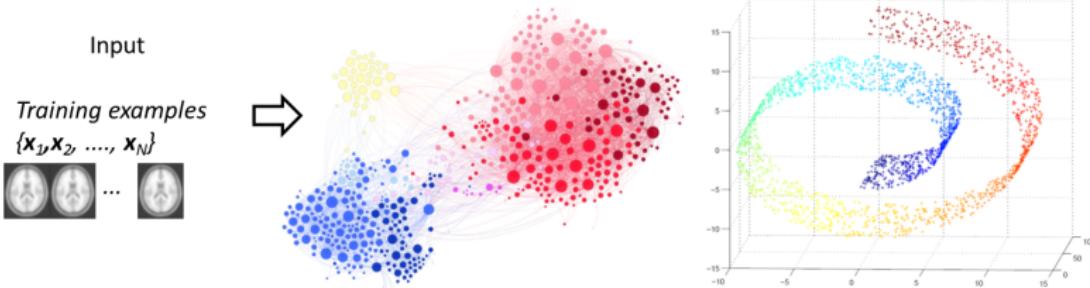
Types of pattern recognition



Supervised learning involves learning a mapping between input and output:

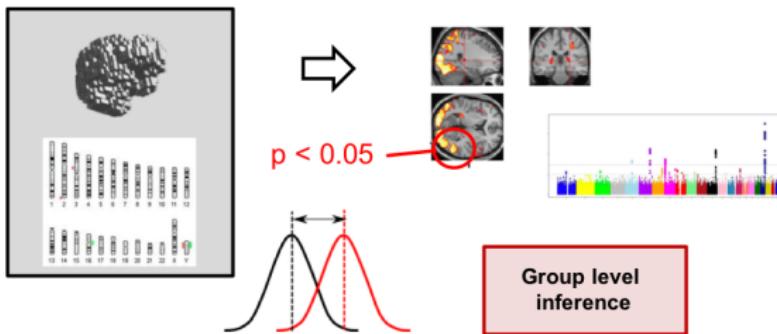


In **Unsupervised** learning, algorithms are not provided with output labels and must learn to structure the data by applying heuristics





Mass univariate association testing (SPM, GWAS)

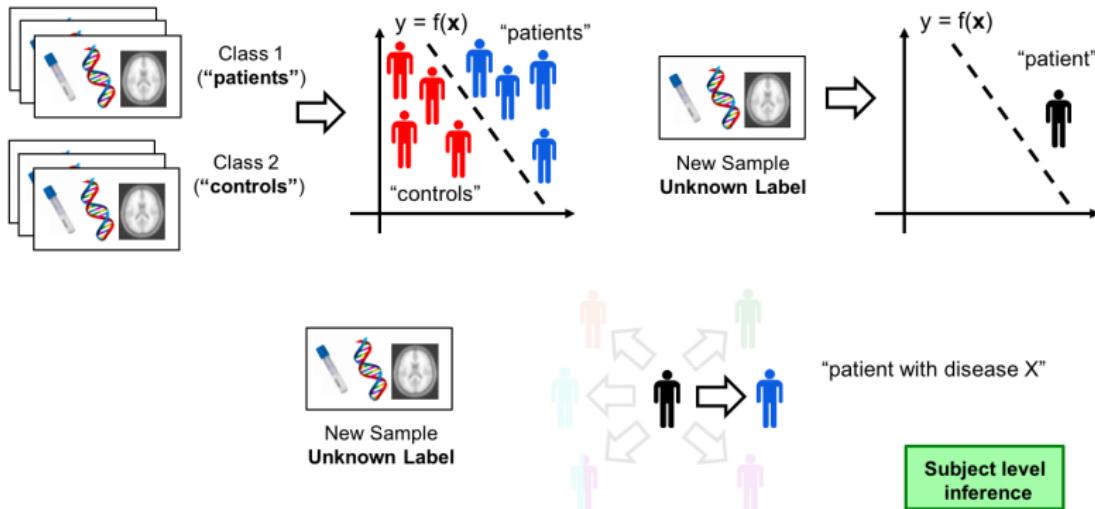


- Useful for understanding mechanisms
- For clinical decision making this does not suffice. It is necessary to make predictions about individuals

Predicting disease state



Making subject level predictions of diagnosis and outcome

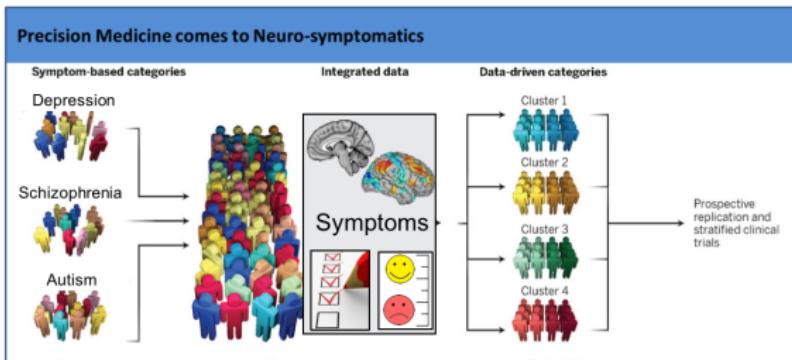


Useful to find a measure of overall group separation

Machine Learning in Psychiatry: Stratification



Tackling the clinical and biological heterogeneity of psychiatric disorders



Insel et al. (2015)

Outline



1 Introduction to Machine Learning

2 Basics of Pattern Recognition Analyses

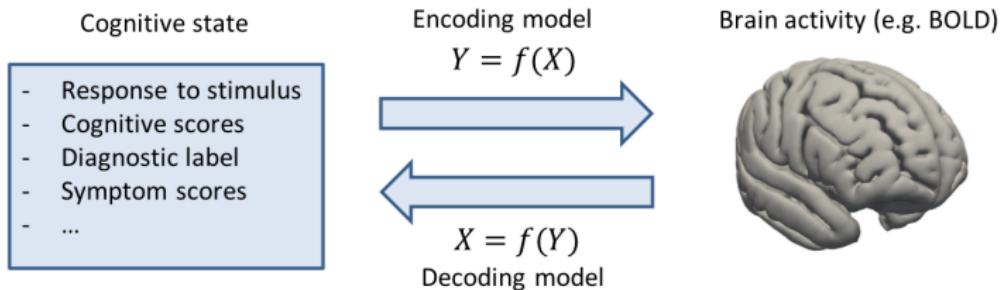
3 Applications in Psychiatry

4 Conclusions

Encoding and Decoding



Which brain regions are activated
in a certain task condition?



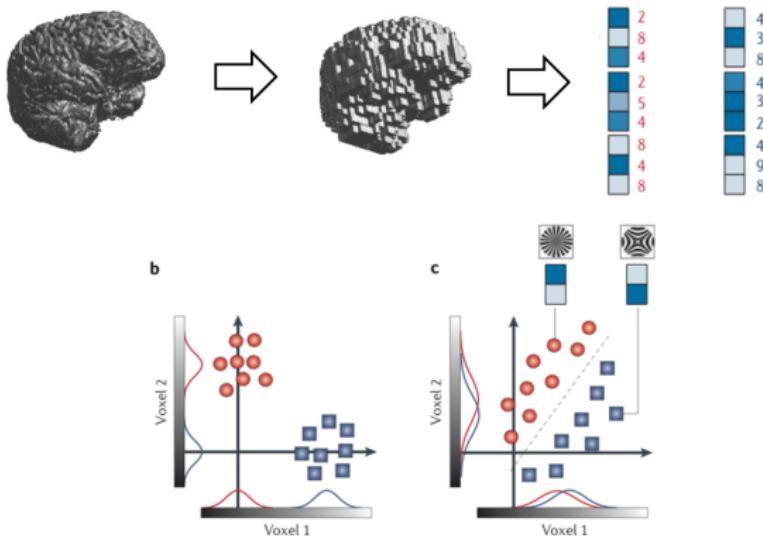
To which clinical group
does this subject belong?

- Also called Generative/Recognition models
- \neq Generative/Discriminative models in machine learning
- This distinction relates to the brain, not to the methods

Multivariate models



Sensitivity for spatially distributed (or multivariate) effects:



Haynes and Rees (2006)

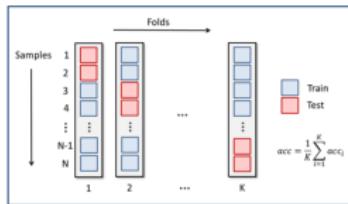
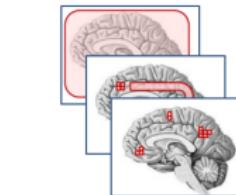
Stages of supervised pattern recognition analysis



1. Feature extraction and/or feature selection

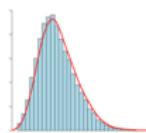


2. Classification / Regression using cross-validation



3. Performance evaluation

$$acc = \frac{1}{K} \sum_{i=1}^K acc_i$$



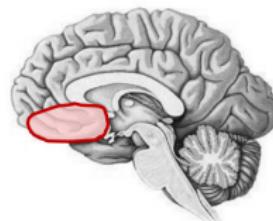
Feature selection and feature construction



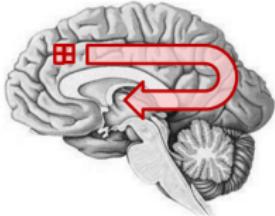
Whole Brain



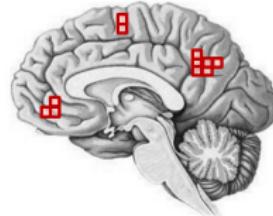
Region of Interest



Searchlight



Feature selection



- Can also construct features (e.g. using ICA/ PCA,...)
- Or learn features from the data (e.g. deep learning)
- Feature selection should be performed on training data only!



Notation

$$\mathcal{D} = \{\mathbf{X}, \mathbf{y}\} \text{ or } \{\mathbf{X}, \mathbf{Y}\}$$
 Dataset
$$\mathbf{X}_{N \times D} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$$
 N samples, D features
$$\mathbf{y} = [y_1, \dots, y_N]^T$$
 Targets
$$\mathbf{w} = [w_1, \dots, w_D]^T$$
 Weights



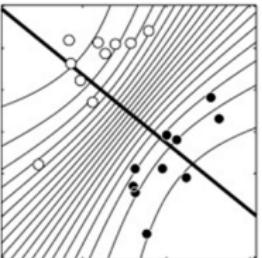
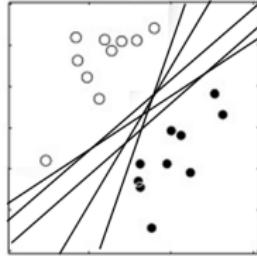
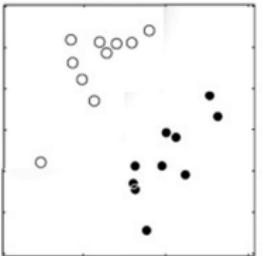
A smorgasbord of different approaches

- Regularisation methods (Penalized linear models, support vector machines, ...)
- Probabilistic approaches (Linear discriminant analysis, Gaussian processes, ...)
- Ensemble methods (Random forests, boosting, ...)
- Neural networks (multi-layer perceptrons, deep learning, ...)

Most methods aim to trade-off data fit with complexity

$$\begin{aligned} f(\mathbf{x}_i, \mathbf{w}) = f_i = \mathbf{x}_i^T \mathbf{w} \quad \Rightarrow \hat{\mathbf{w}} &= \min_{\mathbf{w}} \sum_{i=1}^n \ell(y_i, f_i) + \lambda J(\mathbf{w}) \\ \Rightarrow \hat{\mathbf{w}} &= \min_{\mathbf{w}} -\ln p(\mathbf{w}|\mathbf{y}) \\ &= \min_{\mathbf{w}} \sum_{i=1}^n \ln p(y_i|f_i) + \ln p(\mathbf{w}|\theta) \end{aligned}$$

Choice of pattern recognition algorithm



Integrate over all
Possible decision functions



Gaussian process
Classification (GPC)

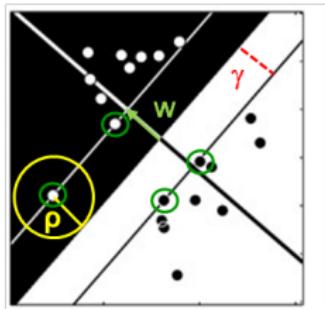
All methods make assumptions!

Ashburner and Klöppel (2011)

Support Vector Machines



- Finds a separating hyperplane that is “optimal” in that it leads to the largest margin between classes (γ)
- Based on the assumption that each point is bounded by unknown noise (ρ)
- New points will be well classified if $\gamma > \rho$
- The hyperplane is uniquely defined by a subset of the most ambiguous data points (“support vectors”)



$$\begin{aligned} \min_{w, \xi, b} \quad & -\gamma + C \sum_{i=1}^N \xi_i \\ \text{s.t.: } \quad & y_i(w^T \phi(x) + b) > \gamma - \xi_i \\ & \xi_i > 0 \\ & \|w\|^2 = 1 \end{aligned}$$

Deep Learning



- 'Deep' neural networks have seen an enormous surge in popularity over the last few years
- Extend 1950s-era neural networks to have many hidden layers
- Now provide state of the art performance in many domains, e.g. computer vision, game playing and perception

LETTER

doi:10.1038/nature14296

Human-level control through deep reinforcement learning

Volodymyr
Mnih¹
Helen King¹
Koray Kavukcuoglu²
David Silver¹
Julian Schrittwieser¹
John Nham¹
Thore Graepel¹

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver¹
Julian Schrittwieser¹
John Nham¹
Thore Graepel¹

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE

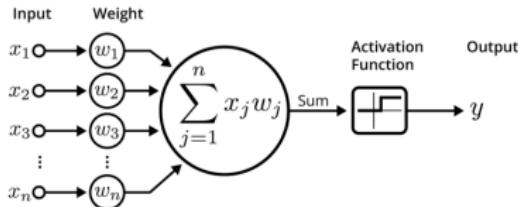
OPEN

Scalable and accurate deep learning with electronic health records

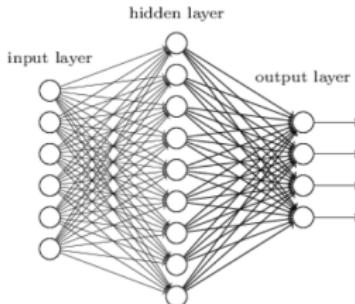
Alvin Rajkomar^{1,2}, Eyal Oren¹, Kai Chen¹, Andrew M. Dai¹, Nissan Hajaj¹, Michaela Hardt¹, Peter J. Liu¹, Xiaobing Liu¹, Jake Marcus¹, Mimi Sun¹, Patrik Sandberg¹, Hector Yee¹, Kun Zhang¹, Yi Zhang¹, Gerardo Flores¹, Gavin E. Duggan¹, Jamie Irvine¹, Quoc Le¹, Kurt Lisch¹, Alexander Mossin¹, Justin Tanuswan¹, De Wang¹, James Wexler¹, Jimbo Wilson¹, Dana Ludwig¹, Samuel L. Volchenboum¹, Katherine Chou¹, Michael Pearson¹, Srinivasan Madabushi¹, Nigam H. Shah¹, Atul J. Butte¹, Michael D. Howell¹, Claire Cui¹, Greg S. Corrado¹ and Jeffrey Dean¹



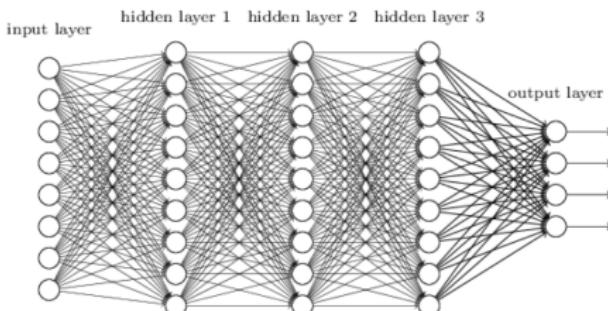
Deep Learning



"Non-deep" feedforward neural network



Deep neural network



- Many variants but “convolutional” networks are popular
- Predominantly supervised learning
- Usually many parameters to optimise (more in lecture 2)

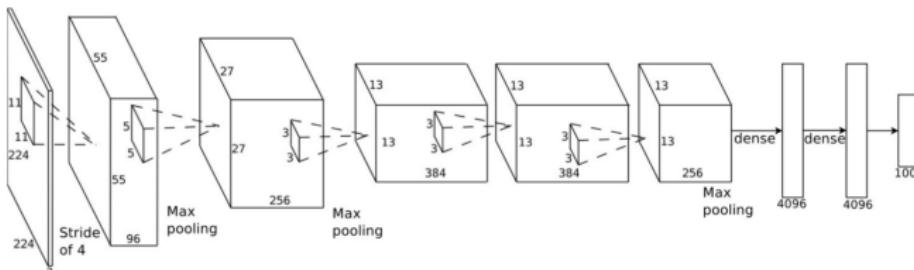


ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

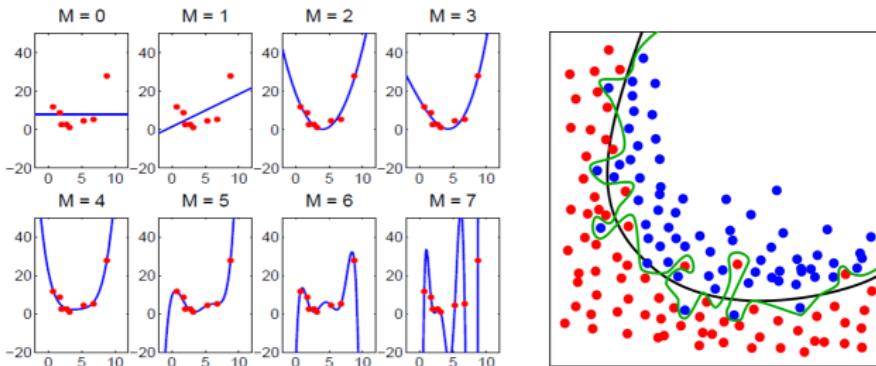


- 7 layer network that won 2012 ImageNet large-scale visual recognition challenge by 10%
- Trained the network on 15 million annotated images from over 22,000 categories
- more than 93,000 citations since 2012!

Overfitting



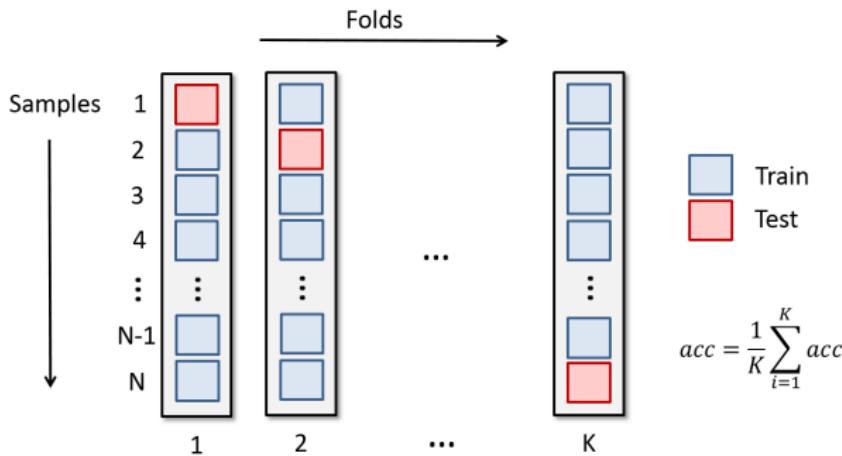
- Occurs when a model performs well on the data that it was estimated or trained on, but poorly on new data
- Can arise in very many ways including improper parameter optimisation or feature selection



Cross-validation



- Testing on unseen data is essential to assess generalizability
- Cross-validation one popular way to do this
- 'K-fold CV': split the data into K approximately equal chunks
- 'Leave-one-out': one sample is left out at a time ($K = N$)



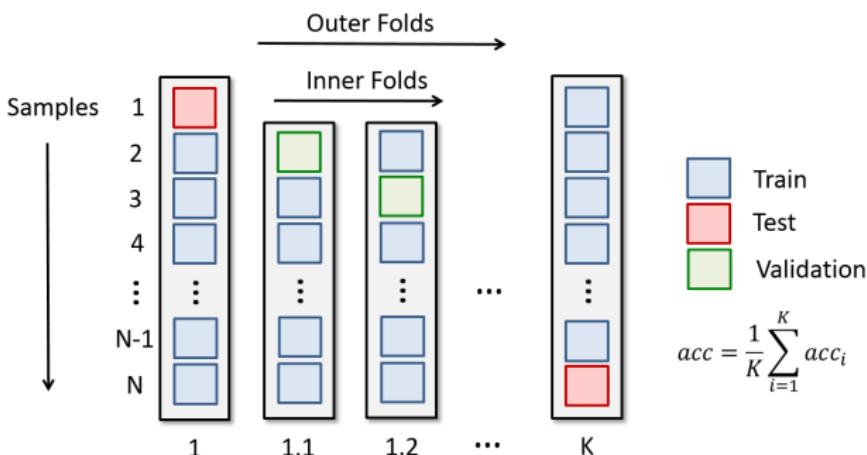
Parameter optimisation



- Most approaches depend on multiple (hyper)parameters
- e.g. regularization parameters in penalized linear models

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} \sum_{i=1}^n \ell(y_i, f_i) + \lambda J(\mathbf{w})$$

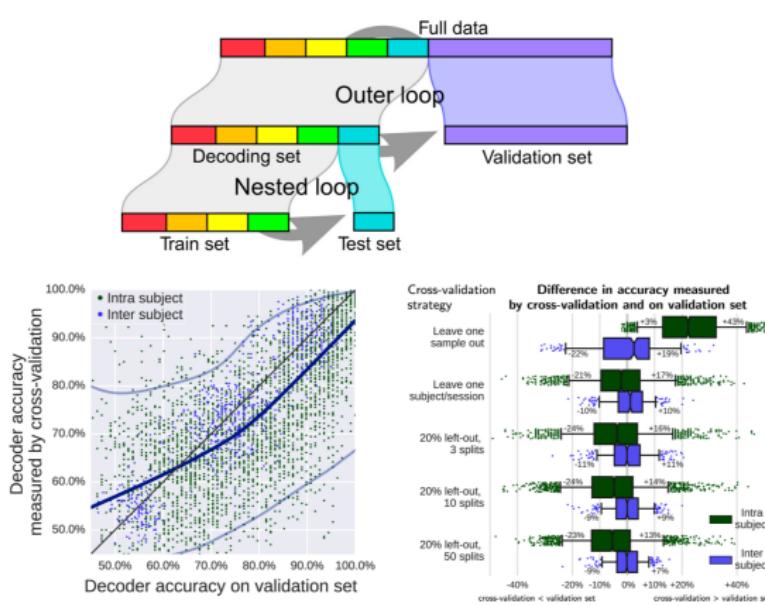
- Standard approach is *nested* cross-validation with a grid search



Multi-stage validation



- Despite being theoretically unbiased, CV can still overfit
- A multistage validation approach protects against this
- CV also invalidates parametric tests (more later)



Varoquaux et al. (2017)

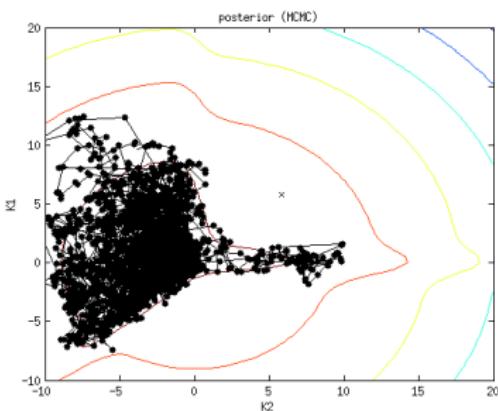
Bayesian parameter optimisation



- Bayesian models also depend on multiple variance/noise hyperparameters

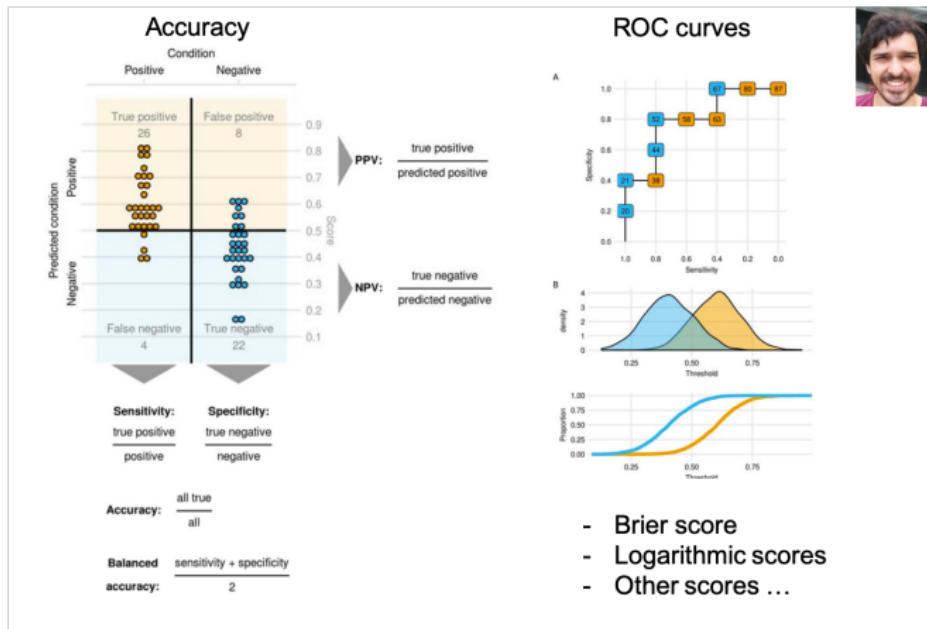
$$p(\mathbf{w}|\mathbf{y}, \theta, \sigma) = \frac{p(\mathbf{y}|\mathbf{w}, \sigma)p(\mathbf{w}|\theta)}{p(\mathbf{y}|\theta, \sigma)}, \quad p(\mathbf{y}|\theta, \sigma) = \int p(\mathbf{y}|\mathbf{w}, \sigma)p(\mathbf{w}|\theta)d\mathbf{w}$$

- Many approaches: nested CV, Empirical Bayes, MCMC ...





① Choice of error measure

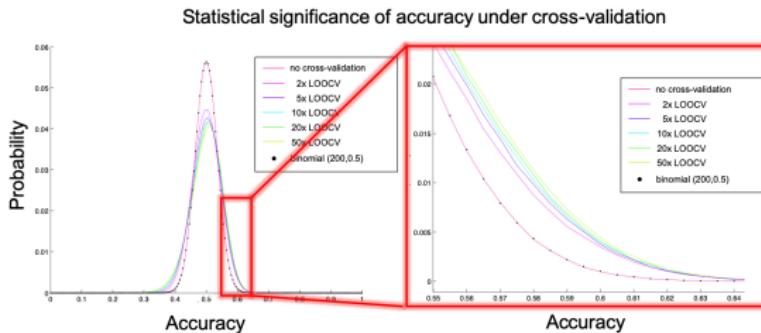


- Also many regression measures (explained variance, MSE,...)
- Different error metrics are sensitive to different aspects (e.g. MSE depends on the scale of the data) *Dinga et al. (2019)*

Choice of error metric



- ② Statistical testing framework. There are various options:
 - Parametric tests (e.g. binomial test, t-test)
 - Randomization tests (permutation, bootstrapping)
- Cross-validation induces dependency between the folds invalidating parametric statistics



- Parametric assumptions may not be met (e.g. interval data)
- Permutation tests must respect *exchangeability*, e.g. site effects, family structure ...

Stelzer et al. (2013); Winkler et al. (2015)

Outline



- 1 Introduction to Machine Learning
- 2 Basics of Pattern Recognition Analyses
- 3 Applications in Psychiatry
- 4 Conclusions

Supervised learning for automated diagnosis and prognosis



Neuroscience and Biobehavioral Reviews 57 (2015) 328–349

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

ELSEVIER

Review

From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics

Thomas Wolfers^{a,b,*}, Jan K. Buitelaar^{c,d}, Christian F. Beckmann^{b,c,e}, Barbara Franke^{a,f}, Andre F. Marquand^{b,g}

NeuroImage 145 (2017) 137–165

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimsg

ELSEVIER

Single subject prediction of brain disorders in neuroimaging:
Promises and pitfalls

Mohammad R. Arbabshirani^{a,b,*}, Sergev Plis^a, Jing Sui^{a,c}, Vince D. Calhoun^{a,d}

nature neuroscience

Building better biomarkers: brain models
in translational neuroimaging

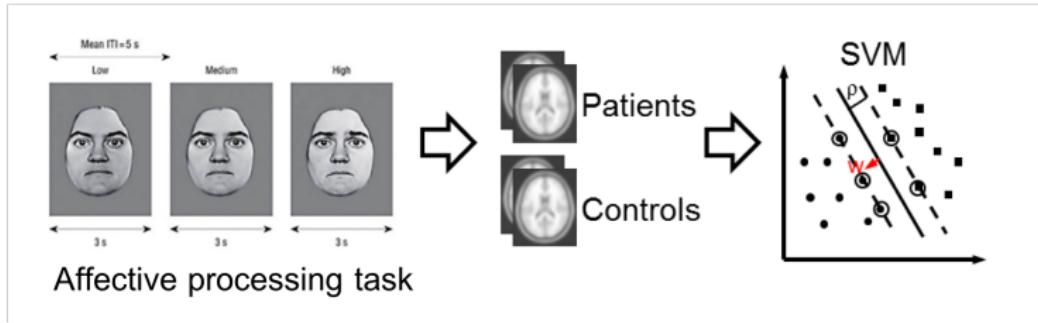
Choong-Wan Woo^{1–4}, Luke J Chang⁵, Martin A Lindquist⁶ & Tor D Wager^{1,4}

Wolfers et al. (2015); Arbabshirani et al. (2017); Woo et al. (2017)

Supervised learning for depression diagnosis

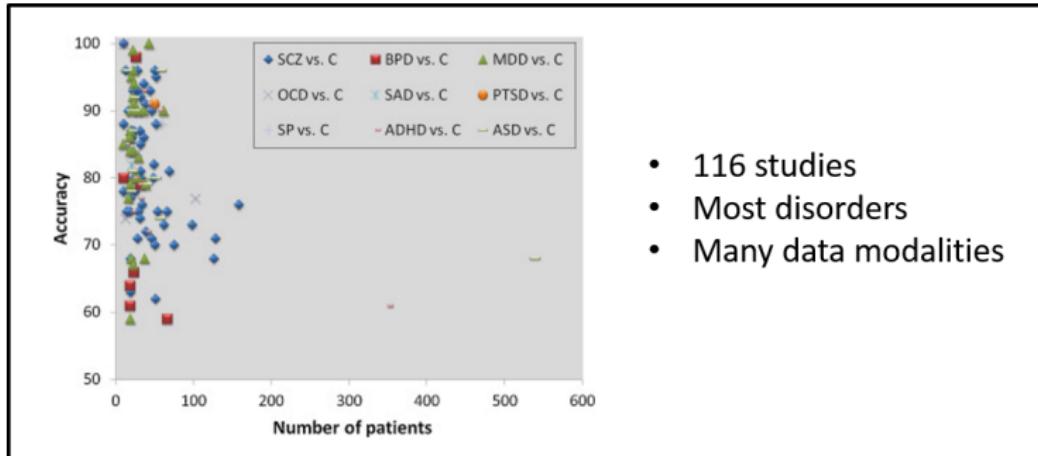


This study was an early application of Pattern recognition to predict disease state in major depression



- Patients could be discriminated from controls with 87% accuracy
- Patients who responded well to fluoxetine could be discriminated from non-responders with 67% accuracy

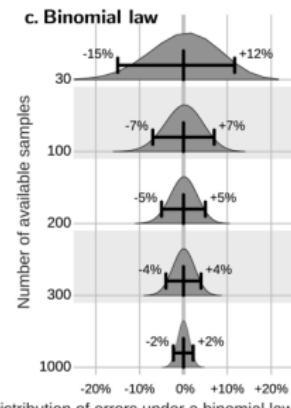
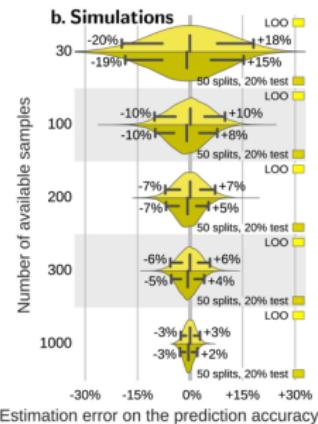
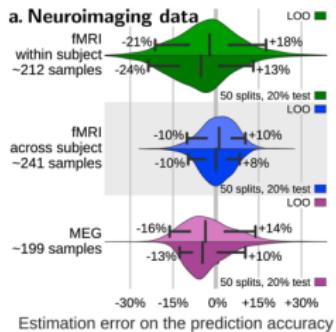
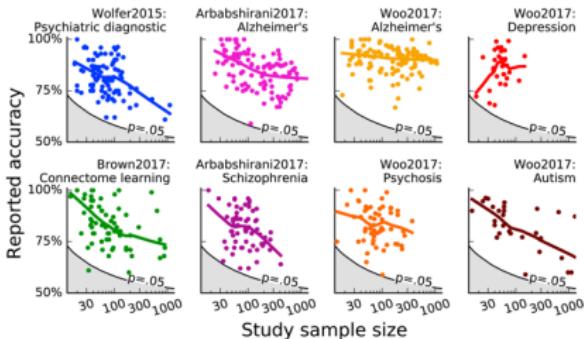
Where are we now?



- 116 studies
- Most disorders
- Many data modalities

- Moderate accuracy, highly variable across studies
- Mostly small samples, minimal validation across cohorts
- Accuracy in small samples is extremely variable
- **Heterogeneity** is a major challenge in clinical cohorts

Cross-validation with small samples



Varoquaux (2017)

Subtyping psychiatric disorders



PNAS

Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD

Damien A. Fair^{a,b,c,1}, Deep...

Departments of ^aBehavioral Neurology and ^bDepartment of Computer Science, University of Iowa, Iowa City, IA 52239

The American Journal of Psychiatry

Current Issue | Archive | About | Residents' Journal | AJP in Advance | Podcast | CME | Author Resources

Back to table of contents

Articles

Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers

Brett A. Clementz, Ph.D., Godfrey D. Pearson, M.D., M...

Published Online: 7 Dec 2012

ARTICLES

nature medicine

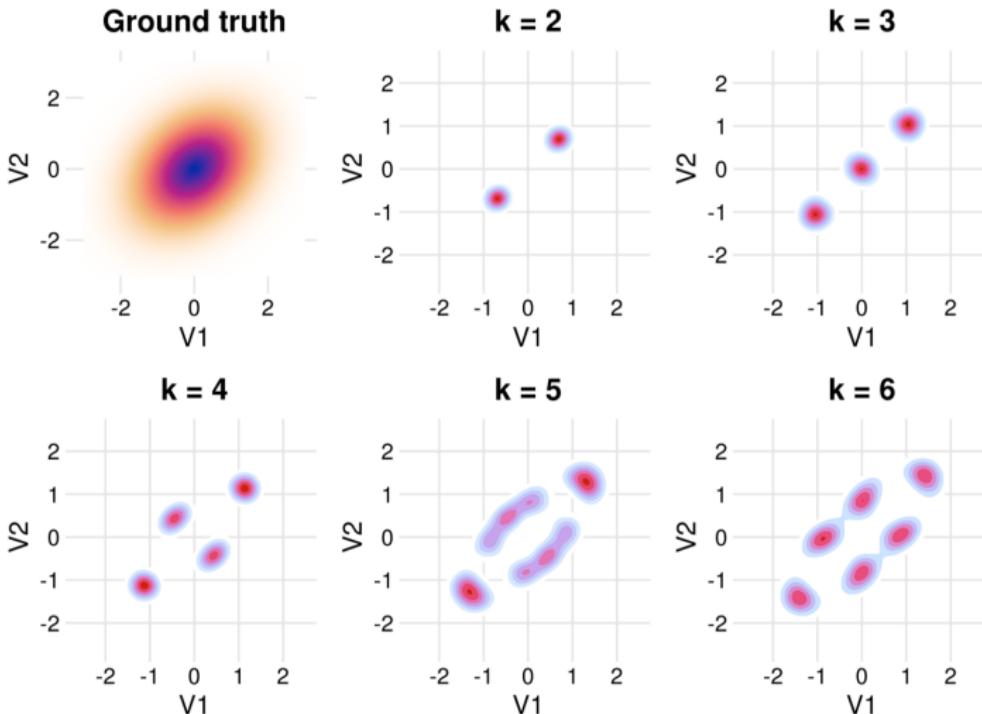
Resting-state connectivity biomarkers define neurophysiological subtypes of depression

Andrew T Drysdale^{1,3}, Logan Grusnick^{4,5}, Jonathan Dowmar², Katharine Dunlop⁶, Farrokh Mansouri⁶, Yue Meng¹, Robert N Fethko¹, Benjamin Zebely⁷, Desmond J Oathes⁸, Amit Etkin^{1,10}, Alan F Schatzberg⁹, Keith Sudheimer⁹, Jennifer Keller⁷, Helen S Mayberg¹¹, Faith M Gunning^{2,12}, George S Alexopoulos^{2,12}, Michael D Fox¹³, Alvaro Pascual-Leone¹³, Henning U Voss¹⁴, BJ Casey¹⁵, Marc J Duhin^{1,2} & Conor Liston^{1,3}

Validation of clusters is difficult:

- Clustering always gives a result and there is no clear measure of success (e.g. stability? separability? predictive ability?)
- Rarely test against the 'null' hypothesis that there are no clusters in the data
- Clustering using symptoms may not map onto biology

Issues with clustering

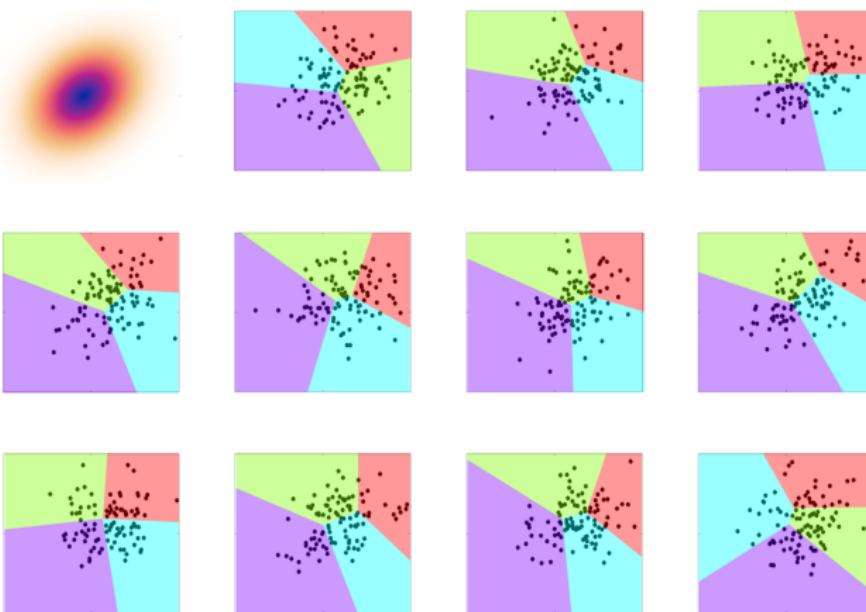


Graphic courtesy of Richard Dinga

Issues with clustering



Ground truth



Graphic courtesy of Richard Dinga

Outline



- 1 Introduction to Machine Learning
- 2 Basics of Pattern Recognition Analyses
- 3 Applications in Psychiatry
- 4 Conclusions



- PR is a powerful tool to perform single subject inference and detect spatially distributed effects
- Useful in clinical neuroscience for:
 - ① Making predictions at the subject level (e.g. prognosis)
 - ② Stratifying psychiatric disorders
 - ③ Estimating mappings between brain and behaviour
- Validation of models is extremely important to ensure generalisability
- More on that in Lecture 2 ...

References

- Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137 – 165, 2017. ISSN 1053-8119. doi: <https://doi.org/10.1016/j.neuroimage.2016.02.079>. Individual Subject Prediction.
- J. Ashburner and S. Klöppel. Multivariate models of inter-subject anatomical variability. *NeuroImage*, 56(2): 422–439, 2011.
- Richard Dinga, Brenda W.J.H. Penninx, Dick J. Veltman, Lianne Schmaal, and Andre F. Marquand. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv*, 2019. doi: 10.1101/743138. URL <https://www.biorxiv.org/content/early/2019/08/22/743138>.
- C. H. Fu, J. Mourao-Miranda, S. G. Costafreda, A. Khanna, A. F. Marquand, S. C. Williams, and M. J. Brammer. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry*, 63(7):656–62, 2008.
- J. D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, 7(7):523–34, 2006.
- Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400 – 410, 2011. ISSN 1053-8119. Multivariate Decoding and Brain Reading.
- Johannes Stelzer, Yi Chen, and Robert Turner. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage*, 65:69–82, 2013.
- Gael Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, (In press), 2017.
- Gael Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andres Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166 – 179, 2017. ISSN 1053-8119. doi: <https://doi.org/10.1016/j.neuroimage.2016.10.038>. Individual Subject Prediction.
- Anderson M. Winkler, Matthew A. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. Multi-level block permutation. *NeuroImage*, 123:253 – 268, 2015. ISSN 1053-8119. doi: <https://doi.org/10.1016/j.neuroimage.2015.05.092>.
- T. Wolfers, J. K. Buitelaar, C. F. Beckmann, B. Franke, and A. F. Marquand. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews*, in press, 2015.
- Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. Building better biomakers: brain models in translational neuroimaging. *Nature Neuroscience*, (20):365–377, 2017.

