Pattern Recognition Methods: basics

Fábio Ferreira

PRoNTo

Course 2018
May 14th- 15th
UCL, London

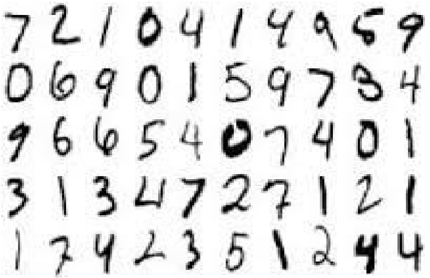
Outline

- **Pattern Recognition**
  - Concepts & Advantages
- **Supervised Learning**
  - Framework
  - Linear Predictive Function
  - Classification & Regression
- **Challenges in Neuroimaging**
- **Linear Models/Machines in PRoNTo**

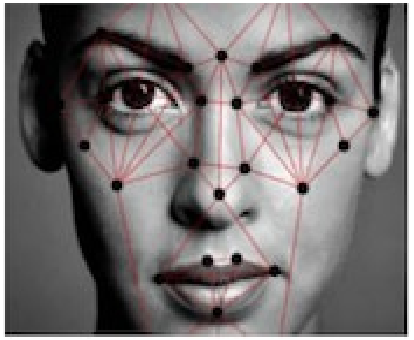
Pattern Recognition: Concepts

Pattern recognition aims to find patterns in the data which can be used to extract meaningful information to make predictions

Digit Recognition



Face Recognition



Finance

Advertising and Business Intelligence

Google Ads

Recommendation Engines

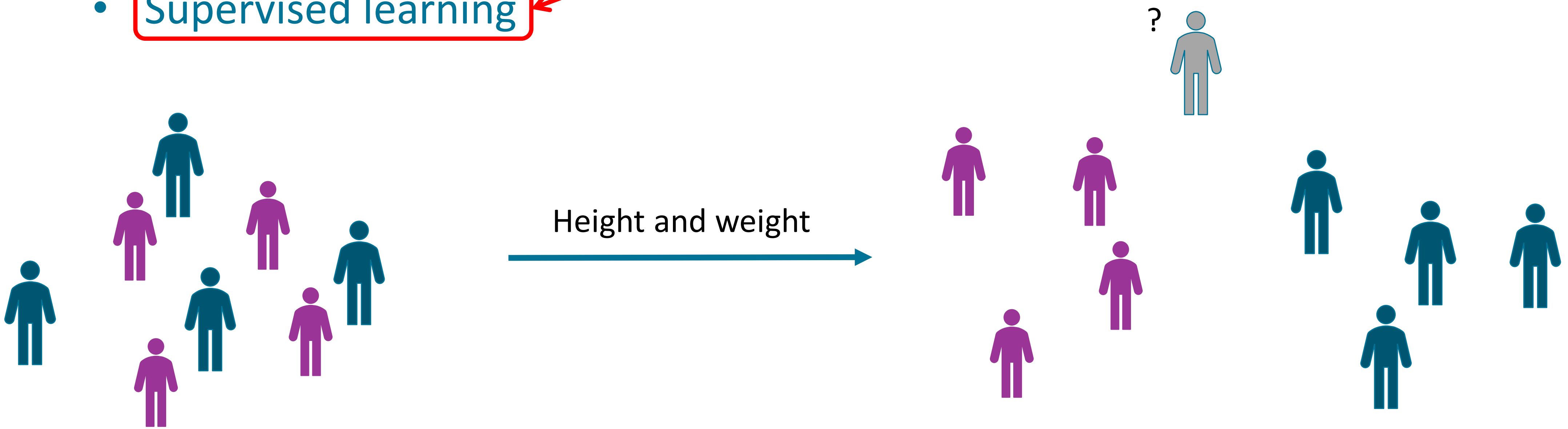
NETFLIX amazon

Pattern Recognition: Concepts

Types of learning procedures:

Currently implemented in PRoNTo

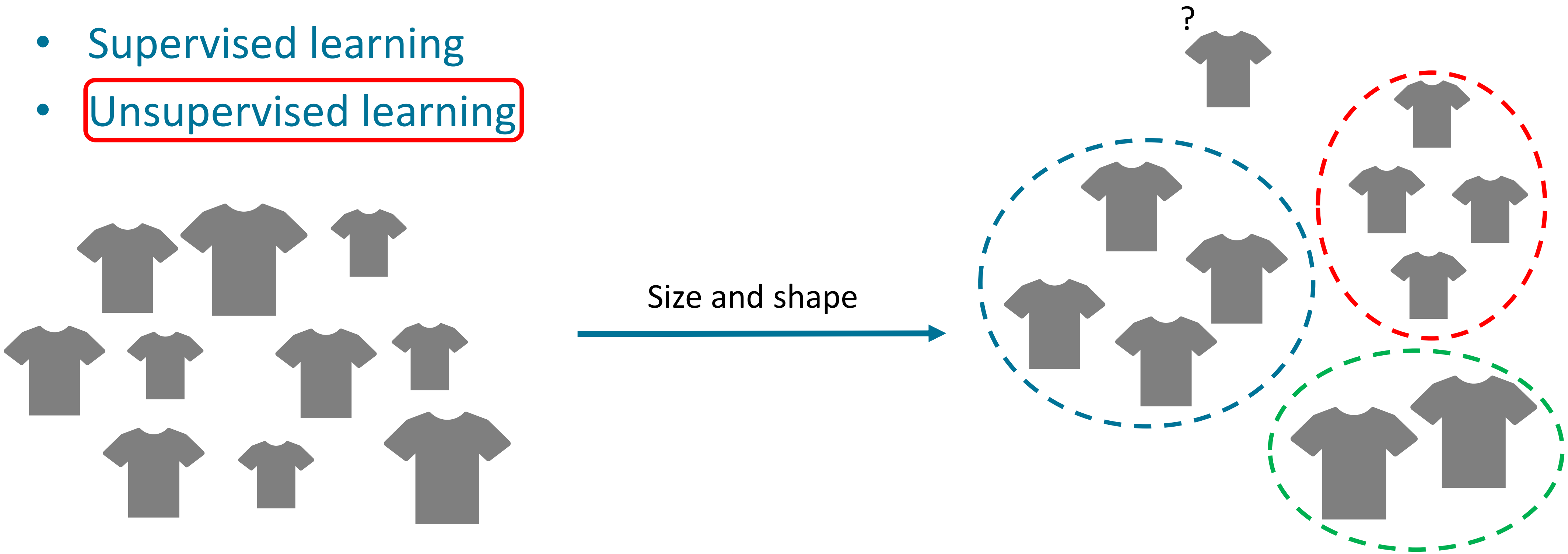
- Supervised learning



Pattern Recognition: Concepts

Types of learning procedures:

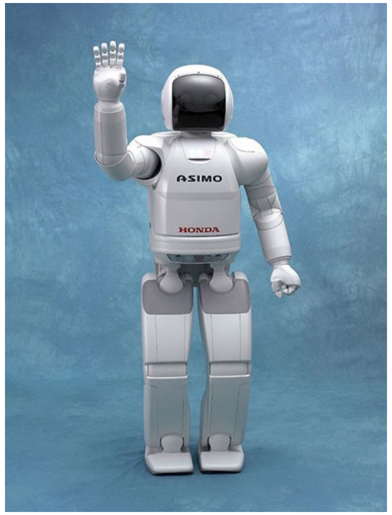
- Supervised learning
- Unsupervised learning



Pattern Recognition: Concepts

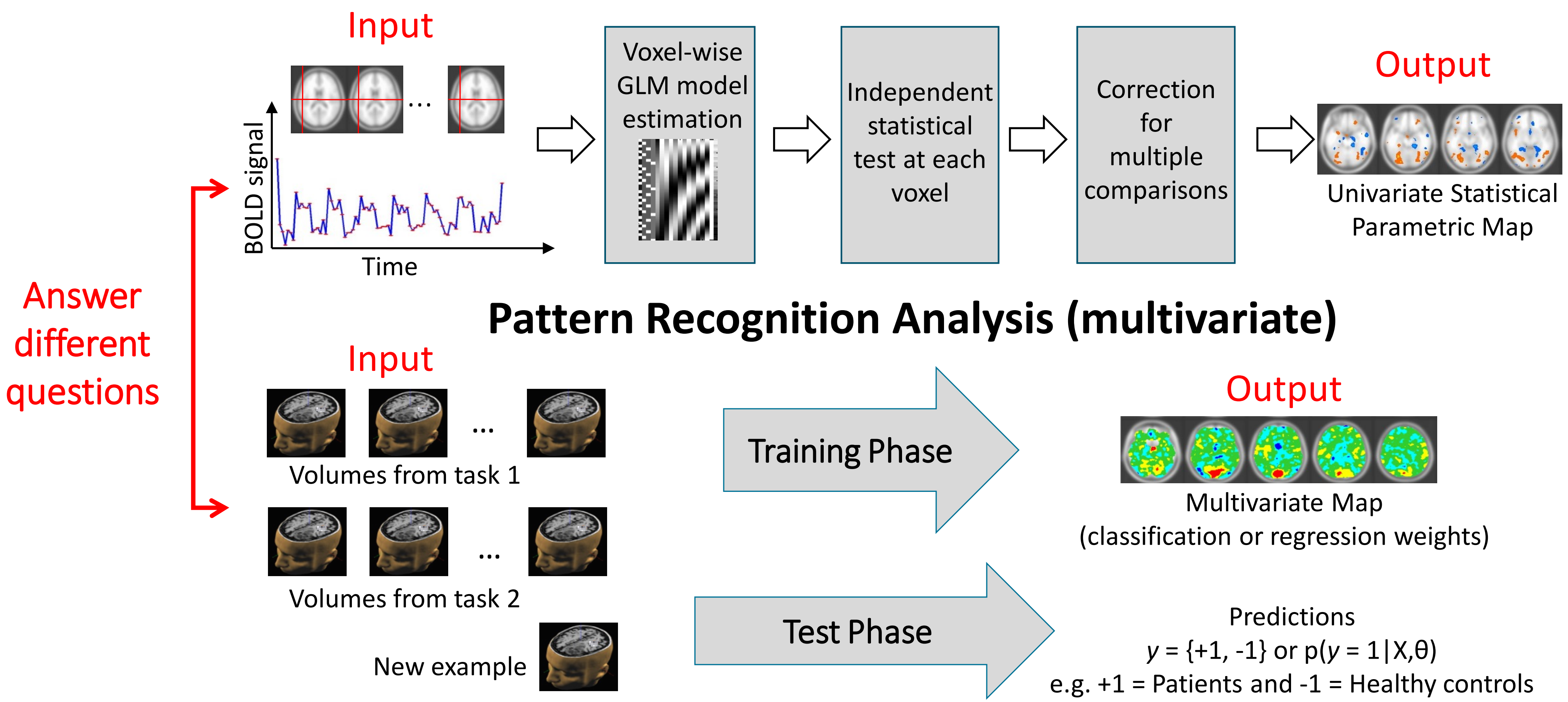
- Types of learning procedures:

  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning



Pattern Recognition Analysis

Standard Statistical Analysis (mass-univariate)



Pattern Recognition Analysis: advantages

- **Explore the multivariate nature of neuroimaging data**
- **Can be used to make predictions for new examples**

Supervised Learning: Framework

Input (Examples) → **No mathematical model available** → Output (labels or targets)

$\mathbf{x}_1 \longrightarrow y_1$

$\mathbf{x}_2 \quad y_2$

$\mathbf{x}_3 \quad y_3$

Machine Learning Methodology

Computer-based procedures that learn a function from a set of examples

Training Examples

$\{\mathbf{x}_1, y_1\}, \cdots, \{\mathbf{x}_n, y_n\} \longrightarrow$

Learning/Training Phase

Estimate a predictive function f such that

$$f : \mathbf{x}_i \rightarrow y_i$$

$\longrightarrow$ Predictive function

f

Test Example

$\mathbf{x}_* \longrightarrow$

Testing Phase

Prediction

$\longrightarrow$ Predicted value

$f(\mathbf{x}_*) = y_*$

Supervised Learning in Neuroimaging

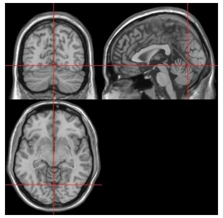
Machine learning models:
Enable predictions from brain imaging

Healthy
vs.
Disease

Cognitive state #1
vs.
Cognitive state #2

Clinical Score

Feature extraction

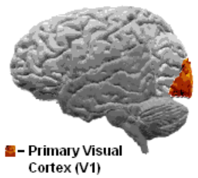


Whole brain volume

Feature Vector

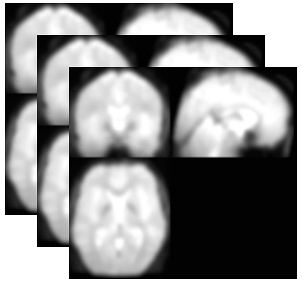
Dimensionality = number of voxels

Region of interest (ROI)



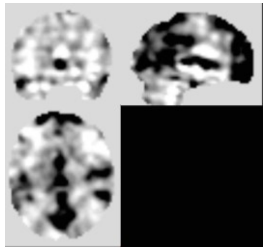
Feature Vector

Dimensionality = number of voxels within the ROI



fMRI

General Linear Model



Beta/Contrast images

Feature Vector

Dimensionality = number of voxels

Linear predictive function

- Linear predictive functions (classifier or regression) are parameterized by a weight vector **w** and a bias term *b*

- We learned **w** and *b* during the training phase by solving an optimisation problem

- The general equation for making predictions for a test example $\mathbf{x}_*$ is:

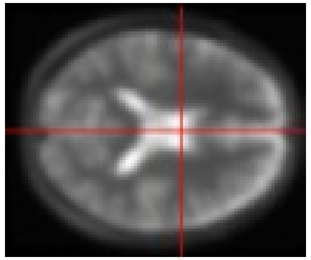
$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

Linear predictive function: prediction

Estimated (**w,** *b*) → | 5 | 2 | -6 | -1 |

× × × ×

New example (**x***)



→ | 1 | 2 | -2 | 4 |

Predictive function

$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

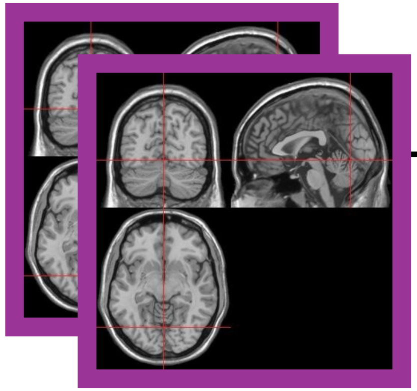
$$f(\mathbf{x}_*) = (5 \cdot 1) + (2 \cdot 2) + (-6 \cdot -2) + (-1 \cdot 4) + 0$$

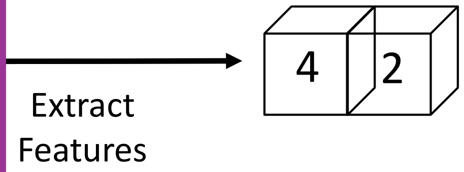
$$f(\mathbf{x}_*) = 5 + 4 + 12 - 4 = 17$$

$f(\mathbf{x}_*)$ is the predicted score for regression or the distance to the decision boundary for classification models.

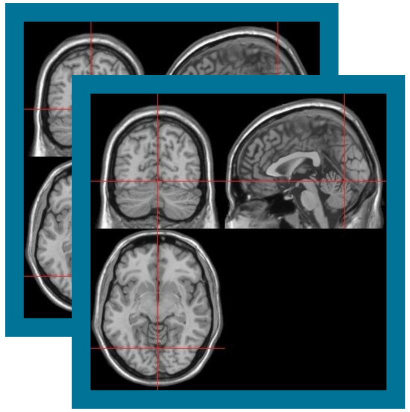
Classification problem

Class 1

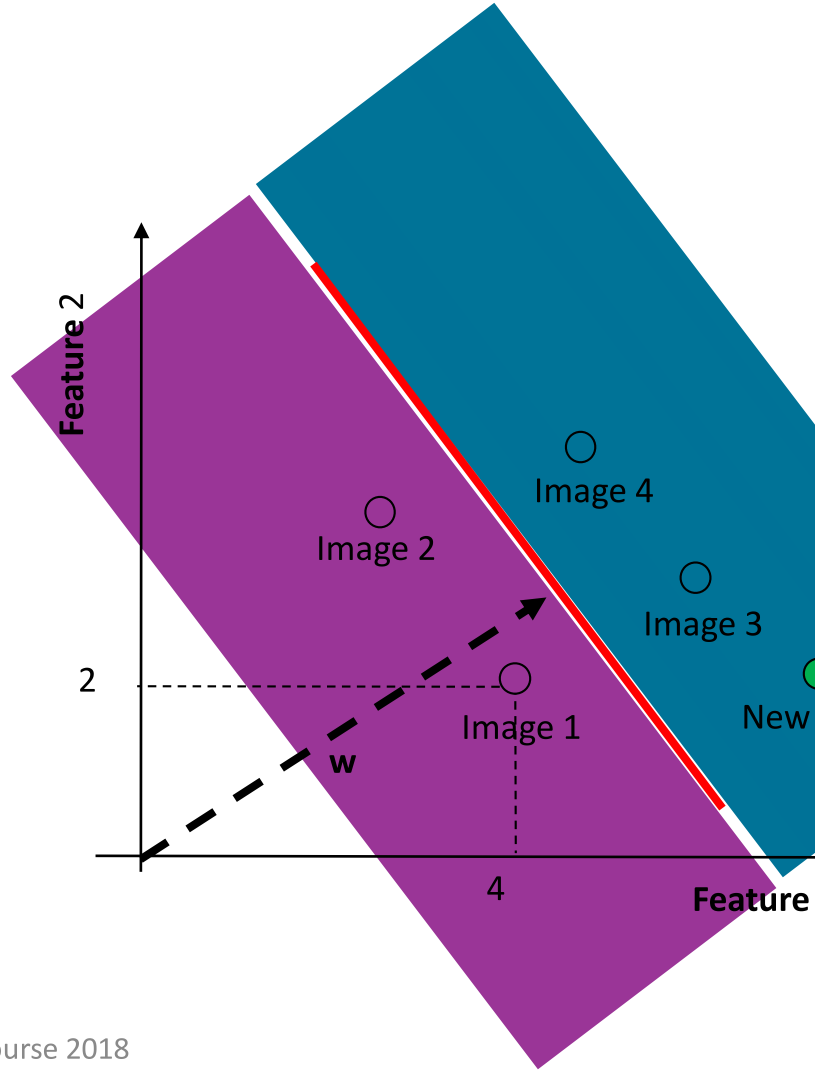




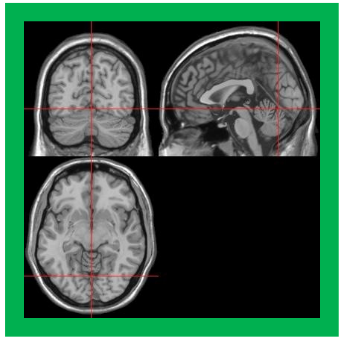
Class 2



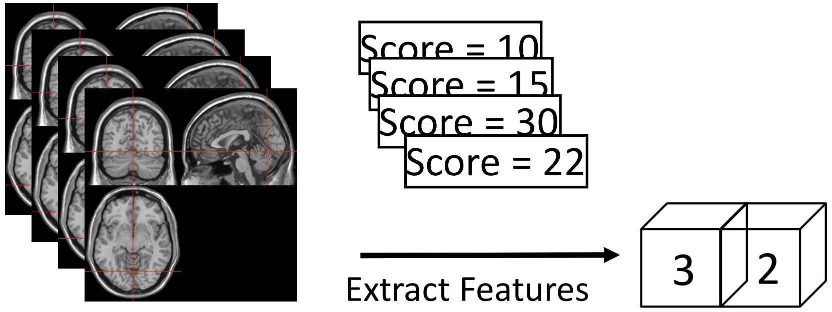
Linear classifiers are parameterized by a weight vector **w** and a bias term *b*.



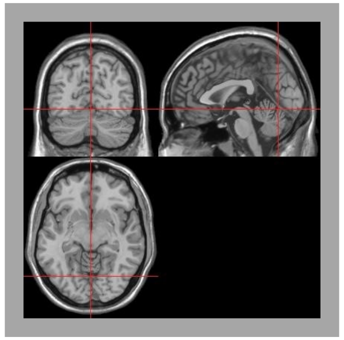
Test example



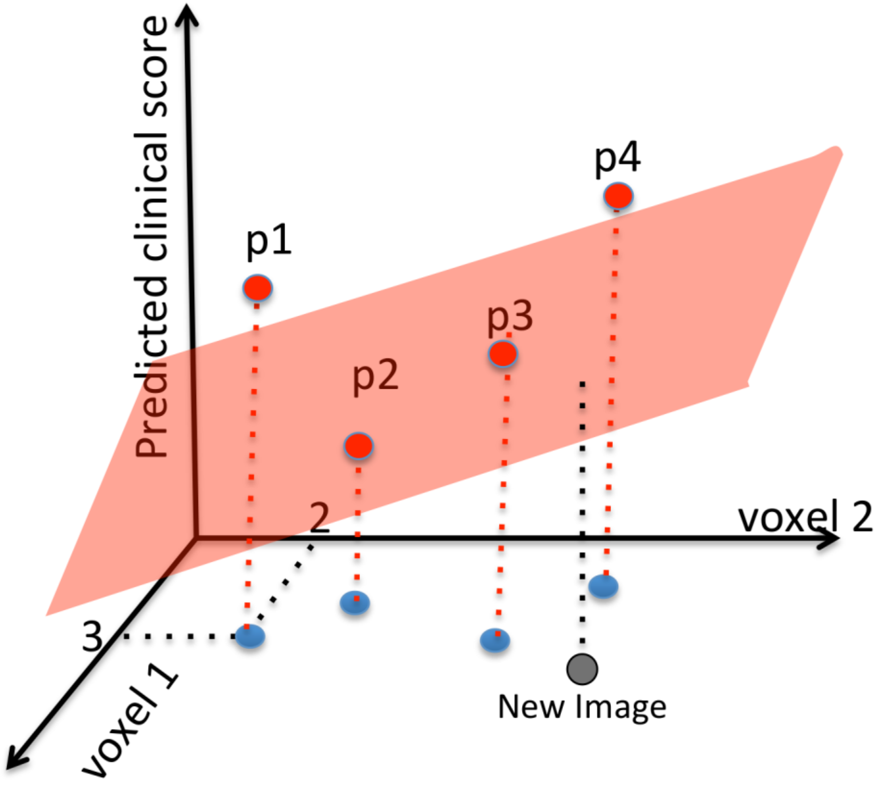
Regression problem



Test example



Linear regression models are also parameterized by a weight vector **w** and a bias term *b*.



Challenges in Neuroimaging

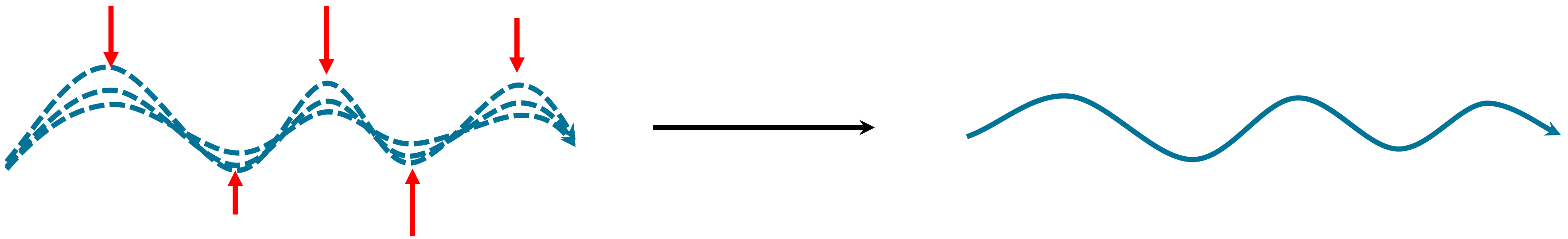
- In neuroimaging applications often the dimensionality of the data (e.g. number of voxels) is higher than the number of examples - **ill-conditioned problems**.

- Possible solutions:
  - –Regions of interest (ROIs)
  - –Feature selection strategies
  - –Searchlight
  - –**Regularisation + Kernel Methods**

Regularisation

- **Regularisation** is a technique used in an attempt to **solve ill-posed problems** and to **prevent overfitting** in statistical/machine learning models.

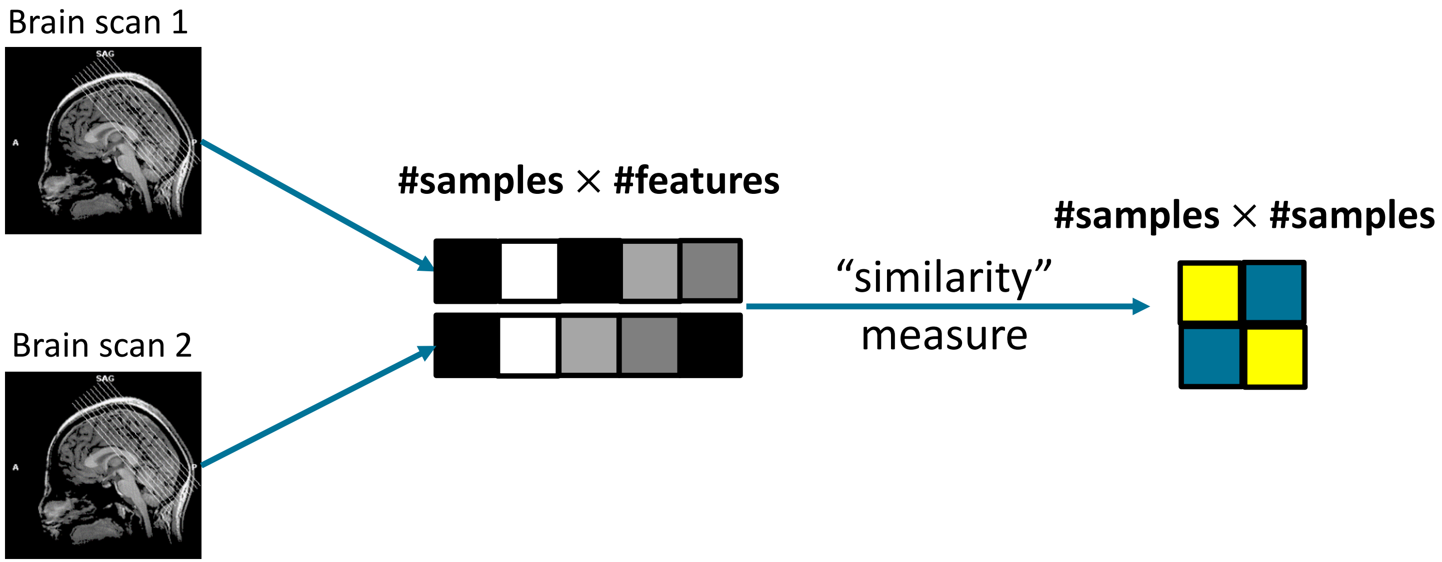
- Regularised methods find **w** by adding an additional constraint to the optimisation problem.



- Different machine learning algorithms solve different optimisation problems using different constraints (e.g. Kernel Ridge Regression (KRR), Support Vector Machine (SVM))

Challenges in Neuroimaging

How can we solve the high-dimensional problem efficiently?



Kernel methods

- **Kernel methods** provide a powerful and unified framework for investigating general types of relationships in the data (e.g. classification and regression)

- **Consists of two parts:**
  - Computation of the kernel matrix (compute all similarities)
  - Apply a learning algorithm based on the kernel matrix

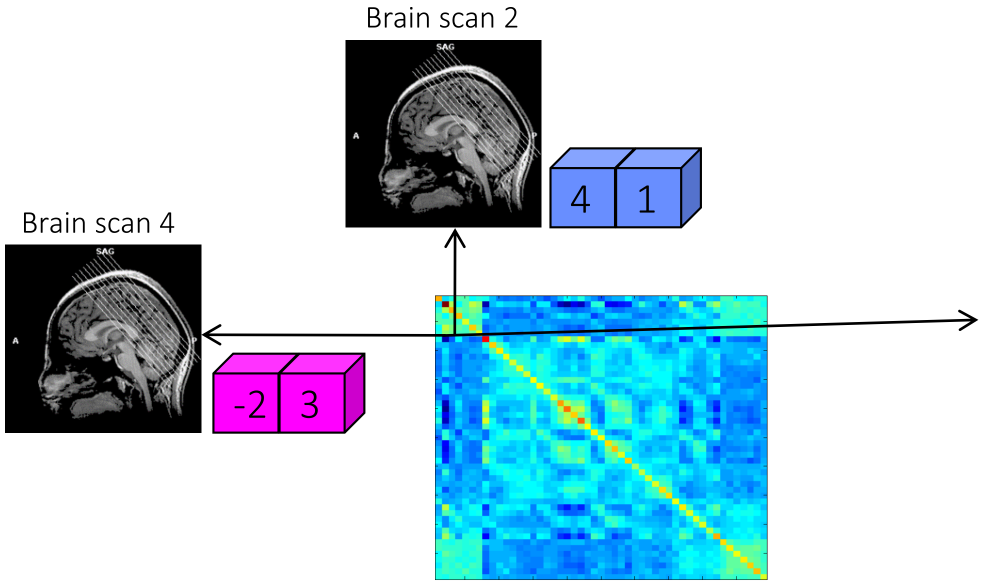
- **Main advantage:**
  - Computational efficiency

Linear Kernels

Kernel Function ("similarity" measure)

- Kernel is a function that, given **x** and **x*∗***, returns a real number characterizing their similarity
- A simple type of similarity measure between two vectors is a dot product (**linear kernel**)

$$\kappa(x, x_*) = \langle x \cdot x_* \rangle$$



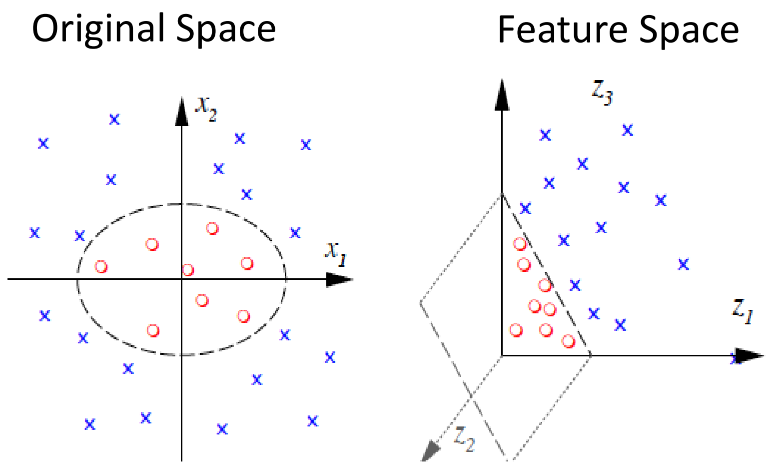
Dot product = (4 × -2)+(1 × 3) = -5

Non-linear Kernels

- There are more general “similarity measures”, i.e. non-linear kernels: Gaussian kernel, Polynomial kernel.

$$\boldsymbol{\phi} : \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x}) \in F$$

$$\kappa(\boldsymbol{x}, \boldsymbol{x}_*) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}_*) \rangle$$



$$\Phi : R^2 \to R^3$$

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt(2) x_1 x_2, x_2^2)$$

- Non-linear kernels are used to map the data to a higher dimensional space as an attempt to make it linearly separable.

Advantages of linear models

- Neuroimaging data usually are high-dimensional and the sample sizes are usually small, therefore non-linear kernels may not bring benefits.

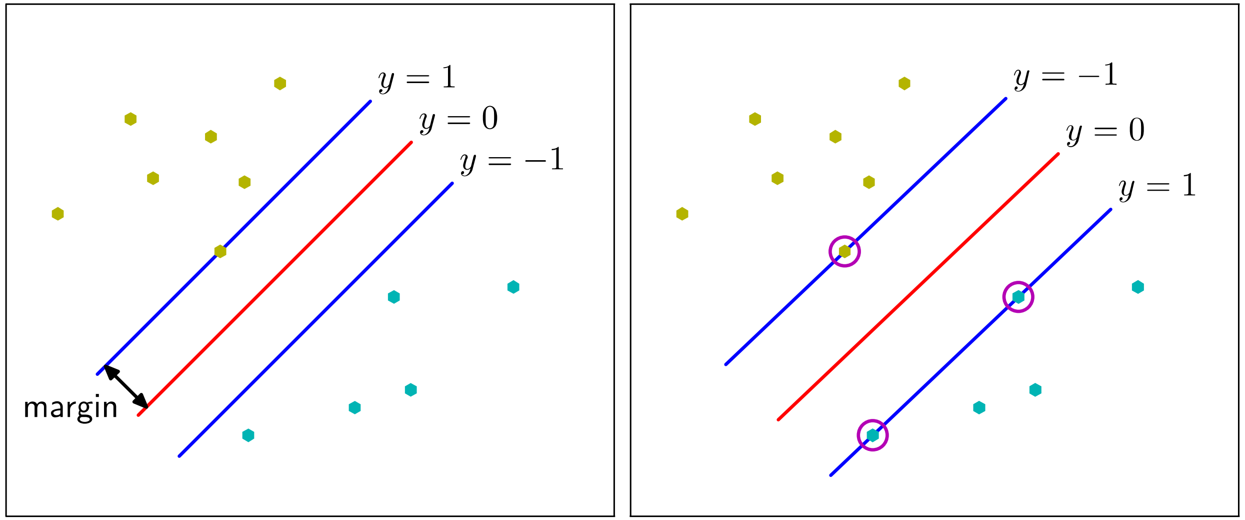
- Linear models reduce the risk of overfitting the data and allow direct extraction of the weight vector as an image (i.e. predictive map).

- Non-linear models usually have more hyperparameters that must be optimised which increases the computational times.

Linear Models/Machines in PRoNTo

- **Non-probablistic models:**
  - Support Vector Machine (SVM) (classification)
  - Kernel Ridge Regression (KRR) (regression)
  - Multiple Kernel Learning (MKL) (classification and regression)
- **Probablistic models:**
  - Relevance Vector Machine (RVM) (regression)
  - Binary (Multiclass) Gaussian Process (GP) (classification and regression)

Support Vector Machine (SVM)



- Gives good results for most problems
- Sparse solution in terms of examples (support vectors)
- Provides hard predictions

Linear Models/Machines in PRoNTo

Kernel Ridge Regression (KRR)

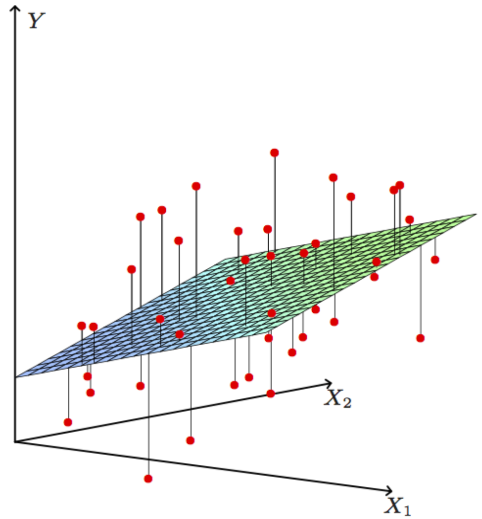


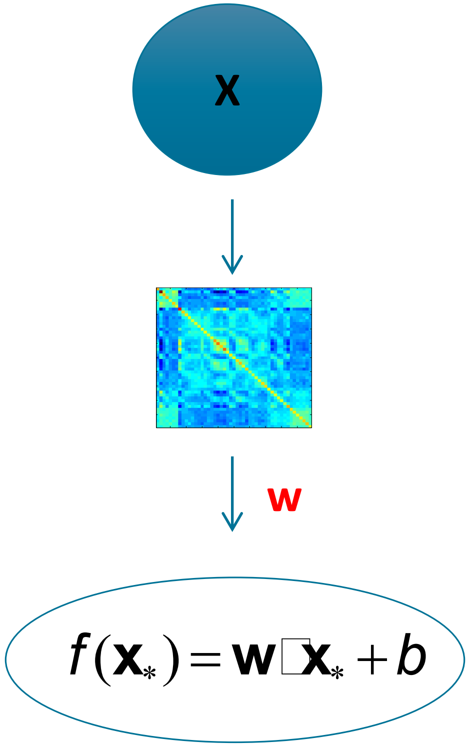
Illustration of a linear least squares fitting with X ∈ IR^2. We seek the linear function of X that minimizes the sum of squared residuals from Y.

Hastie, Tibshirani & Friedman, 2009

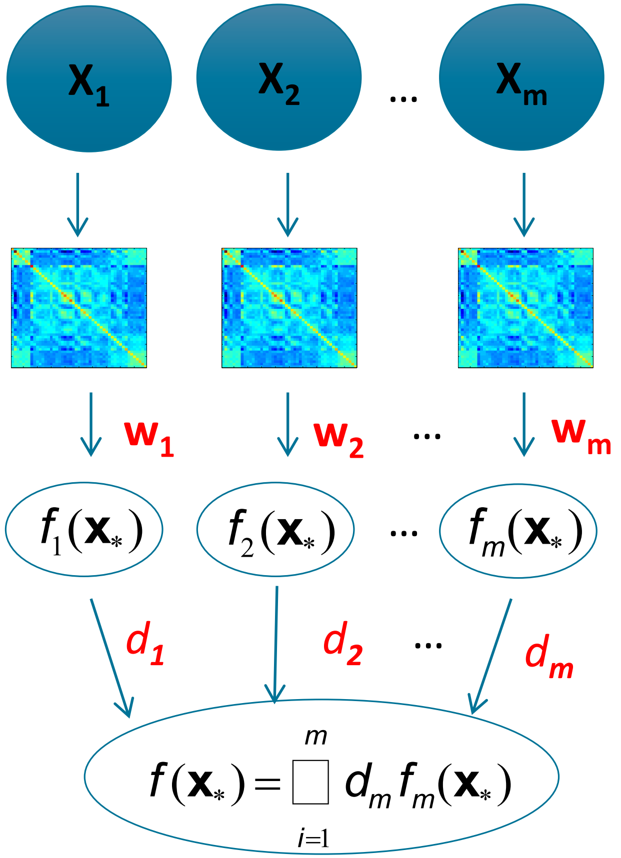
- Ridge regression consists in solving the optimization problem of a linear least squares regression by imposing a regularisation constraint.

Multiple Kernel Learning (MKL)

Single kernel SVM

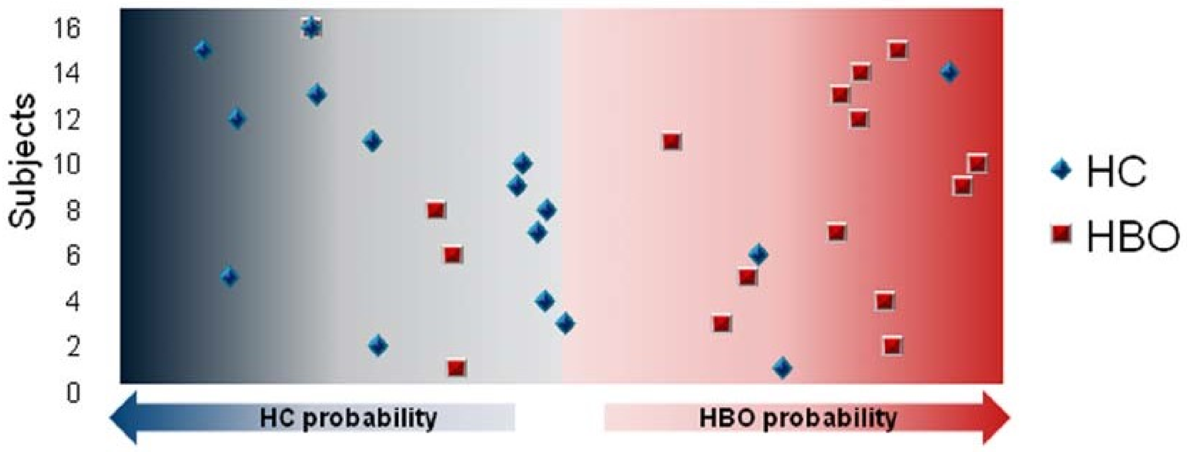


Multiple kernel SVM



Linear Models/Machines in PRoNTo

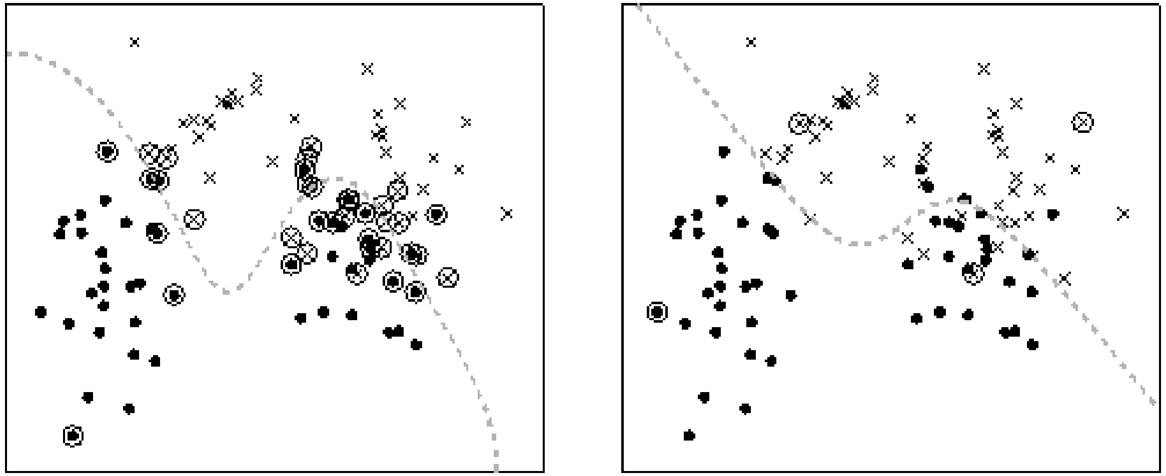
Gaussian Process Classifier – Binary/Multiclass



- Provides probabilistic class predictions (soft predictions)
- Natural extension to direct multi-class classification
- It does not find sparse solutions

Linear Models/Machines in PRoNTo

Relevance Vector Machine (RVM)



- Probabilistic: apply a Bayesian treatment to SVM
- It finds sparser solutions (relevance vectors) than SVM
- For large datasets, the training times can be longer than SVM

Take home message

What is my question?

Localisation vs prediction?

Classification vs regression?

Which machine/model?

References

PRoNTo papers:

- Schrouff J*, Rosa MJ*, Rondina J, Marquand A, Chu C, Ashburner J, Phillips C, Richiardi J, Mourao-Miranda J. PRoNTo: Pattern Recognition for Neuroimaging Toolbox, Neuroinformatics, February 2013. *co-first authors
- Schrouff J*, Monteiro, JM*, Portugal L, Rosa MJ, Phillips C, Mourao-Miranda J. Embedding Anatomical or Functional Knowledge in Whole-Brain Multiple Kernel Learning Models Neuroinformatics, 2018

Reviews:

- Pereira, Mitchell, Botnivik (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage,*45, S199-S209
- Haynes (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, 87(2), 257-270

Books:

- Hastie , Tibishirani, Friedman (2009). Elements of Statistical Learning. *Springer*
- Bishop, Jordan, Kleinberg, Schölkopf (2006). Pattern Recognition and Machine learning. *Springer*
- Shawe-Taylor and Cristianini (2004). Kernel Methods for Pattern Analysis. *Cambridge University Press.*
- Schölkopf and Smola (2001). Learning with Kernels. *The MIT Press.*

References

Machines/Models:

- Burges (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.
- Rasmussen, Williams (2006) Gaussian Processes for Machine Learning. *The MIT Press*
- Tipping (2001) Sparse Bayesian Learning and the Relevance Vector Machine *Journal of Machine Learning Research,* 1, 211-244
- Breiman (1996) Bagging Predictors Machine Learning, 24, 123-140
- Dietterich, Bakiri (1995) Solving multiclass learning problem via error-correcting output codes. Journal of Artificial Intelligence Research, 2: 263-286
- Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. Journal of Machine Learning Research, 9, 2491-2521
- Marquand (2010) Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage, 49(3), 2178-2189

Thank you!

Questions?