Pattern Recognition for Neuroimaging Toolbox

Pattern Recognition Methods: Basics

João M. Monteiro

Based on slides from Jessica Schrouff and Janaina Mourão-Miranda

PRoNTTo course
UCL, London, UK
2017
Outline

- Pattern Recognition
  - Concepts & Framework
- Examples of Linear Models/Machines
  - Classification
  - Regression
- Considerations for Neuroimaging
Pattern Recognition Concepts

Machine learning models:
Enable predictions from brain imaging

Healthy vs. Disease

Cognitive state #1 vs. Cognitive state #2

Clinical Score
Pattern Recognition Concepts

- Pattern recognition aims to assign a label to a given example (test example) based on statistical information extracted from the previous seen examples (training examples).

- The examples to be classified are usually groups of measurements or observations (e.g. brain image), defining points in an appropriate multidimensional vector space.

- Types of learning procedures:
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning, reinforcement learning.

Currently implemented in PRoNTo.
### Pattern Recognition Concepts

- **Samples**: $X$
- **Labels**: $Y$

$$f : X \rightarrow y$$

$$f : x_* \rightarrow y_*$$
Pattern Recognition Framework

Input (brain scans)

\[ x_1, x_2, x_3 \]

No mathematical model available

Output (control/patient)

\[ y_1, y_2, y_3 \]

Machine Learning Methodology

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

Training Examples:
\[ \{x_1, y_1\}, \ldots, \{x_n, y_n\} \]

Learning/Training Phase
Estimate a predictive function \( f \) such that
\[ f : x_i \rightarrow y_i \]

Prediction

Test Example \( x_* \)

Testing Phase
Prediction

\[ f(x_*) = y_* \]
Pattern recognition: classification model

Class 1

Label = patient
Label = patient
Label = patient
Label = patient
Label = patient

Class 2

Label = controls
Label = controls
Label = controls
Label = controls
Label = controls

New subject

Predictive function: $f$

Training

Testing

Prediction: Class membership
Pattern recognition: regression model

Training

Predictive function: $f$

New subject

Prediction:
Score = 28
Standard Statistical Analysis (mass-univariate)

Input

Voxel-wise GLM model estimation

Independent statistical test at each voxel

Correction for multiple comparisons

Output

Univariate Statistical Parametric Map

Very different meaning!

Pattern Recognition Analysis (multivariate)

Input

Volumes from task 1

Volumes from task 2

New example

Training Phase

Multivariate Map (classifier’s or regression’s weights)

Test Phase

Predictions

\( y = \{+1, -1\} \) or \( p(y = 1|X, \theta) \)

e.g. +1 = Patients and -1 = Healthy controls
Advantages of Pattern Recognition Analysis

Explore the multivariate nature of neuroimaging data

- MRI/fMRI data are multivariate since most of the brain functions are distributed processes involving a network of brain regions.

- Pattern recognition analysis can yield greater sensitivity than conventional analysis due to its multivariate properties.

Can be used to make predictions for new examples

- Enable clinical applications: previously acquired data can be used to make diagnostic or prognostic for new subjects.
How to extract features from MRI?

Whole brain volume

Region of interest (ROI)

fMRI

General Linear Model

Beta/Contrast images

Feature Vector
Dimensionality = number of voxels

Feature Vector
Dimensionality = number of voxels within the ROI

Feature Vector
Dimensionality = number of voxels
Pattern Classification

Linear classifiers (hyperplanes) are parameterized by a weight vector $\mathbf{w}$ and a bias term $b$. 

Class 1

Class 2

Test example
Pattern Regression

Linear regression models are also parameterized by a weight vector \( \mathbf{w} \) and a bias term \( b \).
Pattern Recognition Models

• Linear predictive models (classifier or regression) are parameterized by a weight vector \( \mathbf{w} \) and a bias term \( b \).

• The weight vector \( \mathbf{w} \) can be expressed as a linear combination of training examples \( \mathbf{x}_i \) (\( N = \) number of training examples).

\[
\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i
\]

• The general equation for making predictions for a test example \( \mathbf{x}_* \) is:

\[
f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b
\]
Weight map or predictive pattern

- The weight vector $\mathbf{w}$ has the same dimensionality of the input data/image and can be plotted as an image.

- The weight vector might provide potential insights into brain function or structure that drives the prediction, but the interpretation should be done with care!
Using the weights for prediction

Predictive function

\[ f(x_*) = w \cdot x_* + b \]

Weight map \((w)\)

New example \((x_*)\)

\[ f(x_*) = 5 \times 1 + 2 \times 2 + (-6 \times -2) + (-1 \times 4) + 0 \]

\[ f(x_*) = 5 + 4 + 12 - 4 = 17 \]

\(f(x_*)\) is the predicted score for regression or the distance to the decision boundary for classification models.
How to interpret the weight vector $w$?

- It is a spatial representation of the predictive function.
- Shows the contribution of each feature/voxel for the prediction.
- Multivariate pattern -> All voxels with weights different from zero contribute to the final prediction (no arbitrary threshold should be applied).
- No local inferences can be made!
Challenges in Neuroimaging

- In neuroimaging applications often the dimensionality of the data is greater than the number of examples (ill-conditioned problems).

- Possible solutions:
  - Region of interest (ROI)
  - Feature selection strategies
  - Searchlight
  - Regularization + Kernel Methods
Regularization

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

- Regularized methods find \( w \) minimizing an objective function consisting of a data fit term \( E \) and a penalty/regularization term \( J \)

\[
\min_{w\in\mathbb{R}^p} \left\{ E(w) + \lambda J(w) \right\}
\]

The data fit term is a error function \( E \)  The regularisation term \( J \)

- Many machine learning algorithms are particular choices of \( E \) and \( J \) (e.g. Kernel Ridge Regression (KRR), Support Vector Machine (SVM)).
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 (mapping into the feature space)
  - A learning algorithm based on the kernel matrix

- **Main advantage:**
  - Computational efficiency
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).
Nonlinear Kernels

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

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

\[
\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3
\]

\[
(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)
\]
Advantage of linear models

- Neuroimaging data are extremely high-dimensional and the sample sizes are very small, therefore non-linear kernels often don’t bring any benefit.

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

• Making predictions with kernel methods

\[ f(x_*) = w \cdot x_* + b \]  \hspace{1cm} \text{Primal representation}

\[ f(x_*) = \sum_{i=1}^{N} \alpha_i x_i \cdot x_* + b \]

\[ f(x_*) = \sum_{i=1}^{N} \alpha_i K(x_i, x_*) + b \]  \hspace{1cm} \text{Dual representation}
Algorithms available in PRoNTo

Kernel methods

• **Classification:**
  ✓ Support Vector Machine (SVM)
  ✓ Multiple Kernel Learning (MKL) Classifier
  ✓ Binary Gaussian Process Classifier (GPC) -> probabilistic
  ✓ Multiclass Gaussian Process Classifier (GPC) -> probabilistic

• **Regression:**
  ✓ Kernel Ridge Regression (KRR)
  ✓ Multiple Kernel Learning (MKL) Regression
  ✓ Relevance Vector Regression (RVR) -> probabilistic
  ✓ Gaussian Process Regression (GPR) -> probabilistic
Support Vector Machine (SVM)

- Sparse solution in terms of examples (support vectors)
- Computational efficient
- Gives good results for most problems
Gaussian Process Classifier – Binary/Multiclass

- Explicit probabilistic framework
- Natural extension to direct multi-class classification
- Provide mechanisms for automatic parameter optimization
Kernel Ridge Regression

- Dual representation of ridge regression, also known as the linear least square regression with Tikhonov regularization (Chu et al. 2011).
Relevance vector machine

Figure 3: SVM (left) and RVM (right) classifiers on 100 examples from Ripley’s Gaussian-mixture data set. The decision boundary is shown dashed, and relevance/support vectors are shown circled to emphasise the dramatic reduction in complexity of the RVM model.

• Probabilistic: apply a Bayesian treatment to SVM

• Similarly to SVM finds a sparse solution (relevance vectors)

• Risk of local minima during optimization
• MKL has been proposed as an approach to learn a decision function based on a predefined set of kernels.
Single kernel SVM

\[ f(x_*) = w \cdot x_* + b \]

Multiple kernel SVM

\[ f(x_*) = \sum_{i=1}^{m} d_m f_m(x_*) \]
Considerations for Neuroimaging

Define your question:
- Classification or regression?
- Specify the subjects and/or conditions
Considerations for fMRI (BOLD)
How to extract features from MRI?
PRoNTo paper:

Reviews:

Books:

Machines:
- Breiman (1996) Bagging Predictors Machine Learning, 24, 123-140