LIÈGE université
GIGA

Machine Learning in Neuroimaging,
what are we talking about?

OHBM 2020 Educational Course

OHBM 2020
A Virtual Experience for Engaging Minds & Empowering Brain Science

fnrs LA LIBERTÉ DE CHERCHER

Christophe Phillips, Ir Ph.D.

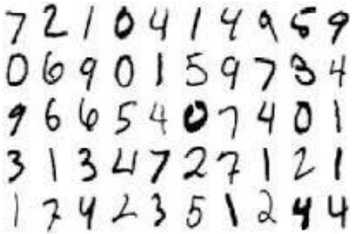
Outline

- Pattern recognition framework
- Mass-univariate vs. pattern recognition analysis
- Linear predictive models: classification and regression
- Regularization & kernel methods
- Validation & inference
- Get home message

Pattern recognition

- Pattern recognition aims to find patterns/regularities in the data that can be used to take actions (e.g. make predictions).

Digit Recognition



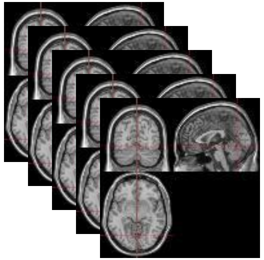
Face Recognition

Recommendation Engines

- Types of Learning:
  - *Supervised* learning: trained with labeled data (classification/regression)
  - *Unsupervised* learning: trained with unlabeled data (clustering)
  - *Reinforcement* learning: actions and rewards (maximize cumulative reward)

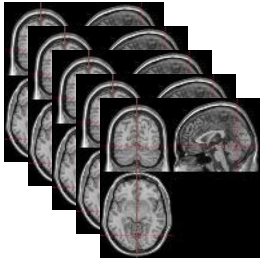
Classification model

Class 1



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

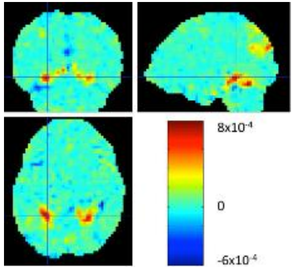
Class 2



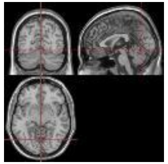
Label = control
Label = control
Label = control
Label = control
Label = control

Training

Predictive function: *f*



New subject

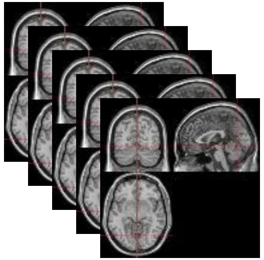


Testing

Prediction:
Class membership
(patient/control)

Regression model

Class 1



Score = 11

Score = 8

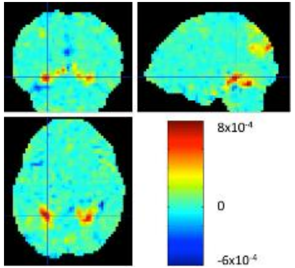
Score = 22

Score = 17

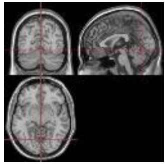
Score = 30

Training

Predictive function: f



New subject



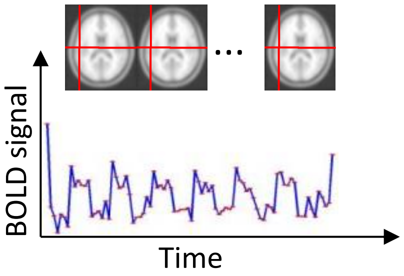
Testing

Prediction:

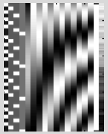
Score = 28

Mass-univariate vs Pattern recognition

Standard Statistical Analysis (mass-univariate)

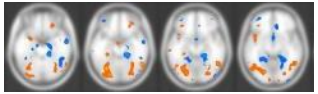


Voxel-wise GLM model estimation



Independent statistical test at each voxel

Correction for multiple comparisons

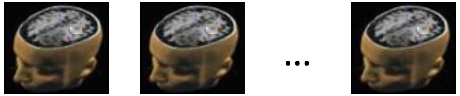


Statistical Parametric Map

Very different meaning!

Pattern Recognition Analysis (multivariate & predictive)

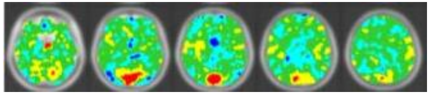
Volumes from task 1



Volumes from task 2

New example

Training Phase



Linear model predictive map
(classification or regression weights)

Testing Phase

Predictions:
task 1 or task 2

Advantages of Pattern Recognition

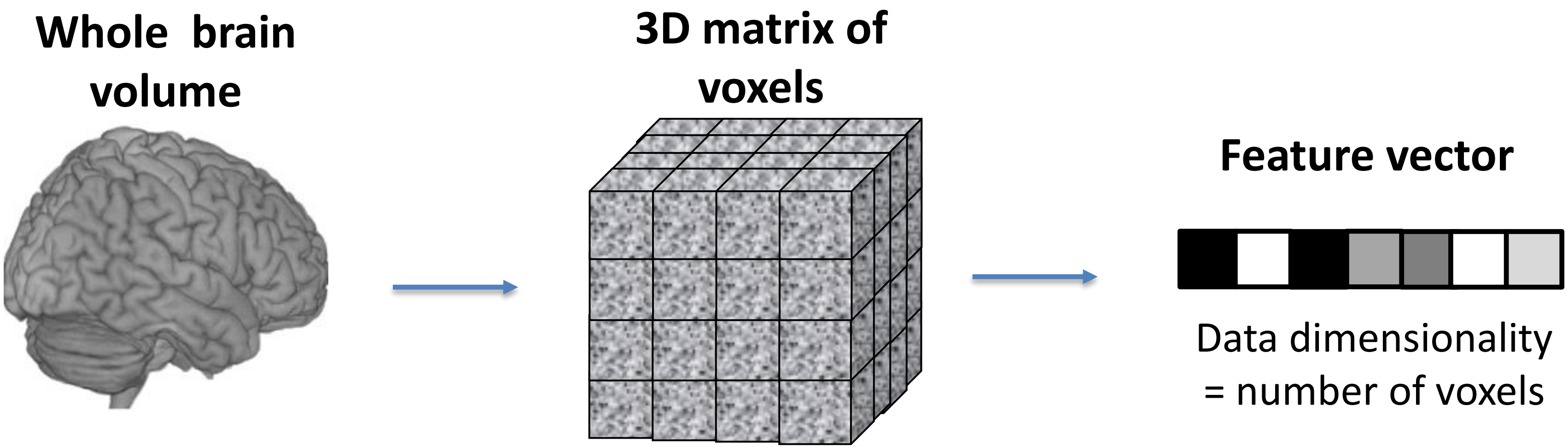
- **Multivariate analysis**: It can be more sensitive to detect spatially distributed effects.

  …but no local inferences.

- **Predictive framework**: Provides predictions for new examples (e.g. new subjects/images).

  …but it typically requires more data!

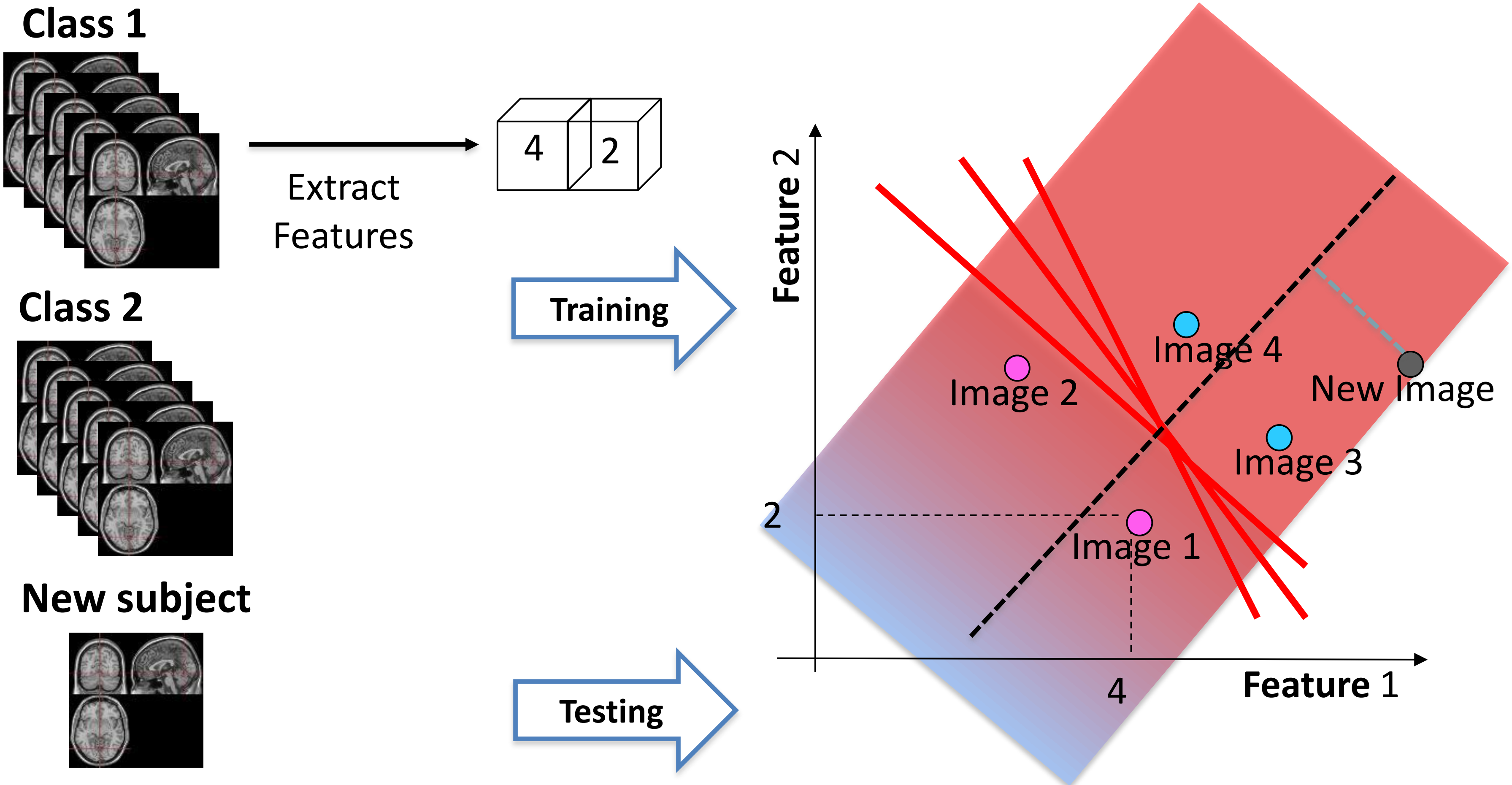
Extracting features from neuroimaging



Other type of features:

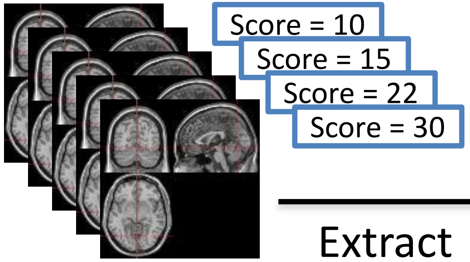
- Volumes of regions of interest (ROIs)
- Connectivity measures
- Cortical Thickness
- …

Classification model

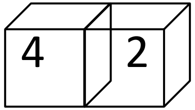


Regression model

Class 1

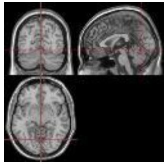


Extract Features

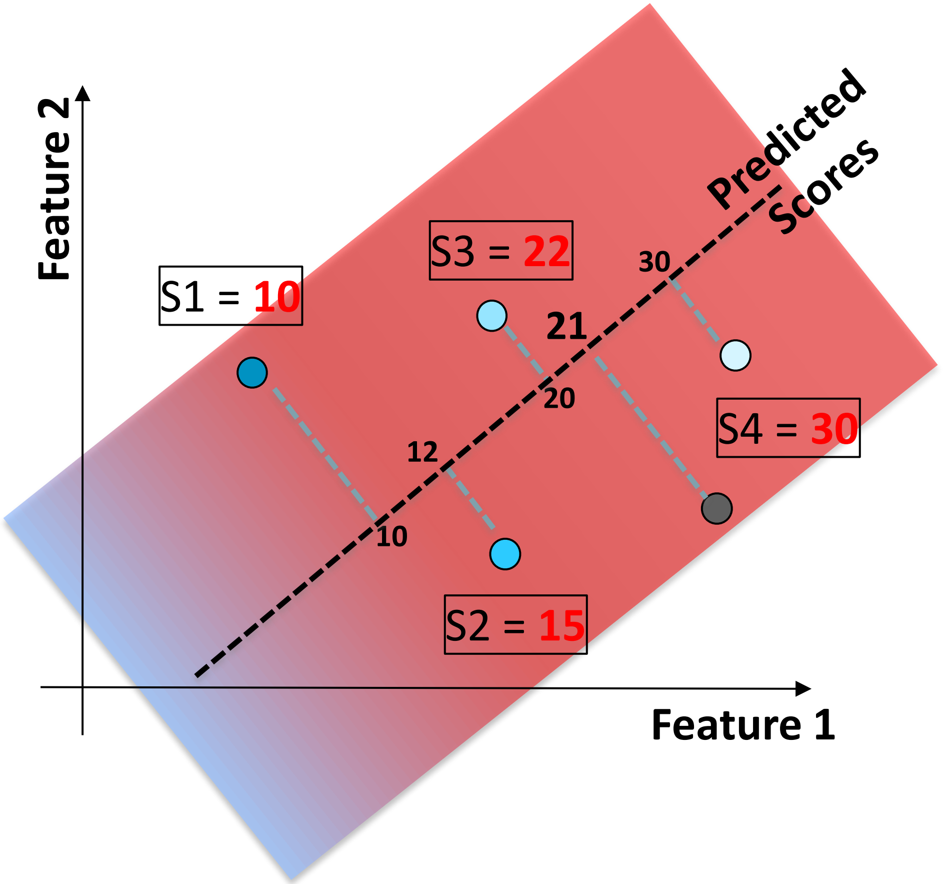


Training

New subject



Testing



Linear predictive models

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

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

Parameters learned/estimated from training data

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

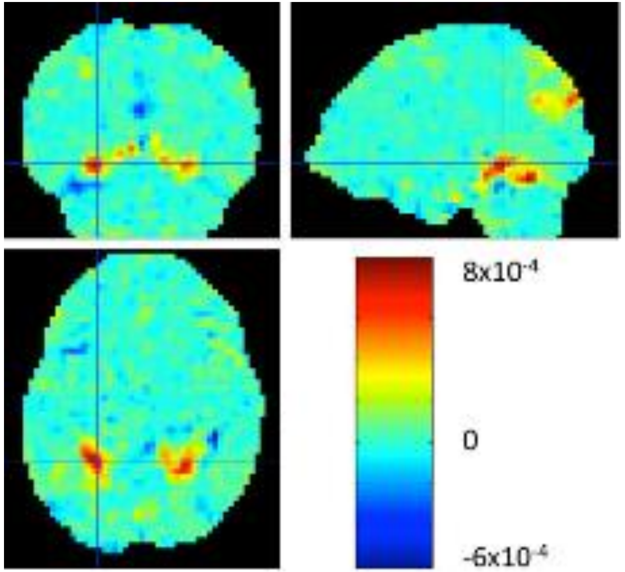
- In the linear case $\boldsymbol{w}$ can be expressed as a linear combination of training examples $\boldsymbol{x}_i$ (N = number of training examples)

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i$$

Weight maps or predictive patterns

Linear prediction model:

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



- Shows the relative contribution of each feature for the decision
- No local inferences can be made!

Pattern recognition in neuroimaging

Common issue with neuroimaging applications:

#features (e.g. voxels) $\gg$ **#samples (e.g. subjects)**

$\Rightarrow$ ill-conditioned problems!

Possible solutions:

- Decrease the number of features
  - Region of interest (ROIs)
  - Feature selection strategies **(DANGER of double dipping!)**
  - Searchlight
- Regularization + Kernel Methods

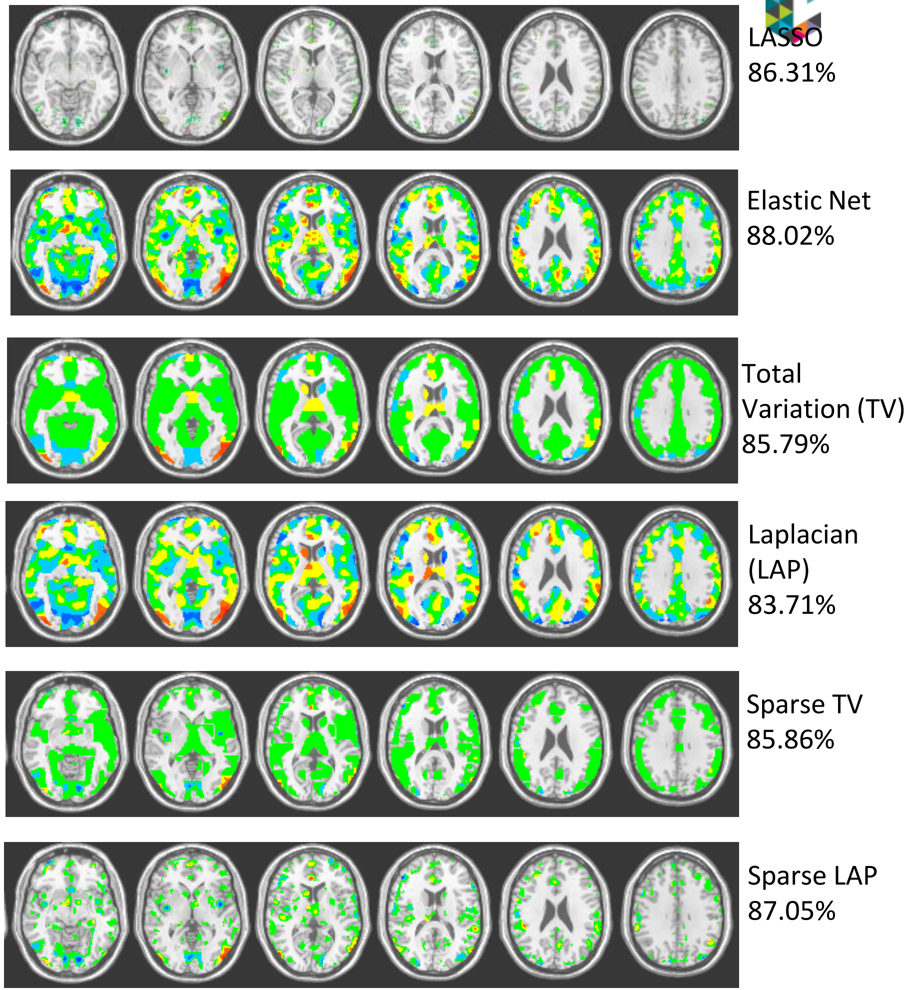
Regularization

- To find a unique solution & avoid overfitting
- Balance between data-fit L & penalty J terms

$$\min_{\boldsymbol{w}\in\mathbb{R}^p}\{L(\boldsymbol{w}) + \lambda\, J(\boldsymbol{w})\}$$

- Different choices of L and J lead to different solutions!

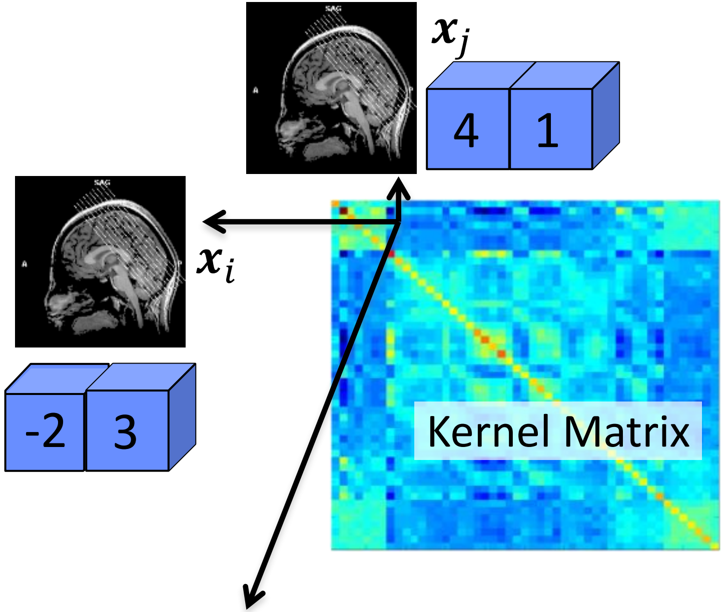
Example: Square loss + different J



Baldassarre et al.(2017)

Kernel Methods

- General framework for classification & regression models
- Relies on 2 parts
  - kernel function $k_{ij} = \mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$
  - algorithm relying on kernel formalism
- Advantages
  - general approach for regularization
  - computational efficiency
  - "kernel trick" (linear & non-linear kernels) to measure "sample similarity"



Linear kernel $\equiv$ dot product

$$k_{ij} = \boldsymbol{x}_i \cdot \boldsymbol{x}_j = (4 * -2) + (1 * 3) = -5$$

Kernel methods & Multi-kernel learning

$$f(\boldsymbol{x}_*) = \boldsymbol{w} \cdot \boldsymbol{x}_* + b \quad \text{where } \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i$$

$$\rightarrow \; f(\boldsymbol{x}_*) = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i \cdot \boldsymbol{x}_* + b$$

$$\rightarrow \; f(\boldsymbol{x}_*) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_*) + b$$

- Example of kernel methods:

  Support Vector Machines (SVM), Kernel Ridge Regression (KRR), Gaussian Process (GP), Kernel Fisher Discriminant, Relevance Vector Regression,…

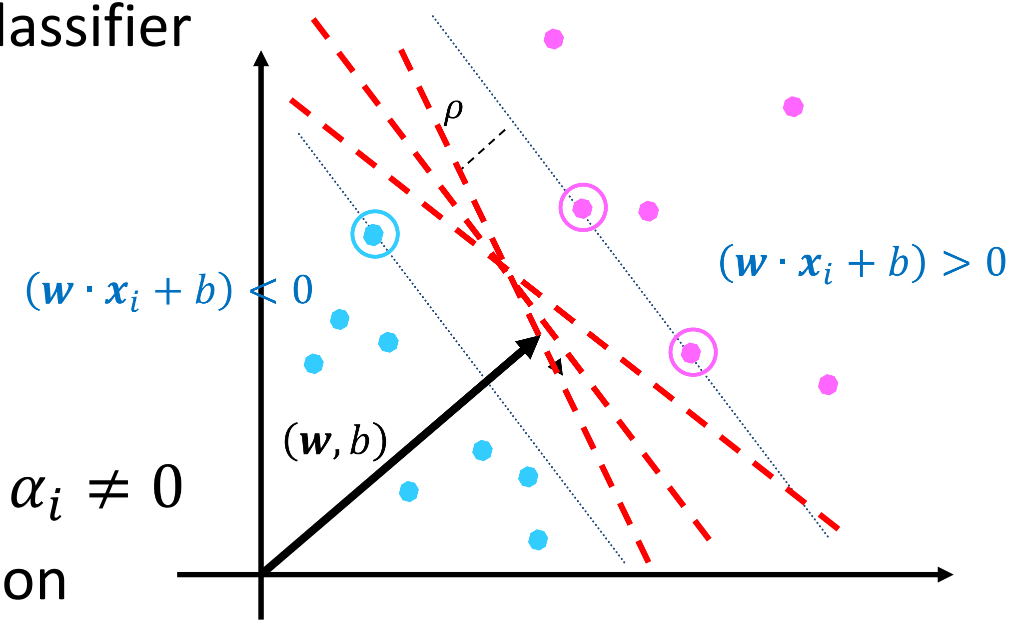
- "Multi-kernel learning" $\equiv$ combine M sub-kernels

  $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{M} d_m K_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $d_m \geq 0$ and $\sum_{m=1}^{M} d_m = 1$

  then learn kernel weight d_m and decision function $(\boldsymbol{w}, b)$.

Support Vector Machine

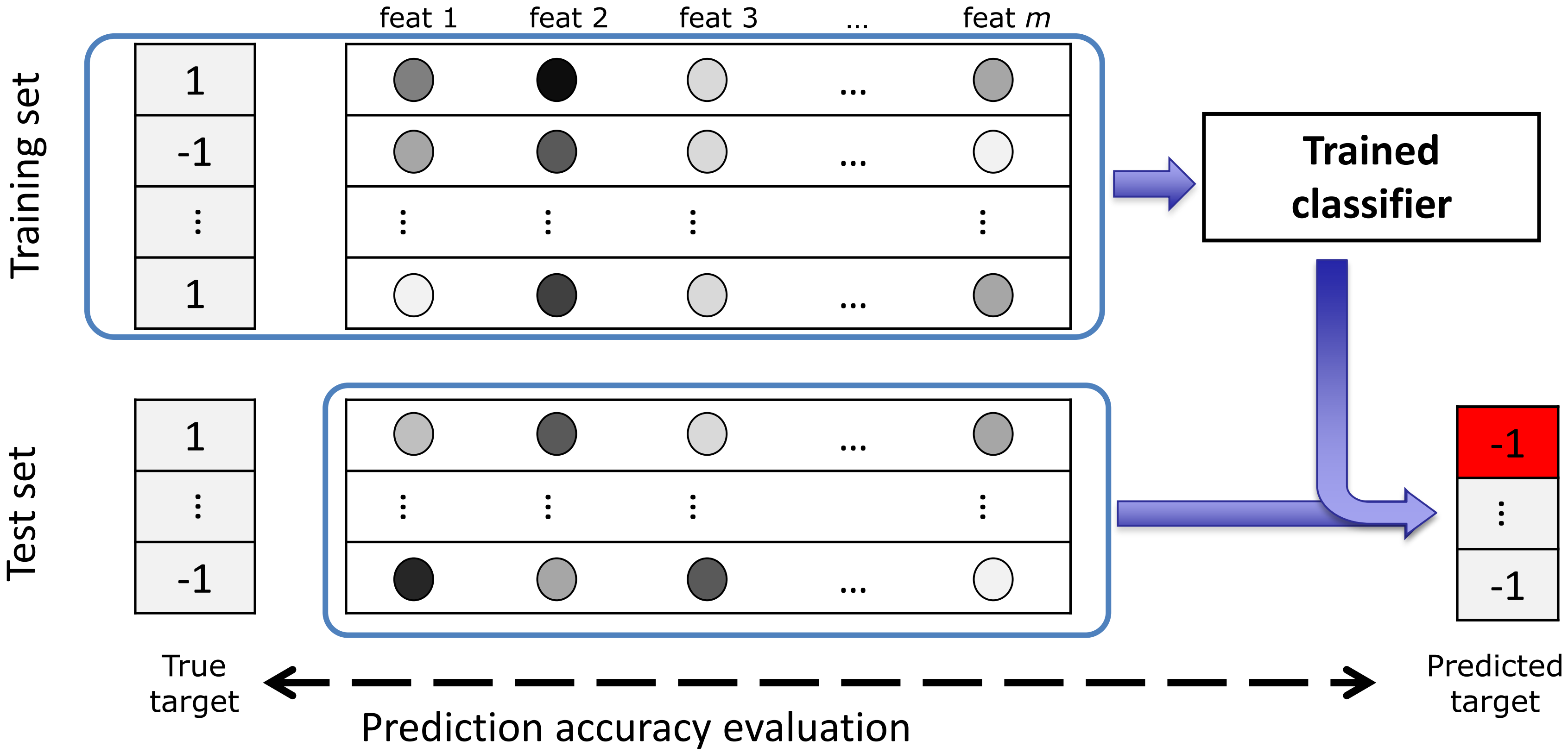
- Relies on kernel representation
- "maximum margin" ρ classifier



- "Support vectors" have $\alpha_i \neq 0$
- Fast & resilient estimation
- …but only "hard binary" prediction!

Validation principle

→ Out of sample prediction!



Prediction assessment

- Classification → confusion matrix
  - Accuracy: total, class specific, or balanced

    $A_{\text{tot}} = \frac{A+D}{A+B+C+D}$, $A_{c1} = \frac{A}{A+B}$ & $A_{c0} = \frac{D}{C+D}$,

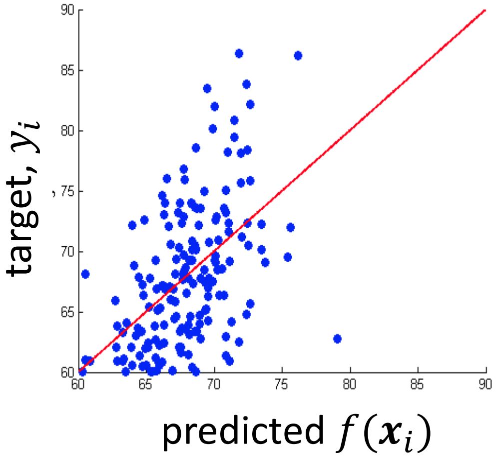
    or $A_{bal} = \frac{A_{c1}+A_{c0}}{2}$
  - Sensitivity & specificity
  - Positive/negative predictive value

		Predicted class	
		1	0
Actual class	1	A	B
	0	C	D

- Regression → mean squared error

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2$$

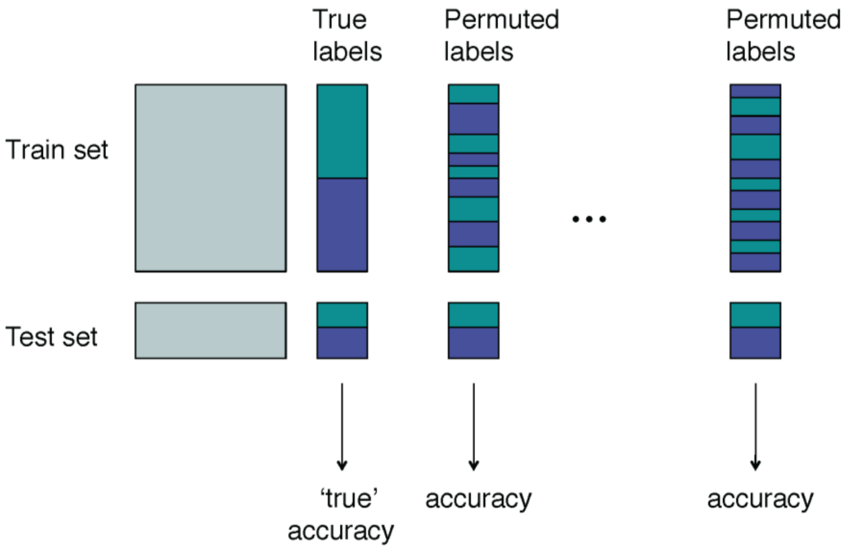
(or correlation between true & predicted scores)



Model inference

- H_0: “no link between features and target”
- Test statistic, e.g. cross-validation (CV) accuracy A
- Estimate distribution of test statistic under H_0
  - → Random permutation of labels
  - → Estimate CV accuracy, A_m
  - → Repeat M times
- Calculate p-value p as

$$p = \frac{1}{M} \sum_{m=1}^{M} \#(A_m \geq A)$$



Conclusions

Univariate

- 1 voxel
- Target → Data
- Look for difference or correlation
- General Linear Model
- GLM inversion
  → parameter & error terms
- Inference on contrast of interest
- Voxel/cluster activation inference
  → localisation

Multivariate

- 1 volume
- Data → Target
- Look for similarity or score
- Specific machine
- Machine training
  → machine parameters
- Prediction accuracy with CV
- Sample label prediction inference
  → no localisation

@CodeWisdom

"A computer is like a mischievous genie.
It will give you exactly what you ask for,
but not always what you want."

- Joe Sondow

References

- **Reviews:**
  - Haynes and Rees (2006) Decoding mental states from brain activity in humans. Nat. Rev. Neurosci., 7, 523-534
  - Pereira, Mitchell, Botnivik (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage,45, S199-S209
- **Books:**
  - Hastie , Tibishirani, Friedman (2003) Elements of Statistical Learning. Springer
  - Shawe-Taylor and Christianini (2004) Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press.
  - Bishop, Jordan, Kleinberg, Schölkopf (2006) Pattern Recognition and Machine learning. Springer
- **Machines:**
  - 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
  - Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. Journal of Machine Learning Research, 9, 2491-2521.

Thank you for your attention!