WEEK 6
SUPERVISED LEARNING

Oualid Benkarim, PhD
Boris Bernhardt, PhD
Bratislav Misic, PhD

Outline

1. What is supervised learning?
2. Linear models
3. The Bias-Variance trade-off
4. Other supervised approaches
5. Cross-validation and model selection
6. Regularization
7. Papers

Outline

Supervised learning

Learn a function f that maps data $\mathbf{x}$ to labels y based on known data-label pairs:

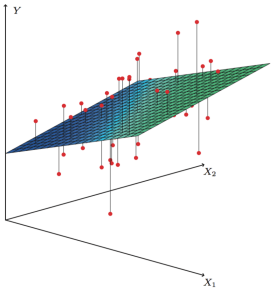
$$y = f(\mathbf{x}) + \epsilon$$

Supervised learning

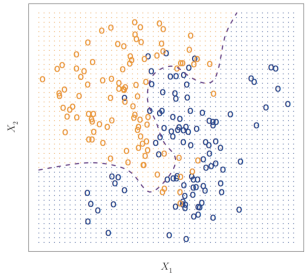
Learn a function f that maps data $\mathbf{x}$ to labels y based on known data-label pairs:

$$y = f(\mathbf{x}) + \epsilon$$

Regression



Classification

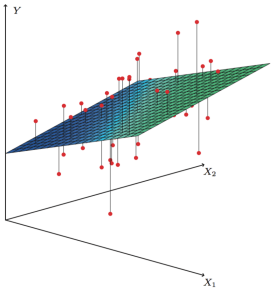


Supervised learning

Learn a function f that maps data $\mathbf{x}$ to labels y based on known data-label pairs:

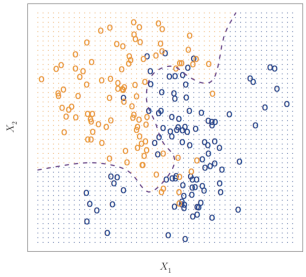
$$y = f(\mathbf{x}) + \epsilon$$

Regression



- $y \in \mathcal{R}$
- Metrics: MSE, MAE, RMSE,...

Classification



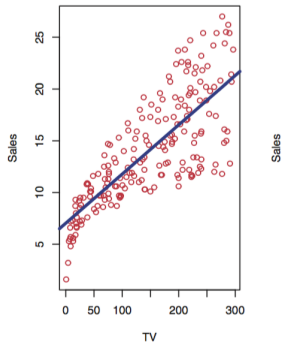
- y categorical
- Metrics: Accuracy, sensitivity,...

Outline

Linear models

Linear and logistic regression

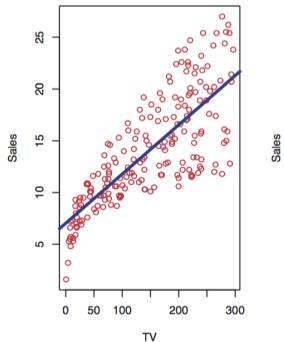
- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$



Linear models

Linear and logistic regression

- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$



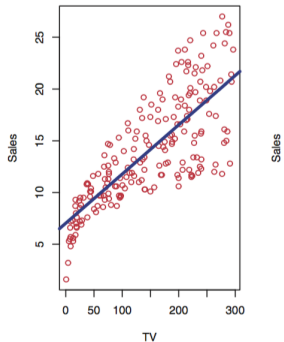
- OLS:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

Linear models

Linear and logistic regression

- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$



- OLS:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

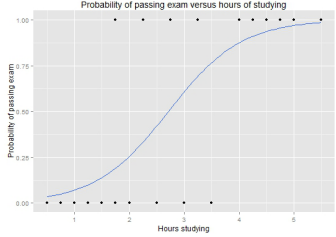
- Analytical solution:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Linear models

Linear and logistic regression

- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$
- Logistic regression: $\hat{y} = S(f(\mathbf{x})) = S(\mathbf{w}^T \mathbf{x})$



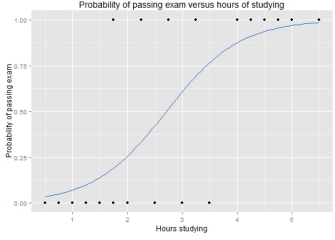
Linear models

Linear and logistic regression

- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$
- Logistic regression: $\hat{y} = S(f(\mathbf{x})) = S(\mathbf{w}^T \mathbf{x})$

- Sigmoid:

$$p(y = 1|\mathbf{x}) = S(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$



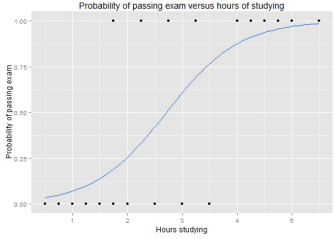
Linear models

Linear and logistic regression

- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$
- Logistic regression: $\hat{y} = S(f(\mathbf{x})) = S(\mathbf{w}^T \mathbf{x})$

- Sigmoid:

$$p(y = 1|\mathbf{x}) = S(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$



- Assume $y \in \{-1, 1\}$:

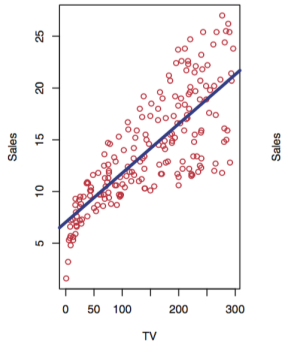
$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \log(1 + e^{-y_i \hat{y}_i})$$

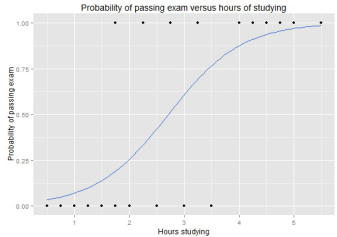
- GD, Newton's method

Linear models

Linear and logistic regression

- Linear regression: $\hat{y} = f(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$
- Logistic regression: $\hat{y} = S(f(\mathbf{x})) = S(\mathbf{w}^T \mathbf{x})$



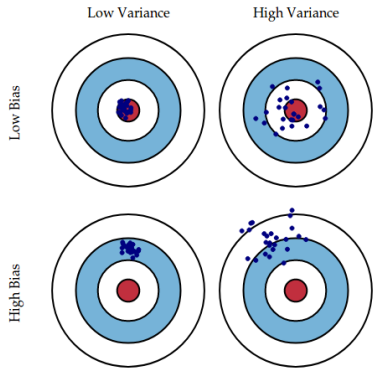


Outline

The Bias-Variance trade-off

Prediction error:

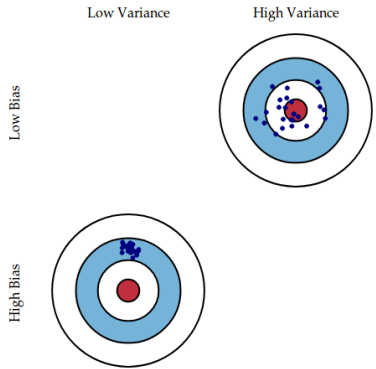
- Bias: difference between model prediction and true target
- Variance: variability of model prediction
- Other: noise, unknown factors, ...



The Bias-Variance trade-off

Prediction error:

- Bias: difference between model prediction and true target
- Variance: variability of model prediction
- Other: noise, unknown factors, ...

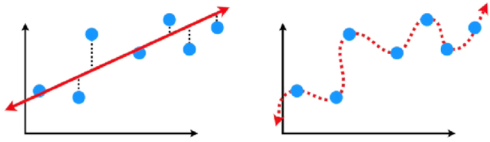


The trade-off

Cannot minimize both bias and variance!

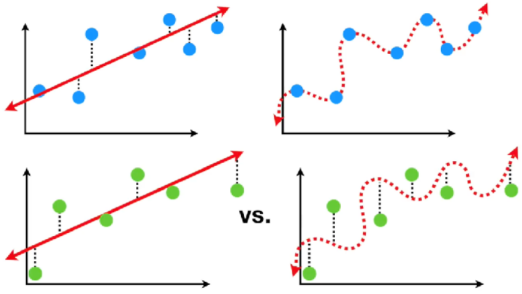
The Bias-Variance trade-off

Under- vs Overfitting



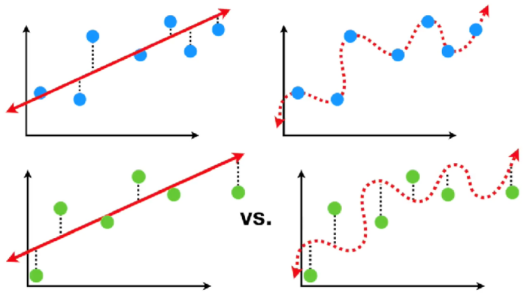
The Bias-Variance trade-off

Under- vs Overfitting



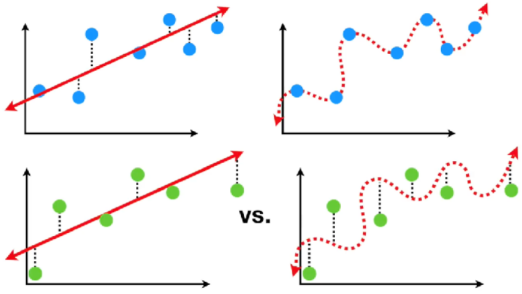
The Bias-Variance trade-off

Under- vs Overfitting



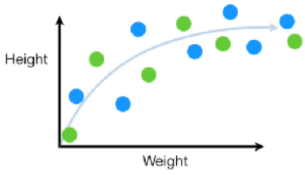
The Bias-Variance trade-off

Under- vs Overfitting



Underfitting
(high bias)

Overfitting
(high variance)



Just right!

Outline

Naive Bayes

- Given a test sample $\mathbf{x} = (x_1, \ldots, x_n)$, assign class label c such that

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, p(c \mid x_1, \ldots, x_n)$$

Naive Bayes

- Given a test sample $\mathbf{x} = (x_1, \ldots, x_n)$, assign class label c such that

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, p(c \mid x_1, \ldots, x_n)$$

- Apply Bayes rule:

$$p(c \mid \mathbf{x}) = \frac{p(c)p(\mathbf{x} \mid c)}{p(\mathbf{x})}$$

  where:
  - $p(c)$ is the class prior
  - $p(\mathbf{x} \mid c)$ is the likelihood
  - $p(\mathbf{x}) = \sum_{c \in C} p(c)p(\mathbf{x} \mid c)$ is constant

Naive Bayes

- Given a test sample $\mathbf{x} = (x_1, \ldots, x_n)$, assign class label c such that

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, p(c \mid x_1, \ldots, x_n)$$

- Apply Bayes rule:

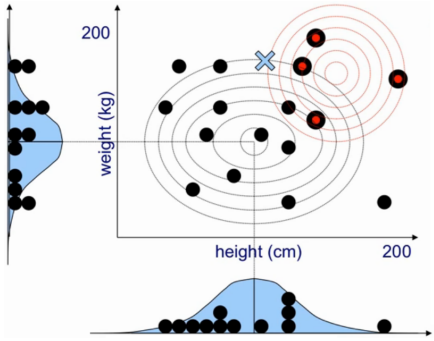
$$p(c \mid \mathbf{x}) = \frac{p(c)p(\mathbf{x} \mid c)}{p(\mathbf{x})}$$

- Be naive:

$$\begin{aligned} p(c \mid x_1, \ldots, x_n) &\propto p(c, x_1, \ldots, x_n) \\ &\propto p(c)p(x_1 \mid c) \cdots p(x_n \mid c) \\ &\propto p(c) \prod_{i=1}^{n} p(x_i \mid c). \end{aligned}$$

Gaussian Naive Bayes

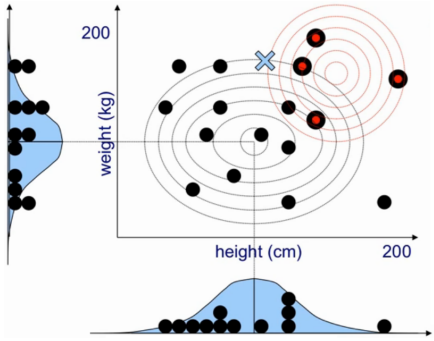
Prior and likelihood in continuous case



Gaussian Naive Bayes

Prior and likelihood in continuous case

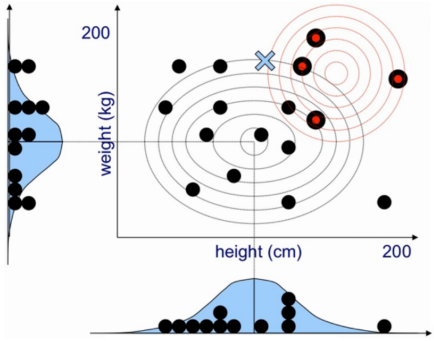
1. Prior probabilities:
   - $p(c_1) = \frac{4}{4+12} = 0.25$
   - $p(c_2) = \frac{12}{4+12} = 0.75$



Gaussian Naive Bayes

Prior and likelihood in continuous case

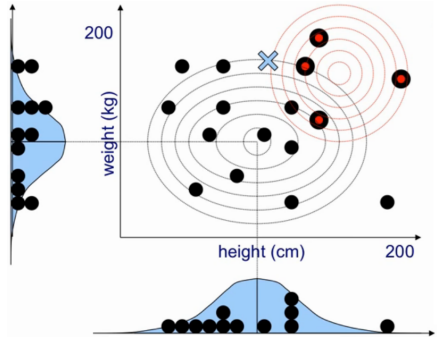
1. Prior probabilities:
   - $p(c_1) = \frac{4}{4+12} = 0.25$
   - $p(c_2) = \frac{12}{4+12} = 0.75$
2. Likelihood:
   - $p(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_{i,c}^2}} e^{-\left(\frac{(x_i - \mu_{i,c})^2}{2\sigma_{i,c}^2}\right)}$



Gaussian Naive Bayes

Prior and likelihood in continuous case

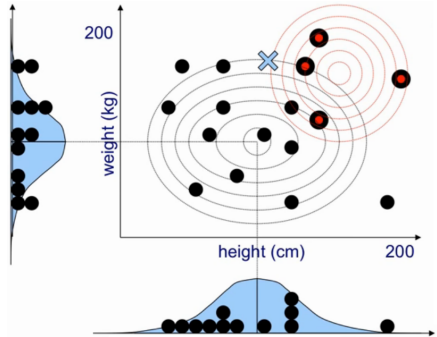
1. Prior probabilities:
   - $p(c_1) = \frac{4}{4+12} = 0.25$
   - $p(c_2) = \frac{12}{4+12} = 0.75$
2. Likelihood:
   - $p(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_{i,c}^2}} e^{-\left(\frac{(x_i-\mu_{i,c})^2}{2\sigma_{i,c}^2}\right)}$
   - $p(\mathbf{x} \mid c_1) = p(x_1 \mid c_1) p(x_2 \mid c_1)$
   - $p(\mathbf{x} \mid c_2) = p(x_1 \mid c_2) p(x_2 \mid c_2)$



Gaussian Naive Bayes

Prior and likelihood in continuous case

1. Prior probabilities:
   - $p(c_1) = \frac{4}{4+12} = 0.25$
   - $p(c_2) = \frac{12}{4+12} = 0.75$
2. Likelihood:
   - $p(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_{i,c}^2}} e^{-\left(\frac{(x_i - \mu_{i,c})^2}{2\sigma_{i,c}^2}\right)}$
   - $p(\mathbf{x} \mid c_1) = p(x_1 \mid c_1) p(x_2 \mid c_1)$
   - $p(\mathbf{x} \mid c_2) = p(x_1 \mid c_2) p(x_2 \mid c_2)$

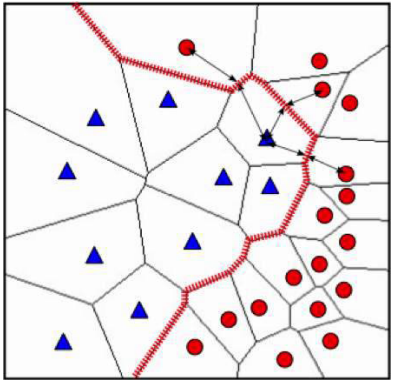


$$p(c_1 \mid \mathbf{x}) = \frac{p(c_1) p(\mathbf{x} \mid c_1)}{p(c_1) p(\mathbf{x} \mid c_1) + p(c_2) p(\mathbf{x} \mid c_2)}$$

K-Nearest Neighbor

Prediction based on k closest training examples

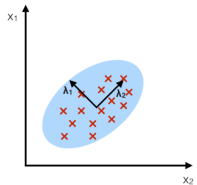
- Define distance
- Hyperparameter: k
- Majority voting
- Lazy learner
- Fast implementation: kd-tree, ...



Linear Discriminant Analysis

Project data onto space with maximum class separation

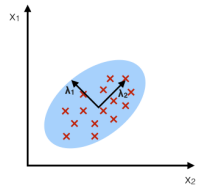
PCA:
component axes that
maximize the variance



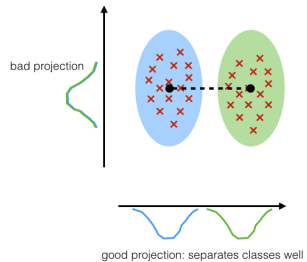
Linear Discriminant Analysis

Project data onto space with maximum class separation

PCA:
component axes that maximize the variance

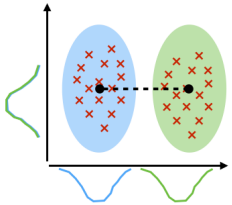


LDA:
maximizing the component axes for class-separation

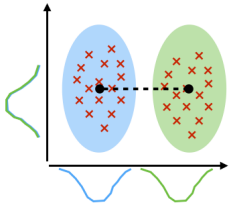


Linear Discriminant Analysis

1. Compute scatter matrices:
   - between-class: maximize
   - within-class: minimize

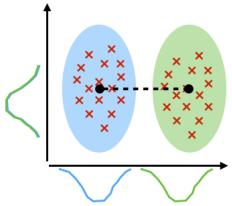


Linear Discriminant Analysis



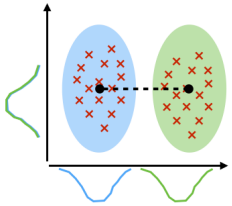
1. Compute scatter matrices:
   - between-class: maximize
   - within-class: minimize
2. Compute eigenvectors $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d)$ and corresponding eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_d)$

Linear Discriminant Analysis



1. Compute scatter matrices:
   - between-class: maximize
   - within-class: minimize
2. Compute eigenvectors $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d)$ and corresponding eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_d)$
3. Choose k eigenvectors with the largest eigenvalues: $\mathbf{W} \in \mathcal{R}^{d \times k}$

Linear Discriminant Analysis

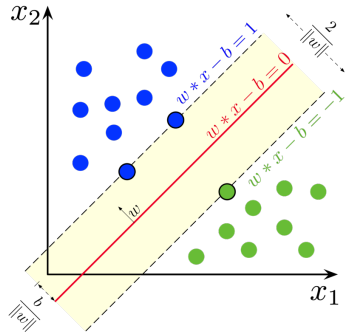


1. Compute scatter matrices:
   - between-class: maximize
   - within-class: minimize
2. Compute eigenvectors $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d)$ and corresponding eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_d)$
3. Choose k eigenvectors with the largest eigenvalues: $\mathbf{W} \in \mathcal{R}^{d \times k}$
4. Transform original samples onto new subspace:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

Support vector machines

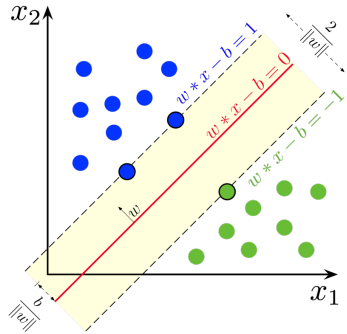
Find hyperplane that maximizes margin between classes



Support vector machines

Find hyperplane that maximizes margin between classes

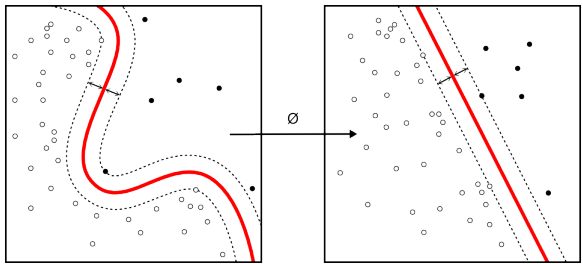
- Generalization
- Slack variables



Support vector machines

Find hyperplane that maximizes margin between classes

- Generalization
- Slack variables
- Linear separability in high-dimensional space
- Kernel trick

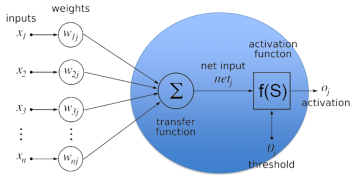


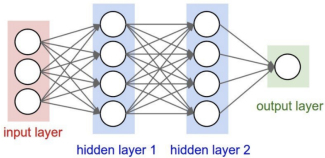
Other learners

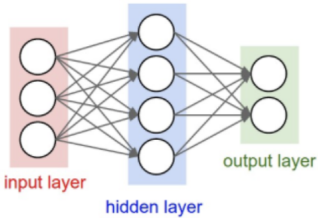
- Decision trees
- Random forests
- AdaBoost
- Neural networks (shallow)
- Deep NN

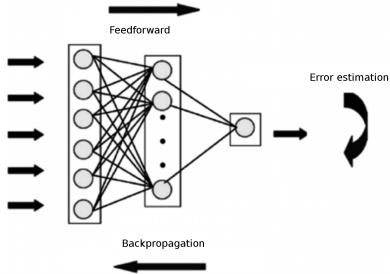
“No Free Lunch” (Wolpert & Macready, 1997)

From ANN to DL



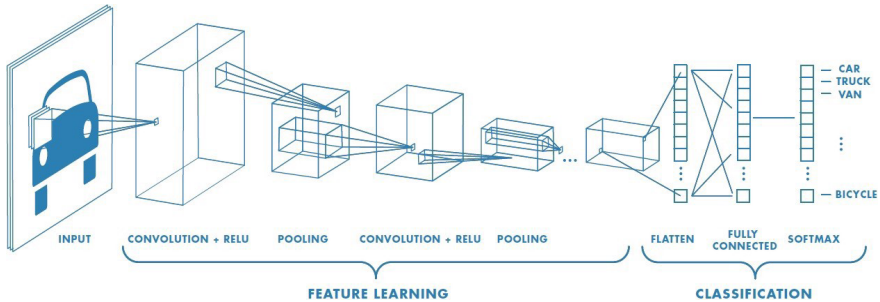






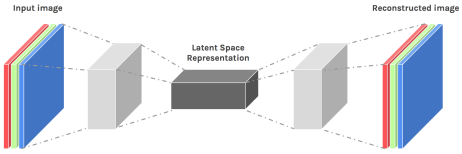
From ANN to DL

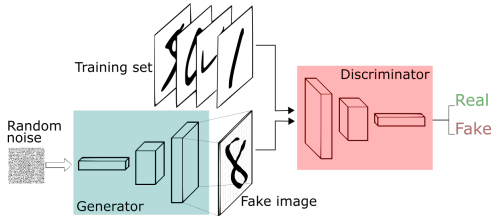
CNN

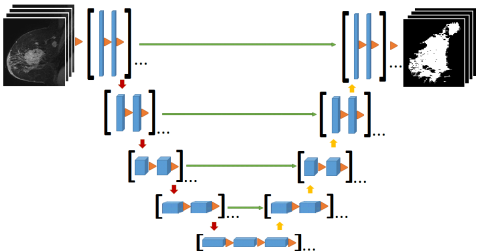


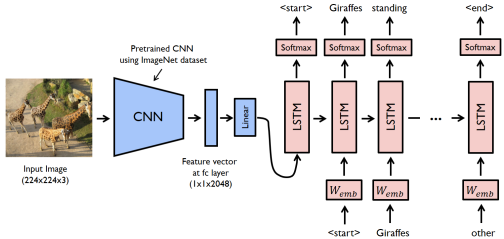
From ANN to DL

Other architectures





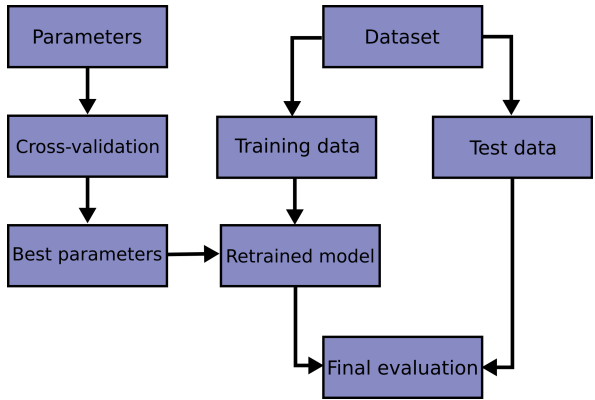




Outline

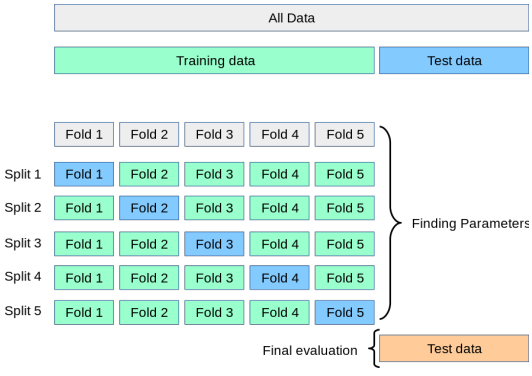
Cross-validation and model selection

General approach



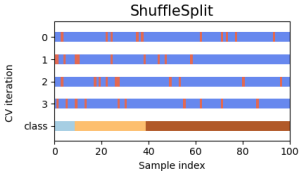
Cross-validation and model selection

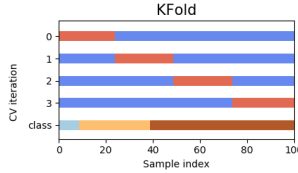
General approach

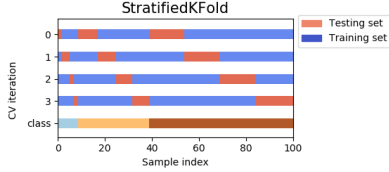


- Split-half
- Leave-one-out
- K-Fold
- etc

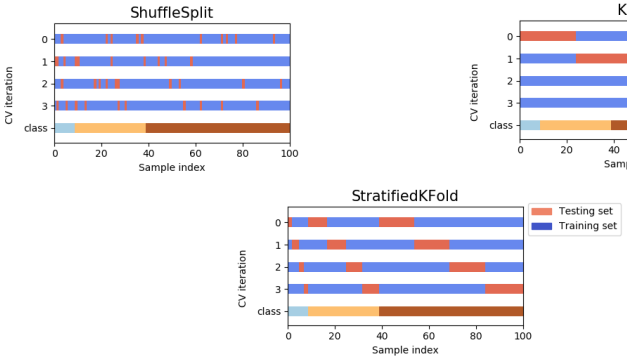
Cross-validation and model selection





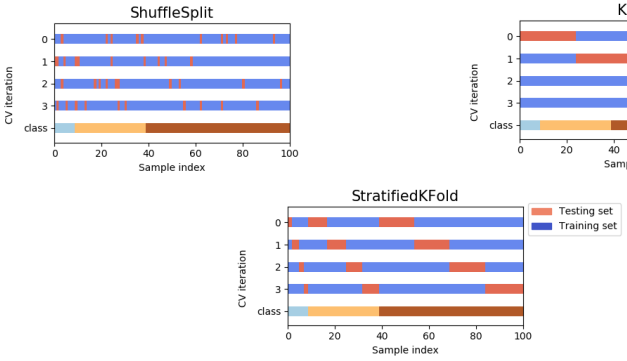


Cross-validation and model selection



- Stratified sampling:
  - Preserve class proportions
  - Deal with imbalance

Cross-validation and model selection



- Stratified sampling:
  - Preserve class proportions
  - Deal with imbalance
- What about regression?

Outline

Regularization

Sparse linear models

1. OLS: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$

Regularization

Sparse linear models

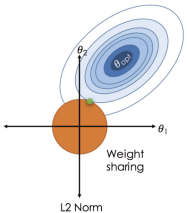
1. OLS: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$

Sparsity-inducing norms:

Regularization

Sparse linear models

1. OLS: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$
2. Ridge: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$



Sparsity-inducing norms:

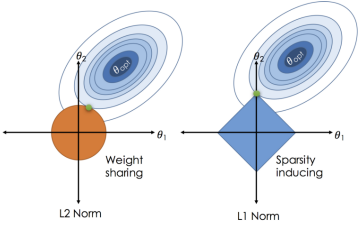
- ℓ_2-norm:

$$\|\mathbf{w}\|_2 = \sqrt{\sum_i w_i^2}$$

Regularization

Sparse linear models

1. OLS: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$
2. Ridge: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$
3. Lasso: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$



Sparsity-inducing norms:

- ℓ_2-norm:

$$\|\mathbf{w}\|_2 = \sqrt{\sum_i w_i^2}$$

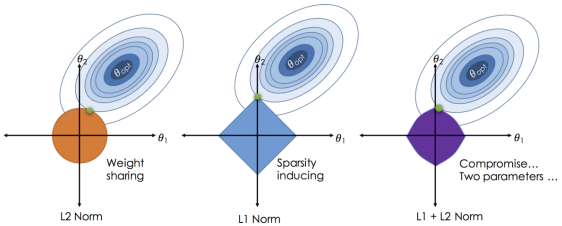
- ℓ_1-norm:

$$\|\mathbf{w}\|_1 = \sum_i |w_i|$$

Regularization

Sparse linear models

1. OLS: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$
2. Ridge: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$
3. Lasso: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$
4. Elastic Net: $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 + \frac{\lambda_2}{2} \|\mathbf{w}\|_2^2$



Sparsity-inducing norms:

- ℓ_2-norm:

$$\|\mathbf{w}\|_2 = \sqrt{\sum_i w_i^2}$$

- ℓ_1-norm:

$$\|\mathbf{w}\|_1 = \sum_i |w_i|$$

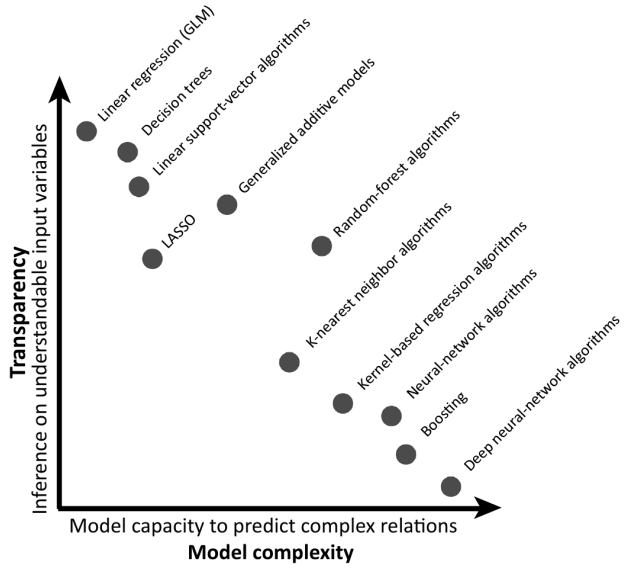
Outline

BZDOK

Table 1. The Inference–Prediction Continuum of Modeling Goals (*cf* Figure 1)

Inference <----------> Prediction	
Commonly Used Tools for Inference Goals Null hypothesis significance testing to compute *P* values for specific target variables. Tools for this purpose include, for example, ANOVA, the *t* test, or χ^2 test. Increasingly popular alternatives include false discovery rate and Bayesian posterior inference, as well as some pattern-learning algorithms (e.g., feature importance scores from random-forest algorithms).	Commonly Used Tools for Prediction Goals Empirical validation schemes to compute prediction accuracy of the built model as a whole. Exemplary tools include support-vector machines, random-forest algorithms, and other ensemble and boosting techniques, the rapidly evolving 'deep'-learning algorithms, as well as ordinary and penalized linear regression.
Knowledge-Guided Candidate variables are often hand-picked by the investigator in a targeted fashion based on existing substantive knowledge. Research questions are explicitly articulated before data collection in a carefully controlled experiment. The chosen variables are evaluated by an often simple but inflexible model that ideally is prespecified by the investigator before seeing the data. However, data dredging, and thus a high false-positive rate, are common in practice.	Pattern-Guided A large and diverse array of 'found' variables is typically considered in the statistical analysis in a heuristic data-led fashion. It can be unknown how the data were generated, and the exact research question may be detailed as the data are being analyzed. The adaptive and sometimes very flexible model extracts a general prediction rule directly from the data in the spirit of 'letting the data speak for themselves'.
Explainable Narrative Statements about the specific contribution of individual input variables are the priority. Such claims of variable relevance are often more readily available in simple linear-regression models. Accordingly, these models tend to be preferred in the context of inference such that every single parameter, and its corresponding unit, can be cleanly attributed its share of the explained variance. Usually, the meaning of each parameter should be readily understood, and hence the model often allows for a simplified narrative; statements are centered on single parameters rather than on the prediction performance of the collective model parameters.	Opaque Black Box Although simple linear-regression models may perform reasonably well in terms of predictive power, if the goal is to maximize prediction accuracy, it is often beneficial to exploit complex non-additive associations in the data. In many real-world situations the target variable depends on the input variables in convoluted ways, which can hinder assigning to single input variables a clear relative contribution to the output, and model parameters are often treated as instrumental intermediates to achieve high prediction performance without necessarily aiming to assign specific meaning to each parameter estimate *per se*.

Formally Justified Many traditional analysis techniques were rigorously characterized and validated by mathematical theory; simple linear models lend themselves well to theoretical model criticism, and carry well-understood modeling limits; another benefit of formal performance guarantees is the typically lower computational load.	Empirically Justified Predictive models can be explicitly and quantitatively evaluated by applying the entire set of estimated model parameters to unseen independent, newly generated, or future observations or individuals; formal performance guarantees are often challenging; these models are often informally validated by means of more computationally demanding cross-validation, bootstrapping, and other resampling schemes.
Data-Efficient Many methods from classical statistics were designed long ago to handle data that are scarce, as well as being laborious and expensive to collect.	Data-Hungry Compared to classical statistics methods, many sophisticated predictive approaches require more data, especially when complex non-linear relationships are to be modeled and more hyperparameters need to be tuned; comparably more data also tend to be needed if each observation has many input variables, and if random noise is expected to be prominent (e.g., medical data).
Problem-Tailored Each approach is designed to solve a particular data-analysis question, typically based on problem-specificprobabilistic and distributional assumptions about how the investigator believes the data have come about.	Versatile Approaches are devised to provide useful solutions to various types of data and data-analysis questions.


Linear regression (GLM)
Decision trees
Linear support-vector algorithms
Generalized additive models
LASSO
Random-forest algorithms
K-nearest neighbor algorithms
Kernel-based regression algorithms
Neural-network algorithms
Boosting
Deep neural-network algorithms
Transparency
Inference on understandable input variables
Model capacity to predict complex relations
Model complexity

Box 1. Stages of Translating Predictive Approaches in Brain Research into Practice

(i) Model Building

To fit the parameters of the chosen predictive model, one first needs empirical measurements from the brain systems of interest. One common preparatory analysis is to probe variable–variable relationships using pairwise correlation plots. Another is to estimate genetic relatedness between the participants using principal component analysis of their genomic profiles. In behavioral experiments in animals or humans, exploratory data summaries can identify collinearity in response times. Such collinearity in response times foreshadows hindered statements about the relevance of individual experimental conditions (i.e., inference), but hardly affects forecasting condition response latencies in new participants (i.e., prediction).

(ii) Internal Validation

These procedures guard against overly optimistic modeling performances. Internal validation procedures, unlike external procedures (point iii), do not require new and independent data and are based only on the original subject sample or dataset that was used during model building [65]. Cross-validation and bootstrapping are resampling schemes ([21], chapter 7) that can estimate metrics of model quality [47], such as expected prediction accuracy for future data, uncertainty of parameter estimates, and variability of prediction errors. Indeed, 'working scientists often find the most interesting aspect of the analysis in the lack of fit rather than the fit itself' ([16], p. 92). Nevertheless, interindividual variability may still be underappreciated by using such internal validations alone [24].

(iii) External Validation

For stronger validation, predictive associations identified from the original subject sample or dataset need to be ascertained in other individuals or in datasets measured later [60,64,66]. Successful application of a predictive model of disease risk, for instance, requires validation in different groups of individuals [24,29]. This step is important to combat reproducibility issues [67]. Currently,external model validations are not done as often as they should be [68]. However, it is important to comprehensively benchmark the value of each predictive approach for clinicians, policymakers, and clinical guidelines [69]. For instance, external validation may be performed in different geographical areas, time periods, and settings (e.g., secondary vs primary care). Generally, some authors have proposed that 'the most stringent external validation involves testing a final model developed in one country or setting on subjects in another country or setting at another time. This validation would test whether the data collection instrument was translated into another language properly, whether cultural differences make earlier findings nonapplicable, and whether secular trends have changed associations or base rates' ([16], chapter 5.3.1).

(iv) Generalizability and Transposability

When evaluating the predictions of a model on new individuals, the more different these individuals are from the original subject sample, the stronger the test for generalizability [59,65]. Prediction accuracies are typically lower than in preceding steps. For instance, our ability to predict the clinical utility of drugs tends to be hindered for particular groups of patients, including women, children, and the elderly. Common comorbidities are also frequently under-represented or intentionally excluded in clinical studies. Meta-analysis methods can be useful for summarizing and examining the predictive performance of a model across different scenarios. Large datasets from multiple studies and electronic health records or registry databases provide promising opportunities for examining the generalizability of predictive approaches [70].

To enhance reproducibility, accurate and complete reporting is imperative for studies applying predictive models. Such reporting is crucial for being able to critically appraise predictive models, to perform acid-test validations of them, to evaluate their impact, and ultimately to translate them into clinical practice [27,71].

CHANG

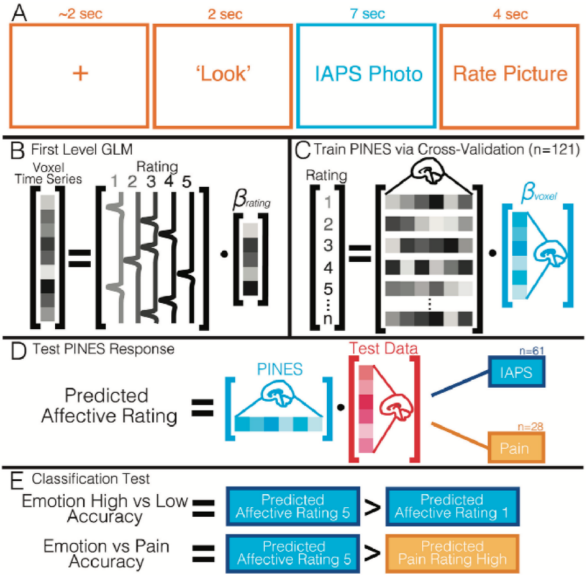


Fig 1. Experimental paradigm and analysis overview. Panel A depicts the sequence of events for a given trial. Participants view an initial fixation cross and then are instructed to look at the picture (compared to reappraise). Participants then see a photo and are asked to rate how negative they feel on a likert scale of 1–5. Panel B illustrates the temporal data reduction for each rating level using voxel-wise univariate analysis and an assumed hemodynamic response function. Panel C: these voxels are then treated as features and trained to predict ratings using LASSO-PCR with leave-one-subject-out cross validation. Subject's data for each rating is concatenated across participants. Panel D: this multivoxel weight map pattern can be tested on new data using matrix multiplication to produce a scalar affective rating prediction. Panel E: we calculated two different types of classification accuracy: (a) the ability to discriminate between high (rating = 5) and low (rating = 1) affective ratings and (b) the ability to discriminate between high affective and high pain data.

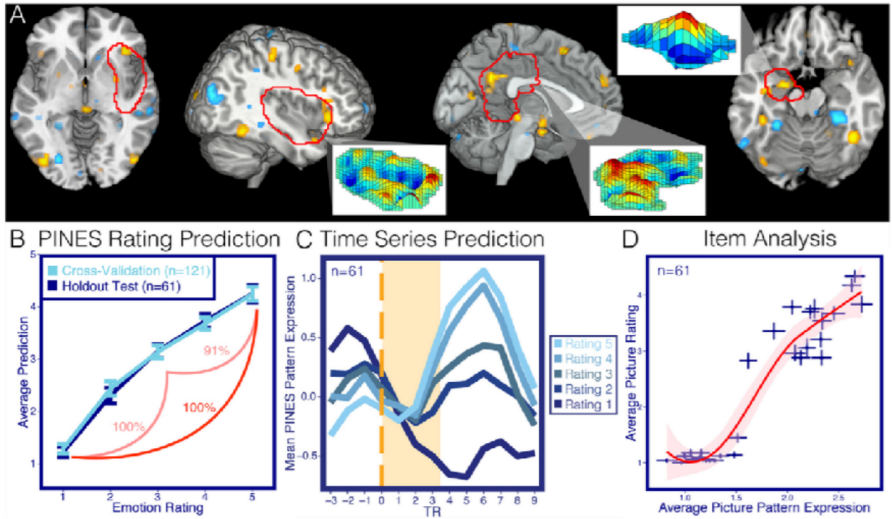


Fig 2. PINES. Panel A depicts the PINES pattern thresholded using a 5,000 sample bootstrap procedure at $p < 0.001$ uncorrected. Blowout sections show the spatial topography of the pattern in the left amygdala, right insula, and posterior cingulate cortex. Panel B shows the predicted affective rating compared to the actual ratings for the cross validated participants ($n = 121$) and the separate holdout test data set ($n = 61$). Accuracies reflect forced-choice comparisons between high and low and high, medium, and low ratings. Panel C depicts an average peristimulus plot of the PINES response to the holdout test dataset ($n = 61$). This reflects the average PINES response at every repetition time (TR) in the timeseries separated by the rating. Panel D illustrates an item analysis which shows the average PINES response to each photo by the average ratings to the photos in the separate test dataset ($n = 61$). Error bars reflect ±1 standard error.

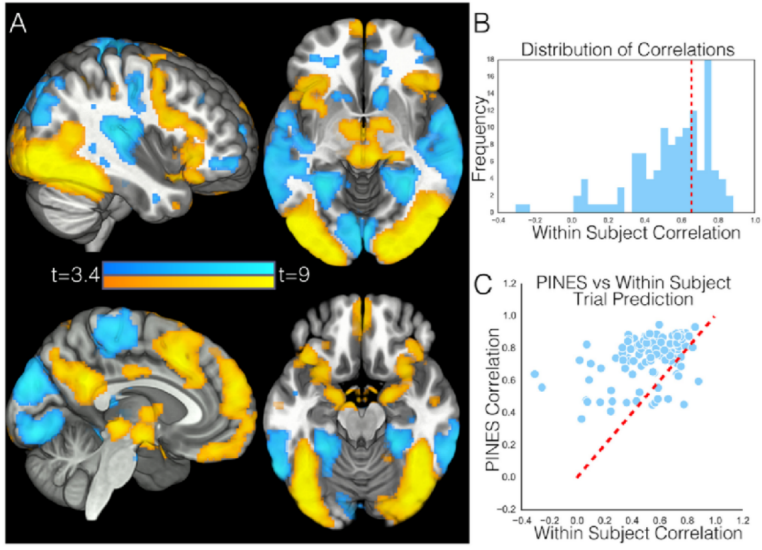


Fig 3. Within participant emotion prediction. This figure depicts results from our within-participant analysis, in which the PINES was retrained separately for each participant to predict ratings to individual photos. Panel A shows the voxels in the weight map that are consistently different from zero across participants using a one sample *t* test thresholded at $p < 0.001$ uncorrected. Panel B shows a histogram of standardized emotion predictions (correlation) for each participant. The dotted red line reflects the average cross validated PINES correlation for predicting each photo's rating. Panel C depicts how well each participant's ratings were predicted by the PINES (y-axis) versus an idiographically trained, cross-validated map using their individual brain data (x-axis). Each point on the graph reflects one participant. The dotted red line reflects the identity line. Any data point above the identity line indicates that the participant was better fit by the PINES than their own weight map.

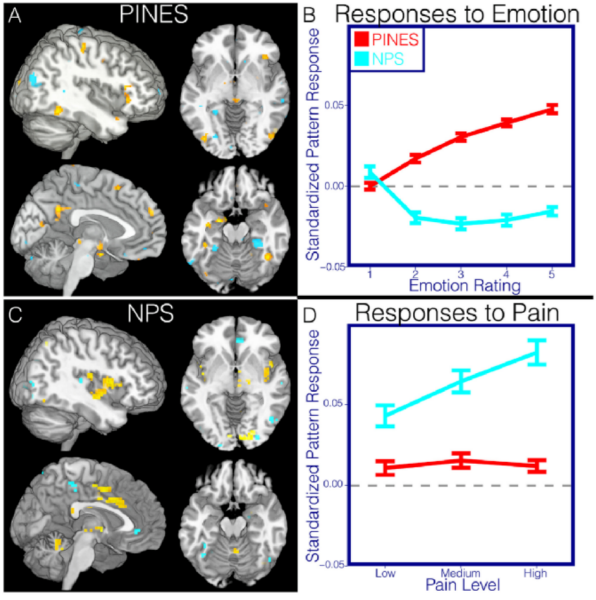


Fig 4. Affective and pain responses to PINES and NPS. This figure illustrates differences in the spatial topography in the thresholded PINES and NPS patterns and their predictions in independent emotion ($n = 61$) and pain ($n = 28$) test data. Panel A depicts the PINES thresholded at $p < 0.001$ uncorrected (see Fig 2). Panel B depicts the average standardized PINES and NPS pattern responses at each level of emotion calculated using a spatial correlation. Error bars reflect ±1 standard error. Panel C depicts the NPS thresholded at false discovery rate (FDR) $q < 0.05$ whole-brain corrected. Panel D depicts the average standardized PINES and NPS pattern responses at each pain level calculated using a spatial correlation. Error bars reflect ±1 standard error.

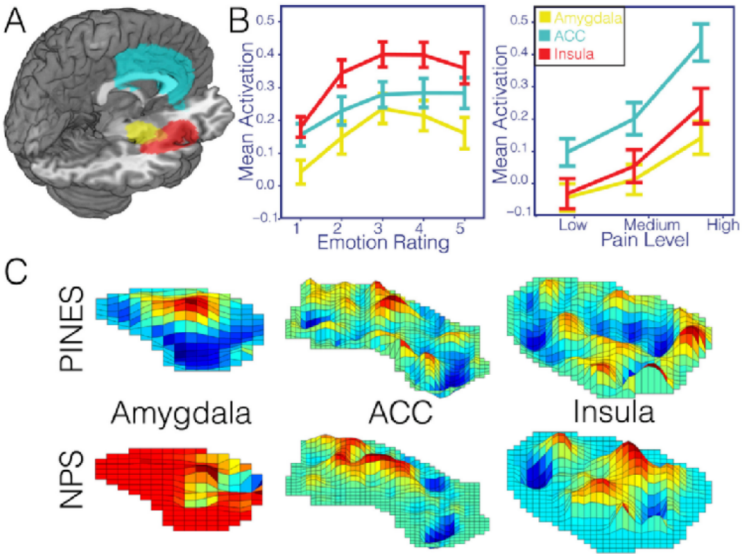


Fig 5. Region of interest analysis. Panel A illustrates the spatial distribution of the three anatomical ROIs used in all analyses (amygdala = yellow, insula = red, ACC = cyan). Panel B depicts the average activation within each ROI across participants for each level of emotion and pain in the emotion hold out ($n = 61$) and pain test datasets ($n = 28$). Error bars reflect ±1 standard error. Panel C illustrates the spatial topography of the PINES and NPS patterns within each of these anatomical ROIs. While these plots show one region, correlations reported in the text reflect bilateral patterns.

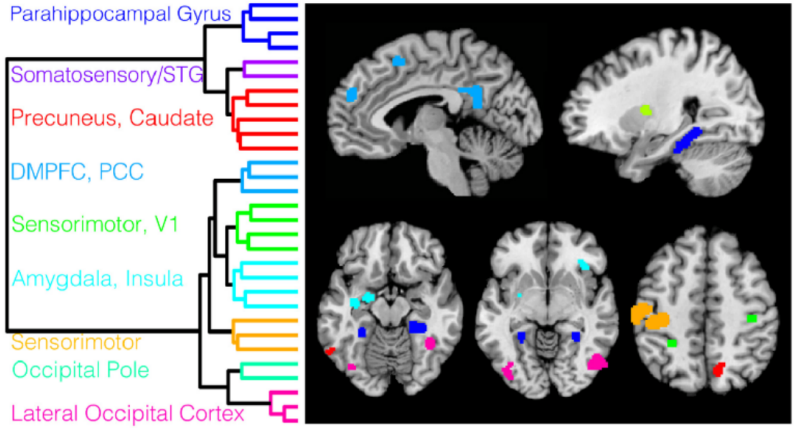


Fig 6. PINES clustering based on shared patterns of connectivity. This figure depicts the results of the hierarchical clustering analysis of the functional connectivity of the largest regions from the $p < 0.001$ thresholded PINES pattern. Clusters were defined by performing hierarchical agglomerative clustering with ward linkage on the trial-by-trial local pattern responses for each region using Euclidean distance. Data were ranked and normalized within each participant and then aggregated by concatenating all 61 subjects' trial x region data matrices. Panel A depicts the dendrogram separated by each functional network. Panel B depicts the spatial distribution of the networks. Colors correspond to the dendrogram labels.

Table 1. Patten sensitivity and specificity.

	Map	Emotion 5 versus 1 (SE)	Pain High versus Low (SE)	Emotion versus Pain (SE)†	Emotion Correlation (SE)	Pain Correlation (SE)
Pattern						
	PINES	93.6 (2.6%)+	60.7 (8%)	93.2 (2.9%)+	0.92 (0.01)	0.64 (0.11)
	Neurologic Pain Signature (NPS)	27.7 (5%)+*	82.1 (5.6%)+	10.7 (3.6%)+*	-0.35 (0.06)	0.91 (0.04)
Average Region of Interest (ROI)						
	Amygdala	55.3 (6%)*	64.3 (8%)	50.5 (5.8%)*	0.31 (0.07)	0.62 (0.09)
	Anterior Cingulate (ACC)	55.3 (5.6%)*	75 (6.7%)+	50.5 (5.8%)*	0.26 (0.07)	0.9 (0.02)
	Insula	55.3 (6%)*	78.6 (6.2%)+	45.6 (5.7%)*	0.32 (0.07)	0.92 (0.02)
Network						
	Visual	50 (6.5%)*	57.1 (8%)	78.6 (4.7%)+*	-0.01 (0.08)	0.22 (0.13)
	Somatomotor	36.2 (6.2%)+*	71.4 (7.1%)+	28.1 (5.2%)+*	-0.38 (0.06)	0.78 (0.09)
	Dorsal Attention	57.4 (6.4%)*	71.4 (6.2%)+	61.2 (5.6%)*	0.34 (0.07)	0.57 (0.12)
	Ventral Attention (Salience)	51.1 (6%)*	71.4 (6.2%)+	13.5 (3.9%)+*	0.14 (0.07)	0.56 (0.13)
	Limbic	57.4 (6%)*	35.7 (8%)	53.4 (5.8%)*	0.28 (0.06)	-0.5 (0.13)
	Frontoparietal	51.1 (5.8%)*	60.7 (7.6%)	42.7 (5.7%)*	0.29 (0.07)	0.34 (0.13)
	Default	63.8 (5.4%)+*	57.1 (7.6%)	70.8 (5.3%)+*	0.34 (0.06)	-0.03 (0.15)

All balanced accuracies reported in this table result from single-interval classification on the test dataset ($n = 47$; see S1 Table for forced-choice test). Analyses involving Level 5 and/or Level 1 comparisons exclude participants that did not rate any stimuli with that label. Accuracy values reflect the ability to discriminate the conditions compared, but are signed, so that values >50% indicate the proportion of participants for which high intensity was classified as greater than low intensity, for high vs. low analyses, or emotion was greater than pain, for Emotion vs. Pain analyses. Values < 50% indicate the proportion of participants for which low intensity was classified as greater than high intensity or pain was classified as greater than emotion. For example, the 10.7% emotion classification of the NPS in the Emotion vs. Pain analysis should be interpreted as a 89.3% hit rate in discriminating pain from emotion. Correlations reflect Pearson correlations between participant's pattern responses to levels of affective intensity and self-reported ratings averaged across participants.

†Please note that this column does not reflect accuracy but rather percent classified as emotion.

+Indicates that accuracy is significantly different from chance (50%), using a two-tailed dependent binomial test.

*Indicates accuracy significantly different from PINES performance using a two-sample two-tailed z-test for proportions (only tested on Emotion 5 versus 1 and Emotion versus Pain columns).

Table 2. Single-cluster and "virtual lesion" analysis.

	Map	nVoxels	Emotion 5 versus 1 (SE)	Pain H versus L (SE)	Emotion versus Pain (SE)	Emotion Correlation (SE)	Pain Correlation (SE)
Pattern							
	PINES	328796	93.5 (2.4%)	60.7 (6.5%)	93.2 (2.9%)	0.92 (0.01)	0.64 (0.11)
	PINES ($p < .001$)	5303	91.5 (3%)$^{+}$	67.9 (7.5%)$^{+}$	97.2 (1.9%)$^{+}$	0.89 (0.01)	0.51 (0.13)
Single Cluster							
	Visual (LOC)	981	83 (4.3%)$^{+*}$	64.3 (7.1%)$^{+}$	85.4 (4.1%)$^{+}$	0.73 (0.03)	0.56 (0.12)
	Somatosensory and superior temporal gyrus (STG)	308	59.6 (5.8%)*	32.1 (7.1%)$^{+}$	61.2 (5.5%)*	0.12 (0.07)	-0.66 (0.11)
	Sensorimotor and V1	335	57.4 (6.2%)*	67.9 (7.5%)$^{+}$	57.3 (5.7%)*	0.23 (0.07)	0.8 (0.07)
	DMPFC and PCC	318	70.2 (5.4%)$^{+*}$	60.7 (7.5%)	70.8 (5.3%)$^{+*}$	0.47 (0.03)	0.61 (0.1)
	Sensorimotor and Cerebellum	1227	78.7 (4.5%)$^{+*}$	60.7 (7.5%)	93.2 (2.9%)$^{+}$	0.72 (0.04)	0.39 (0.14)
	Parahippocampal Gyrus	1025	51.1 (6.4%)*	39.3 (7.1%)	39.9 (5.7%)*	-0.05 (0.07)	-0.43 (0.13)
	Occipital Pole	118	55.3 (6.7%)*	53.6 (8%)	85.4 (4.1%)$^{+}$	0.29 (0.03)	0.22 (0.14)
	Precuneus and Caudate	537	48.9 (6.2%)*	28.6 (7.1%)$^{+}$	53.4 (5.8%)*	-0.15 (0.07)	-0.82 (0.06)
	Amygdala and Insula	454	59.6 (6%)*	75 (6.7%)$^{+}$	54.4 (5.7%)*	0.39 (0.05)	0.76 (0.08)
Virtual Lesion-Cluster Removed	Visual (LOC)	4322	85.1 (4%)$^{+}$	46.4 (8.4%)	96.1 (2.3%)$^{+}$	0.72 (0.05)	-0.17 (0.13)
	Somatosensory and STG	4995	91.5 (3%)$^{+}$	64.3 (8%)	93.2 (2.9%)$^{+}$	0.87 (0.01)	0.67 (0.11)
	Sensorimotor and V1	4968	95.7 (2.1%)$^{+}$	50 (8%)	97.2 (1.9%)$^{+}$	0.9 (0.01)	0.08 (0.15)
	DMPFC and PCC	4985	89.4 (3.4%)$^{+}$	57.1 (8.7%)	97.2 (1.9%)$^{+}$	0.9 (0.01)	0.37 (0.14)
	Sensorimotor and Cerebellum	4076	91.5 (3%)$^{+}$	60.7 (8.4%)	96.1 (2.3%)$^{+}$	0.84 (0.02)	0.56 (0.11)
	Parahippocampal Gyrus	4278	85.1 (4%)$^{+}$	67.9 (7.1%)$^{+}$	96.1 (2.3%)$^{+}$	0.83 (0.02)	0.62 (0.11)
	Occipital Pole	5185	93.6 (2.6%)$^{+}$	64.3 (7.8%)	97.2 (1.9%)$^{+}$	0.89 (0.01)	0.46 (0.14)
	Precuneus and Caudate	4766	89.4(3.4%)$^{+}$	66.1(7.8%)$^{+}$	96.1(2.3%)$^{+}$	0.85(0.02)	0.76(0.07)
	Amygdala and Insula	4849	91.5(3%)$^{+}$	57.1(8.4%)	97.2(1.9%)$^{+}$	0.9(0.01)	0.25(0.15)

All balanced accuracies reported in this table result from single interval classification on the test sample ($n = 47$; see S2 Table for forced-choice test). Analyses involving Level 5 and/or Level 1 comparisons exclude participants that did not rate any stimuli with that label. Accuracy values reflect the ability to discriminate the conditions compared, but are signed so that values >50% indicate the proportion of participants for which high intensity was classified as greater than low intensity for high vs. low analyses, or emotion was greater than pain for Emotion vs. Pain analyses. Values < 50% indicate the proportion of participants for which low intensity was classified as greater than high intensity or pain was classified as greater than emotion. For example, the 10.7% emotion classification of the NPS in the Emotion vs. Pain analysis should be interpreted as a 89.3% hit rate in discriminating pain from emotion. Correlations reflect Pearson correlations between participant's pattern responses to levels of affective intensity and self-reported ratings averaged across participants.

$^{+}$Indicates that accuracy is significantly different from chance (50%) using a two-tailed binomial test.

*Indicates accuracy is significantly different from PINES performance using a two-sample, two-tailed z-test for proportions (only tested on Emotion 5 versus 1 and Emotion versus Pain columns).