Machine learning in neuroimaging: promises and pitfalls

Tal Yarkoni

Department of Psychology, UT-Austin

Machine Learning Summer School 2015
Goals

• Apply some of the concepts/methods you’ve learned
• Identify some gotchas and caveats
• Compare research objectives in ML vs. many sciences
In the beginning…

• I.e., in 1990

• There was a scanner
The Blood-Oxygen-Level-Dependent signal (BOLD)

• At rest, hydrogen protons spin freely

• We impose a strong magnetic field to align the protons, then “pulse” with an RF coil

• Neurons temporarily revert to their original states, releasing energy when they relax back to low-energy state

• The MRI machine picks up this signal

• Strength of signal depends on local tissue properties
  • Influenced by ratio of oxygenated to deoxygenated blood

• When neurons are active in an area, local blood flow increases, and so does the amount of oxygenated blood
The hemodynamic response function

- Neurons are fast
- Blood flow is slow (peaks ~6 seconds after activity)
- Need transfer function relating them
- Canonically modeled as a double-gamma function
- Many variants
- HRF varies!
Basis functions

- True amplitude diff. [A - B]
- True delay diff. [A - B]
- True duration diff. [A - B]

- Canonical HRF
- Canonical HRF + temporal derivative
- Smooth FIR
- Inverse Logit Model

Brain mapping

- Emotion
- Memory
- Language
- Prefrontal cortex
- Claustrum
- LTP
- Pyramidal cells
The classical approach

• How can we map specific cognitive processes onto specific brain regions/networks?

• Subtraction logic + mass univariate analysis
Subtraction logic

- Can isolate cognitive processes via experimental manipulation
- Assumption of pure insertion: we can cleanly add a specific processing step to a given task without affecting anything else
Subtraction logic

Mass univariate analysis

• Treat every point in the brain (“voxel”) as its own universe
• Run same analysis everywhere (~200k times)
• Create nice colorful whole-brain images

Some pretty results

Wojciulik, Kanwisher, & Driver (1998)  
Knutson et al (2001)
Some pretty results

Subsequent memory

Theory of Mind

Wagner et al. (1998)

Gallagher et al. (2000)
Some problems

- Assumption of pure insertion often fails (Friston et al, 1996)
Some problems

• Sensitivity is often very poor
  • Need to correct for thousands of comparisons
Real effects in population
The effects of sampling...
Some problems

• The brain *isn’t* a massively univariate object

• There’s all kinds of structure!
  • Local spatial correlations
  • Long-distance networks
  • May remind you of an earlier lecture…

• Shouldn’t our analytical methods strive to reflect reality?
The rise of machine learning in fMRI

• Beginning around 2000, people started applying concepts and methods from ML to fMRI data
• First powerful demonstration by Haxby et al (2001)
  • Showed that neural responses to visual categories were widely distributed
Haxby et al (2001)
Early classification studies

- Predict which of 10 visual categories subject was looking at
- Feature selection: select category-sensitive voxels
- Several classifiers

Cox & Savoy (2003)
MVPA vs. univariate approaches

Tom, Fox, Trepel, & Poldrack (2007)
Predicting individual differences...

Dosenbach et al (2010)
…or diagnoses

- Hundreds of studies have trained classifiers to discriminate clinical populations from controls
- Often obtain very high sensitivity & specificity (> 80 - 90%)
- E.g., Just et al (2014) reported 97% accuracy classifying autistic vs. control patients using a GNB classifier
A generative model

• Mitchell et al (2008) trained a classifier to predict entirely new nouns
• Initially learn (from text corpus) an intermediate set of semantic vectors
• Predict brain activity from learned vectors
• Apply to unseen nouns
Modeling semantic space

Movies were shown to subjects

Response to each category was found using regularized linear regression

BOLD responses were recorded from the whole brain using fMRI

Category labels

Category model weights

BOLD responses

Huth et al (2012)
category representation in the brain probably has many dimensions. However, given the limitations of fMRI and a finite stimulus set, we expect that we will only be able to recover the first few dimensions of the semantic space for each individual brain and fewer still dimensions that are shared across individuals. Thus, of the 1,705 semantic PCs produced by PCA on the voxel weights, only the first few will resemble the true underlying semantic space, while the remainder will be determined mostly by the statistics of the stimulus set and noise in the fMRI data.

To determine which PCs are significantly different from chance, we compared the semantic PCs to the PCs of the category stimulus matrix (see Experimental Procedures for details of why the stimulus PCs are an appropriate null hypothesis). First, we tested the significance of each subject's own category model weight PCs. If there is a semantic space underlying category representation in the subject's brain, then we should find that some of the subject's model weight PCs explain more of the variance in the subject's category model weights than is explained by the stimulus PCs. However, if there is no semantic space underlying category representation in the subject's brain, then the stimulus PCs should explain the same amount of variance in the category model weights as do the subject's PCs. The results of this analysis are shown in Figure 3. Six to eight PCs from individual subjects explain significantly more variance in category model weights than do the stimulus PCs ($p < 0.001$, bootstrap test). These individual subject PCs explain a total of 30%–35% of the variance in category model weights. Thus, our fMRI data are sufficient to recover semantic spaces for individual subjects that consist of six to eight dimensions.

Second, we used the same procedure to test the significance of group PCs constructed using data combined across subjects. To avoid overfitting, we constructed a separate group semantic space for each subject using combined data from the other four subjects. If the subjects share a common semantic space, then some of the group PCs should explain more of the variance in the selected subject's category model weights than do the stimulus PCs. However, if the subjects do not share a common semantic space, then the stimulus PCs should explain the same amount of variance in the category model weights as do the group PCs. The results of this analysis are also shown in Figure 3. The first four group PCs explain significantly more variance ($p < 0.001$, bootstrap test) than do the stimulus PCs in four out of five subjects.

Figure 2. Category Selectivity for Two Individual Voxels

Each panel shows the predicted response of one voxel to each of the 1,705 categories, organized according to the graphical structure of WordNet. Links indicate ‘‘is a’’ relationships (e.g., an athlete is a person); some relationships used in the model are omitted for clarity. Each marker represents a single noun (circle) or verb (square). Red markers indicate positive predicted responses and blue markers indicate negative predicted responses. The area of each marker indicates predicted response magnitude. The prediction accuracy of each voxel model, computed as the correlation coefficient ($r$) between predicted and actual responses, is shown in the bottom right of each panel along with model significance (see Results for details).

(A) Category selectivity for one voxel located in the left hemisphere parahippocampal place area (PPA). The category model predicts that movies will evoke positive responses when ‘‘structures,’’ ‘‘buildings,’’ ‘‘roads,’’ ‘‘containers,’’ ‘‘devices,’’ and ‘‘vehicles’’ are present. Thus, this voxel appears to be selective for scenes that contain man-made objects and structures (Epstein and Kanwisher, 1998).

(B) Category selectivity for one voxel located in the right hemisphere precuneus (PrCu). The category model predicts that movies will evoke positive responses from this voxel when ‘‘people,’’ ‘‘carnivores,’’ ‘‘communication verbs,’’ ‘‘rooms,’’ or ‘‘vehicles’’ are present and negative responses when movies contain ‘‘atmospheric phenomena,’’ ‘‘locations,’’ ‘‘buildings,’’ or ‘‘roads.’’ Thus, this voxel appears to be selective for scenes that contain people or animals interacting socially (Iacoboni et al., 2004).
To better understand the overall structure of the semantic space, we created an analogous figure in which category position is determined by the PCs instead of the WordNet graph. Figure 5 shows the location of all 1,705 categories in the space formed by the second, third, and fourth group PCs (Movie S1 shows the categories in 3D). Here, categories that are represented similarly in the brain are plotted at nearby positions. Categories that appear near the origin have small PC coefficients and thus are generally weakly represented or are represented similarly across voxels (e.g., ‘‘laptop’’ and ‘‘clothing’’). In contrast, categories that appear far from the origin have large PC coefficients and thus are represented strongly in some voxels and weakly in others (e.g., ‘‘text,’’ ‘‘talk,’’ ‘‘man,’’ ‘‘car,’’ ‘‘animal,’’ and ‘‘underwater’’). These results support earlier findings that categories such as faces (Avidan et al., 2005; Clark et al., 1996; Halgren et al., 1999; Kanwisher et al., 1997; McCarthy et al., 1997; Rajimehr et al., 2009; Tsao et al., 2008) and text (Cohen et al., 2000) are represented strongly and distinctly in the human brain.

Interpretation of the Semantic Space

Earlier studies have suggested that animal categories (including people) are represented distinctly from nonanimal categories (Connolly et al., 2012; Downing et al., 2006; Kriegeskorte et al., 2008; Naselaris et al., 2009). To determine whether hypothesized semantic dimensions such as animal versus nonanimal are captured by the group semantic space, we compared each of the group semantic PCs to nine hypothesized semantic dimensions. For each hypothesized dimension, we first assigned a value to each of the 1,705 categories. For example, for the dimension animal versus nonanimal, we assigned the value +1 to all animal categories and the value 0 to all nonanimal categories. Then we computed how much variance each hypothesized dimension explained in each of the group PCs. If
Semantic Dimensions

1. Mobile vs. Immobile
2. Animacy
3. Human vs. Non-human
4. Social vs. Non-social
5. Civilization vs. Nature
6. Animal vs. Non-animal
7. Biological vs. Non-biological
8. Place vs. Non-place
9. Real-world Size

Semantic Space and Nine Hypothesized Dimensions

PC 1: Mobile vs. Immobile
PC 2: Animacy
PC 3: Human vs. Non-human
PC 4: Social vs. Non-social

Huth et al (2012)
But keep in mind…

• As always, there are limitations

• We want to avoid that whole “have hammer, must seek nail” thing
What questions are we answering?

• Not exactly the same ones we started out with
• Discussion of neuroanatomy often suspiciously missing from ML papers
  • What does a complex pattern of voxel weights mean?
• The famous “black box” in action?
Overfitting, always

• Cross-validation does not magically protect against overfitting

• Can still overfit during model selection, parameterization, etc.

• Think about reported results in light of prior
  • Performance is bounded by the reliability of the outcome

• Is 95%+ classification of psychiatric phenotypes plausible?
  • Can’t possibly outperform human-generated labels
The ADHD-200 competition

• Challenge: predict ADHD status in ~200 public fMRI datasets
  • Data include resting-state fMRI, anatomical scan, demographics, etc.

• Best performance ~55 - 60% (depending on metric)

• Except… the best team was disqualified

• Why?
  • They didn’t use the brain data at all!
  • “For the record, we tried a pile of imaging-based approaches. As a control, we also did classification with age, gender, etc. but no imaging data. It was actually very frustrating for us that none of our imaging-based methods did better than the no imaging results. It does raise some very interesting issues.” (Matthew Brown, personal communication)
Prediction is hard

• Whelan et al (2014): predict binge drinking at age 16 from behavioral, genetic, neuroimaging data at age 14 (n = 692)

• Almost everything is useful to some degree

• But overall performance is surprisingly modest (ROC AUCs ~ .75)

• By far the best predictors are drinking/smoking at age 14
  • Incremental value of other features is small
Correlation doesn’t imply…

- Do pattern classification approaches bring us closer to causal mechanism?
- A classifier will use *any* information it can to make a prediction!

While a pattern recognition approach shows great promise for extracting large amounts of information from fMRI data and for guiding multivariate exploration of representation in the human brain, one must always remain cautious about the nature of the information that a classifier is using to distinguish different classes of stimuli. The fact that information can be extracted by our analysis does not necessarily mean that this information is used by the brain or that the information is used in the way that we think it is. One must always remain conscious of this concern for all analysis techniques that are fundamentally correlational (including traditional univariate fMRI data analysis).

Cox & Savoy (2003)
Gains and losses, or...?

Jimura & Poldrack (2012)
Causal inference in fMRI experiments

• In some cases, predictive approaches actually make it more difficult to draw causal conclusions

• In classical experimental paradigm (manipulation —> brain —> behavior), which links support causal inference?
Can we ever infer causation just from correlation?

• Maybe…

Figure 2: Identifiable ANM with $Y = \tanh(X) + E$, where $X \sim \mathcal{N}(0, 1)$ and $E \sim \mathcal{N}(0, 0.5^2)$. 

Mooij et al. (2014)
But…

Now suppose that we only have data from the observational distribution $\mathbb{P}_{X,Y}$ (for example, because doing intervention experiments is too costly). Can we then still infer the causal relationship between $X$ and $Y$? We will simplify matters by considering only (a) and (b) in Figure 1 as possibilities. In other words, we assume that $X$ and $Y$ are dependent (i.e., $\mathbb{P}_{X,Y} \neq \mathbb{P}_X \mathbb{P}_Y$), there is no confounding (common cause of $X$ and $Y$), no selection bias (common effect of $X$ and $Y$ that is implicitly conditioned on), and no feedback between $X$ and $Y$ (a two-way causal relationship between $X$ and $Y$). Inferring the causal direction between $X$ and $Y$, i.e., deciding which of the two cases (a) and (b) holds, using only the observational distribution $\mathbb{P}_{X,Y}$ is the challenging task that we consider here. If, under certain assumptions, we can decide upon the causal direction, we say that the causal direction is identifiable from the observational distribution.

Mooij et al. (2014)
Meanwhile, in the real world...

- Can we model causal relationships with fMRI?
- Not directly using the BOLD signal!
  - Why?
- So we have to do some deconvolution
  - Causally model deconvolved neuronal responses
  - Dynamic Causal Modeling (Friston, Harrison, & Penny, 2003)
- But... the HRF varies systematically across people, tasks, brain regions, etc.
- Still have standard omitted variables problem
- Is this approach really plausible?

Schlösser et al (2008)
What’s the “right” description?

- The brain/mind is a high-dimensional object
- Is there a single optimal low-dimensional description?
  - Absolutely not
- And yet, almost every method gives good results!
- E.g., structure of brain networks
“Functional Network Organization of the Human Brain”
“A Whole Brain fMRI Atlas Generated via Spatially Constrained Spectral Clustering”

Craddock et al (2011)
“Correspondence of the brain's functional architecture during activation and rest”

Smith et al. (2011)
“The organization of the human cerebral cortex estimated by intrinsic functional connectivity”

Yeo et al. (2011)
How is this possible?

• Why does almost everything we do look so good?

• Because…
  • A huge amount of structure
  • Massive redundancy

• Remember ‘multiclustering’?

• Most high-D real-world datasets have many parsimonious, interpretable low-D descriptions
  • Often reproducible across datasets, contexts

• Doesn’t mean we have the right generative model!
Parsimony is no guarantee

Ryali et al (2010)
Parsimony is no guarantee

• Importance of a feature depends on what else is in the model
• Is it a good idea to select/reduce features prior to estimation?
• Depends…
Interpretation != generation

Figure 6. Comparison between the Group Semantic Space and Nine Hypothesized Semantic Dimensions

For each hypothesized semantic dimension, we assigned a value to each of the 1,705 categories (see Experimental Procedures for details) and we computed the fraction of variance that each dimension explains in each PC. Each panel shows the variance explained by all hypothesized dimensions in one of the four group PCs. Error bars indicate bootstrap SE. The first PC is best explained by a dimension that contrasts mobile categories (people, nonhuman animals, and vehicles) with nonmobile categories and an “animacy” dimension (Connolly et al., 2012) that assigns high weight to humans, decreasing weights to other mammals, birds, reptiles, fish, and invertebrates, and zero weight to other categories. The second PC is best explained by a dimension that contrasts social categories (people and communication verbs) with all other categories. The third PC is best explained by a dimension that contrasts categories associated with civilization (people, man-made objects, and vehicles) with categories associated with nature.

Huth et al. (2012)
Don’t panic!

• These aren’t ML-specific problems
  • They’re reality problems

• Mostly apply to theoretical science
  • Much less of a concern in applied domains—e.g., clinical prediction

• ML approaches can be used instrumentally to address theoretical questions
  • Often much more powerful than conventional approaches
• Is there a detectable neural signature of seen-but-forgotten items?
• I.e., does the brain encode information we can’t readily retrieve when prompted?
The scale of fMRI activity to be the same across scanners (see the Supplementary Appendix). Sensitivity, specificity, positive predictive value, and decision accuracy are all equivalent in the forced-choice test. The MATLAB code for implementing all analyses is available at http://wagerlab.colorado.edu/.

**RESULTS**

**CROSS-VALIDATED PREDICTION OF PAIN**

In study 1, the neurologic signature included significant positive weights in regions including the bilateral dorsal posterior insula, the secondary somatosensory cortex, the anterior insula, the ventrobasal thalamus, and the thalamus. The signature map, consisting of voxels in which activity reliably predicted pain, is shown in Panel A. The map shows weights that exceed a threshold (a false discovery rate of q<0.05) for display only; all weights were used in prediction. ACC denotes anterior cingulate cortex, CB cerebellum, FUS fusiform, HY hypothalamus, IFJ inferior frontal junction, INS insula, MTG middle temporal gyrus, OG occipital gyrus, PAG periaqueductal gray matter, PCC posterior cingulate cortex, PFC prefrontal cortex, S2 secondary somatosensory cortex, SMA supplementary motor area, SMG supramarginal gyrus, SPL superior parietal lobule, TG temporal gyrus, and THAL thalamus. Direction is indicated with preceding lowercase letters as follows: a denotes anterior, d dorsal, i inferior, l lateral, m middle, mid mid-insula, p posterior, and v ventral. Panel B shows reported pain versus cross-validated predicted pain. Each colored line or symbol represents an individual participant. Panel C shows the signature response versus the pain intensity for heat, pain-anticipation, and pain-recall conditions. Signature-response values were calculated by taking the dot product of the signature-pattern weights and parameter estimates from a standard, single-participant general linear model, with regressors for each condition. The estimates shown are derived from cross-validation, so signature weights and test data are independent. Error bars indicate standard errors. The receiver-operating-characteristic plots in Panel D show the tradeoff between specificity and sensitivity. Lines are fitted curves, assuming gaussian signal distributions. The test of pain versus no pain and the forced-choice test are shown by dashed lines and solid lines, respectively. Performance on the forced-choice test was at 100% for all conditions; thus, the lines are overlapping.

Panel A shows the signature map, consisting of voxels in which activity reliably predicted pain. The map shows weights that exceed a threshold (a false discovery rate of q<0.05) for display only; all weights were used in prediction. ACC denotes anterior cingulate cortex, CB cerebellum, FUS fusiform, HY hypothalamus, IFJ inferior frontal junction, INS insula, MTG middle temporal gyrus, OG occipital gyrus, PAG periaqueductal gray matter, PCC posterior cingulate cortex, PFC prefrontal cortex, S2 secondary somatosensory cortex, SMA supplementary motor area, SMG supramarginal gyrus, SPL superior parietal lobule, TG temporal gyrus, and THAL thalamus. Direction is indicated with preceding lowercase letters as follows: a denotes anterior, d dorsal, i inferior, l lateral, m middle, mid mid-insula, p posterior, and v ventral. Panel B shows reported pain versus cross-validated predicted pain. Each colored line or symbol represents an individual participant. Panel C shows the signature response versus the pain intensity for heat, pain-anticipation, and pain-recall conditions. Signature-response values were calculated by taking the dot product of the signature-pattern weights and parameter estimates from a standard, single-participant general linear model, with regressors for each condition. The estimates shown are derived from cross-validation, so signature weights and test data are independent. Error bars indicate standard errors. The receiver-operating-characteristic plots in Panel D show the tradeoff between specificity and sensitivity. Lines are fitted curves, assuming gaussian signal distributions. The test of pain versus no pain and the forced-choice test are shown by dashed lines and solid lines, respectively. Performance on the forced-choice test was at 100% for all conditions; thus, the lines are overlapping.

Wager et al (2013)
Conclusions

• ML approaches contribute to neuroimaging in many ways

• In applied contexts (e.g., predicting diagnoses/treatments), prediction is key

• When understanding is more important than prediction, ML can still help
  • Doesn’t have to be a black box
  • The trick is to construct problem in the right way

• But keep pitfalls in mind, and remember that the goal matters!