Data science to ask questions in mental health

@kordinglab

Shameless plug: Please read *10 simple rules for structuring papers*
Outline

• I) What ML is used for
• II) ML settings, diagnostics and typical uses
• III) Four ways of doing it wrong
• IV) An aside: video based approaches
• V) Causality FTW
I: ML is getting popular in biomedical science

![Graph showing trends in using ML for biomedical sciences over time. The x-axis represents years from 1992 to 2016, and the y-axis represents the number of occurrences. Two lines are shown: one for publications and one for patents. Both lines show an upward trend, with publications consistently higher than patents.](image-url)
Solve real problems

Depression estimates from mobile phones (with Mohr)
Understand data
e.g. Bialek
Provide a benchmark

Being better than another model does not make a model true.

See Jonas and Kording, Could a neuroscientist understand a microprocessor 2017
Model for brain

see Marblestone, Wayne, Kording, 2017
Model for disease

- Solutions
- Fitting
- Bayes
- ...
- Deep learning
II: Two approaches towards diagnostics

• Measure the right thing
  – e.g. identify antibodies, viral RNA etc

• Measure a lot of stuff (ubiquitous)
  – Google searches (e.g. Flu)
  – Locations
  – New media use
  – Accelerations
  – Etc

• And then get at the relevant stuff through machine learning
Workflow

- Produce data where we know the correct diagnostic
- Train a machine learning system
- Test that our machine learning system works
- Use it to make cheaper/better diagnostics
Typical Supervised ML setting

samples → Training → Model → features | label
Typical Supervised ML setting

- **Samples**: Input data for training
- **Features**: Attributes used for training
- **Label**: Output or target variable

**Training**

**Model**

**Predictions**
A typical example: PHQ9 from phone sensors
Phone sensors, truly ubiquitous

- Accelerometer/ Magnetometer/ Barometer
- Brightness sensor
- GPS
- Screen/ Keyboard
- Microphone
Phone use

People use their phones all the time

With Lonini, Jayaraman
GPS data
Extract GPS Features

- Location Variance
- Number of clusters
- Entropy
- Home Stay
- Circadian Movement
- Transition time
Correlated with PHQ9
Combine them with trivial machine learning!

\[ P(\text{Depressive Symptoms}) = g(b_0 + b_1F_1 + b_2F_2 + \ldots + b_nF_n) \]

While looking for small \( b \)
Somewhat can predict mood

<table>
<thead>
<tr>
<th>Training features</th>
<th>Classification (PHQ9&lt;5 vs PHQ9≥5)</th>
<th>PHQ9 score estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% mean accuracy (SD)</td>
<td>% mean sensitivity</td>
</tr>
<tr>
<td>Usage duration</td>
<td>74.2 (3.4)</td>
<td>64.0</td>
</tr>
<tr>
<td>Usage frequency</td>
<td>68.6 (4.1)</td>
<td>56.4</td>
</tr>
<tr>
<td>All</td>
<td>65.7 (4.9)</td>
<td>55.7</td>
</tr>
</tbody>
</table>
Semantic location

[Bar chart showing average relative time for various activities such as Travel Or Transport, Nightlife Spot, Spiritual, Outdoors & Recreation, Arts & Entertainment, Work, Professional Or Medical, Another’s Home, Food, Home, Shop Or Store, comparing depressed and non-depressed groups.]

Saeb... Koring Mohr
How to do good ML

- SVM/SVR
- kNN
- xgBoost
- Random Forest
- GLM
- Stacking!

This is what all the ML courses teach
Use Auto-ML instead

- Approaches are sufficiently standard that this part can easily be automated, e.g. auto-SKlearn, auto-WEKA

- Implication: knowledge about details of ML techniques will become less relevant for biomedical scientists
Result

• AutoML (autosklearn, Freiburg) is almost always better than published results

• AutoML is usually better than our own results

• It is literally three lines of code
Auto-sklearn is good

<table>
<thead>
<tr>
<th>model</th>
<th>features</th>
<th>accuracy</th>
<th>macro f1</th>
<th>weighted f1</th>
</tr>
</thead>
<tbody>
<tr>
<td>majority baseline</td>
<td>N/A</td>
<td>0.5714</td>
<td>0.1818</td>
<td>0.4156</td>
</tr>
<tr>
<td>random forest</td>
<td>age/gender</td>
<td>0.5714</td>
<td>0.1818</td>
<td>0.4156</td>
</tr>
<tr>
<td>random forest</td>
<td>comm</td>
<td>0.6667</td>
<td>0.4795</td>
<td>0.6254</td>
</tr>
<tr>
<td>random forest</td>
<td>comm + age/gender</td>
<td>0.6667</td>
<td>0.4750</td>
<td>0.6225</td>
</tr>
<tr>
<td>random forest</td>
<td>comm + demo + loc</td>
<td>0.6762</td>
<td>0.4744</td>
<td>0.6326</td>
</tr>
<tr>
<td>auto-sklearn</td>
<td>age/gender</td>
<td>0.5714</td>
<td>0.1818</td>
<td>0.4156</td>
</tr>
<tr>
<td>auto-sklearn</td>
<td>comm</td>
<td>0.6571</td>
<td>0.4731</td>
<td>0.6195</td>
</tr>
<tr>
<td>auto-sklearn</td>
<td>comm + age/gender</td>
<td>0.6905</td>
<td>0.5488</td>
<td>0.6654</td>
</tr>
<tr>
<td>auto-sklearn</td>
<td>comm + demo + loc</td>
<td>0.7095</td>
<td>0.5519</td>
<td>0.6806</td>
</tr>
</tbody>
</table>

Relationship prediction, with Lyle Ungar, Tony Liu
Example uses of ML in Neuroscience

Bensmaia, Miller, 2014
Decoding (Neurons->movement)

Motor Cortex

- Wiener Filter
- Kalman Filter
- XGBoost
- Simple RNN
- Wiener Casc.
- SVR
- Feedforward NN
- GRU
- LSTM
- Ensemble

$R^2$ values for different methods:
- Wiener Filter (WF): 0.6
- Wiener Casc. (WC): 0.6
- Kalman Filter (KF): 0.6
- SVR: 0.8
- XGBoost (XGB): 0.9
- Feedforward NN (FNN): 0.9
- RNN: 1.0
- GRU: 1.0
- LSTM: 1.0
- Ensemble: 1.0
Finding generalizes

Somatosensory Cortex

Hippocampus

R²

WF  WC  KF  SVR  XGB  FNN  RNN  GRU  LSTM  Ens

Wiener Filter  Kalman Filt.  XGBoost  Simple RNN  LSTM
Wiener Casc.  SVR  Feedfrwrd NN  GRU

Ens

Wiener Filter  Kalman Filt.  XGBoost  Simple RNN  LSTM
Wiener Casc.  SVR  Feedfrwrd NN  GRU

Ens
Encoding (movements->neurons)

GLM  Feedfrwrd NN  XGBoost  Ensemble

Motor Cortex

pseudo $R^2$

GLM  FNN  XGB  Ens
Finding Generalizes

Rat Hippocampus

Macaque S1

pseudo $R^2$

GLM FNN XGB Ens

GLM FNN XGB Ens
III: The four deadly sins of machine learning

• (1) Wrong question
• (2) Wrong way of assessing quality
• (3) Wrong way of comparing
• (4) Wrong way of managing
Wrong question

- Most ML people are in CS
- Little knowledge about medicine
- Often ask medically irrelevant question
(2) Wrong way of assessing Quality
e.g. bad crossvalidation

With Lonini, Saeb, Mohr, Jayaraman
Cheating works
Massive overconfidence
Literature review

368 records identified through database searching

368 records after duplicates removed

200 records screened

113 records excluded based on title and/or abstract

87 full-text articles assessed for eligibility

25 full-text articles excluded, because:
1) They involved no prediction algorithm (n=4)
2) They used no cross-validation (n=3)
3) They used one record per subject (n=2)
4) They used personal models (n=9)
5) Their cross-validation type was unknown (n=6)
6) Full-text was not accessible (n=1)

62 studies included in quantitative synthesis (meta-analysis):
Subject-wise (n=34)
Record-wise (n=28)
Cheating helps
No one cares
(3) Wrong way of comparing
e.g. personal baselines

• Variance explained
Personal vs group baselines

(A) StudentLife - stress
(B) Friends and Family - happiness
(C) Friends and Family - stress
Machine learning often does not help
## User Lift

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Problem</th>
<th>Model</th>
<th>Avg. Personal Baseline Error</th>
<th>Avg. Personal Model Error</th>
<th>Avg. User Lift (Error)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL—Stress</td>
<td>binary</td>
<td>Log.Reg.</td>
<td>29.19%</td>
<td>29.09%</td>
<td>0.10</td>
<td>.481</td>
</tr>
<tr>
<td>FaF—Happiness</td>
<td>binary</td>
<td>SVM(rbf)</td>
<td>16.51%</td>
<td>18.67%</td>
<td>-2.17</td>
<td>.967</td>
</tr>
<tr>
<td>FaF—Stress</td>
<td>binary</td>
<td>SVM(rbf)</td>
<td>25.17%</td>
<td>23.35%</td>
<td>1.82</td>
<td>.240</td>
</tr>
<tr>
<td>SL—Stress</td>
<td>regression</td>
<td>Elastic Net</td>
<td>0.75</td>
<td>0.78</td>
<td>-0.03</td>
<td>.988</td>
</tr>
<tr>
<td>FaF—Happiness</td>
<td>regression</td>
<td>Elastic Net</td>
<td>0.81</td>
<td>0.83</td>
<td>-0.02</td>
<td>.999</td>
</tr>
<tr>
<td>FaF—Stress</td>
<td>regression</td>
<td>Elastic Net</td>
<td>1.10</td>
<td>1.13</td>
<td>-0.03</td>
<td>1.000</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0184604.t001
Literature review

Google Scholar Search Emotion (n = 517)

Select the first (n = 200)

Matched inclusion criteria

Included (n = 19)

Excluded (n = 181)

Google Scholar Search Stress (n = 1690)

Select the first (n = 200)

Matched inclusion criteria

Included (n = 20)

Excluded (n = 180)
Machine learning often does not help
Does ML even help?
(4) Wrong way of managing

- Get data
- Give half of it to your ML collaborators
- Lock the other half away
- Get their algorithm
- Then test performance on the parts they have not seen
The many ways of leakage

- By not cross validating
- By cross validating wrongly
- By shared recruitment strategy
- By trainee
An app to track Parkinson’s disease

Can the technology behind cell-phone bowling change the lives of Parkinson’s patients?
IV) Towards computer vision-based automated infant neuromotor disorder diagnosis

Dr. Claire Chambers  Rachit Saluja  Wilson Torres  Dr. Laura Prosser  Dr. Michelle Johnson
Neuromotor developmental disorders cause lifelong disability and can be detected early

5 to 10% children are affected by developmental disorders (Rydz et al., 2005)

**Cerebral Palsy:** 2.11 per 1000 live births (Oskoui et al., 2013)
May be higher, 5 per 1000, in lower and middle income countries (Khandaker et al. 2018)

Early detection is crucial so as to maximize brain plasticity during treatment (Palmer, 2004)
Need for a quantified, sensitive and accessible diagnostic

Early diagnosis

Existing clinical methods (General Movements Assessment) have high specificity and are widely tested, but are:
- qualitative
- expensive
- inaccessible in resource-poor environments

Optic flow assessments:
- give only gross movement features
- not clinician interpretable
Approach

A database of ‘normative’ infant movements

Infant movements in a clinical setting

Compare to assess risk
‘Normative’ infant movements from YouTube

YouTube search terms such as:
- one, two, three, four, five, six months old baby
- ______ weeks old

Inclusion criteria:
- infant is non-occluded
- infants move independently
- infant body is present in the video within the frame
- Duration > 6 sec

385 videos found, and 85 included
Collecting infant movement data in a clinical setting

Data collected in the Children’s Hospital of Philadelphia. Approved by ethics board.

Inclusion criteria:
- Infants cannot yet walk
- Absence of history of cardiac, neurological or orthopedic condition
- Parents provide informed consent

GoPro camera used to record movements while in supine position.

Bayley Infant Neurodevelopment Screener (BINS) was used by clinical to assess neuromotor risk. 19 infants assessed. 5 low-risk, 9 moderate-risk, 5 high-risk.
Using computer vision-based pose estimation to extract infant pose

OpenPose (Cao et al., 2018):
- nose, neck, ears, eyes, shoulders, elbows, wrists, hip, knees, and ankles

OpenPose initially provided messy estimates for infants because:
- infant body proportions are different from adults
- Such infant images are missing from the original training dataset (COCO and MPII)
OpenPose domain adaptation

- Keypoints for ~9000 frames were manually labeled using Vatic.
- 8003 frames in the training set and 1036 frames in the test set.
- The test frames are from videos unseen during training.
- Gradient descent for 75 iterations.
- Minimize the error relative to the ground truth manual labels.

Root mean squared error

Epoch
The network worked better on infants after retraining
The network worked better on infants after retraining
The network worked better on infants after retraining.
Cleaning the infant pose raw data

- outlier removal: interpolate and drop points that are greater than two standard deviations (0.1 s bins)
- smoothing using moving average of 1 sec
- camera movements were dealt with by fixing a reference body part (trunk)
- lengths were normalized by trunk length
Old fashioned features

52 features in all

For the positions of the extremities (wrists/ankles) and joint angles (elbows/knees) on both left and right side of the body, we included:

- median position/angle
- IQR of position/angle
- median speed
- IQR of speed
- IQR of acceleration
- mean entropy
- left-right cross correlation
Naive Bayesian surprise metric

- assumes normal distribution and feature independence
- normalized the metric with respect to the ‘normative’ database
- estimate the log probability that a given infant’s movements are drawn from the ‘normative’ distribution

\[
p(x_1, ..., x_n | \mu_{i,H}, \sigma_{i,H}^2)
= \prod_{i=1}^{n} p(x_i | \mu_{i,H}, \sigma_{i,H}^2)
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{i,H}^2}} e^{-\frac{(x_i-\mu_{i,H})^2}{2\sigma_{i,H}^2}}
\]

\[
\Psi = -\ln p = -\sum_{i=1}^{n} \left( \frac{1}{2} \ln(2\pi\sigma_{i,H}^2) + \frac{(x_i - \mu_{i,H})^2}{2\sigma_{i,H}^2} \right)
\]
Predicted risk corresponds to clinician-assessed risk

Chambers, Seethapathi, et al., 2019. Towards accessible computer vision-based diagnosis of infant neuromotor disorders. (in prep.)
V) causality and pseudo experiments

causality
/kəˈzælədē/ ♦

noun
1. the relationship between cause and effect.
2. the principle that everything has a cause.
Definition of causality

• Let a and b be events

• Causation exists if:

  • if we had changed a to a*, the probability for b would have been different
Why causality is hard: Confounding

E.g. Hormone Replacement Therapy, Buying extra insurance
Why causality is hard: Confounding

E.g. Hormone Replacement Therapy, Buying extra insurance
A continuum of confounding

• No confounders: e.g. atari, imagenet
• Few confounders: starcraft
• Countless confounders: Medicine
• $10^{11}$ confounders: brains
• Countless thresholds
• Few controllable variables
• Everything is confounded
• Big datasets
• The ultimate control problem
Simulate a trivial causal system

\[ x_{t+1} = Ax_t + \epsilon \]

Where

\[ \epsilon \sim \mathcal{N}(0, \Sigma) \]

\[ \Sigma = \text{diag}(nL) \]

Choose A: sparse binary (p=.1), largest SV=.99
Delayed Correlation vs Causation

![Graph showing the relationship between R² and the number of variables]

- X-axis: #variables
- Y-axis: R²
Popular solutions

- (1) Randomized perturbations (Experiments)
  - RL exploration \( \sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \)

- (2) ML Bayesian network/ saturated structural equation model
  \[ p(x) = \prod p(x_i | Pa(x_i)) \]

- (3) Model comparisons
Popular solutions

- (1) Randomized perturbations (Experiments)
  - RL exploration \( \sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \)

- (2) ML Bayesian network/ saturated structural equation model
  \[ p(x) = \prod p(x_i \mid Pa(x_i)) \]

- (3) Model comparisons

- Quasiexperiments
Perturbations

Implicit assumptions: we randomly perturb what we care about

Low-D!, expensive, unethical, dangerous
Model comparison

- Have two models with distinct internal causality
- Choose the one that describes data better (p<.05)

\[ AIC = 2k - 2 \ln(\hat{L}) \]
Saturated structural equations + DAGs

\[ L = \sum \log \lambda_\theta(t_{sp}) + \int \lambda_\theta(t)dt \]

Assumptions: causal sufficiency, correct functional form, ...

Paninski, Pillow, Butts, Sahani, ..., yours truly
Pearl/ DAGs

Fig. 2. A diagram representing the back-door criterion; adjusting for variables \( \{X_3, X_4\} \) or \( \{X_4, X_5\} \) yields a consistent estimate of \( \text{pr}(x_j | \bar{x}_i) \).

Fig. 3. A diagram representing the front-door criterion.
Does the world look like this?
Or this?
Potential outcomes

Untreated \( Y_i(0) \)

Treated \( Y_i(1) \)

Rubens, Imbens, Athey
No bias in RCT

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Y_0</th>
<th>Y_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2</td>
<td>3.7</td>
</tr>
<tr>
<td>2</td>
<td>3.5</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6.3</td>
</tr>
<tr>
<td>4</td>
<td>3.4</td>
<td>6.5</td>
</tr>
<tr>
<td>5</td>
<td>4.1</td>
<td>11.1</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>4.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[ TE = E (Y_1 - Y_0) \approx \frac{1}{N} \sum Y_1(i) - Y_0(i) \]
No bias in RCT

\[ TE = E \left( Y_1 - Y_0 \right) \]

\[ \approx \frac{1}{N} \sum_{i} Y_1(i) - Y_0(i) \]

\[ \approx \frac{1}{N_1} \sum_{S_1} Y_1(i) - \frac{1}{N_0} \sum_{S_0} Y_0(i) \]
Quasiexperiments

Idea: find something that is locally kinda random

Marinescu, Lawlor, Kording, Nature Human Behavior, In press
Estimate effect of certificate of merit

Lawlor, Marinescu, Kording, NHB, in press
Does winning merit certificate help?

\[ Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + \hat{\beta}_3 Z_i (X_i - X_c) + e_i \]

Thistlewaite and Campbell 1960
Sanity checks

- Cheating
  - visible as discontinuity in co-variates
- Fuzziness
  - visible as smooth treatment changes
Variance of RDD estimators

- requires ~3 times as many samples as experiment

\[ \text{Var}_{\text{RDD}}(\alpha_0) \propto \frac{3\sigma^2}{n_{\text{bandwidth}}P^2} \]

- how to choose bandwidth? E.g. crossvalidation
Obvious optimization problem (linear)

Reinforcement learning without Exploration

With Marinescu, Triantafillou, forthcoming
Neural data analysis: intracellular recordings
Preplanned RDD

- Often more ethical: e.g. help the poorest districts
  - instead of random
- same in medicine, apply to those who are highest risk
Instrumental variable

With Mikkel Lepperod
Optogenetics is not local

\[ I \approx \frac{1}{d^2} \quad \quad N \approx d^2 \]
Massive confounding
Instrumental variables

Instrument → Treatment → Outcome

Confounder
Civic engagement - college relation

• Distance to nearest college as instrument

• Does it affect $p(\text{register to vote})$?
Example

Distance to college → attend college → Register

World

Groenwold et al
For us
Instrumental variables

$$
\theta^{IV} = \frac{E[C | A_r = 1] - E[C | A_r = 0]}{E[A | A_r = 1] - E[A | A_r = 0]}
$$

Wald estimator (1940)
Instrumental variables
Many neurons
IV helps. A lot.
Why it matters

- Optogenetics is arguably the best causal tool we have
- But crazy hard (2p) to target individual cells
- Use causal inference tricks to cure confounding
An aside

- Medicine has
  - many thresholds
  - many random assignments (e.g. doctors)

- Confounding literally kills
One more pseudoexperiment: Diff in Diff
Caveats

The lure of causal statements: Rampant mis-inference of causality in estimated connectivity

Mehler & Kording

shoutout: Manjari Narayan (@neurostats)
Take home message

• We really mean causality when we talk about mechanism

• In many cases we provide no relevant information re causality

• Perturbations are gold standard. But do not scale

• Quasiexperiments are important set of approximation ideas
Acknowledgements

• ML
  • Ari Benjamin
  • Hugo Fernandes

• Video tracking
  • Claire Chambers
  • Gaiqing Kong
  • Julian Yarkoni
  • Shaofei Wang

• Bad ML
  • Luca Lonini
  • Sohrob Saeb
  • David Mohr
  • Ben Recht
  • Orianna Demasi

• Causality
  • Ioana Marinescu
  • Pat Lawlor
  • Mikkel Løppernd

• Funding
  • NIH, NSF
Stevenson’s Law

N=56

Doubling Time:
7.4 ± 0.4 years
Getting data from brains

- Typing: 100 bits/s record, 20 bits/s meta
- Eye movement: 20 bits/s
- EEG: .5 bits/s
- EMG Hand movement BMI: 2 bits/s

- Dancing? 200 muscles*8 bits/muscle*100/s = 160k bits/s
Take home: Standard ML

- Work really well, fast
- Challenge people to get better results with brain intuitions
- Set baseline

- Ok, let's talk about non-standard now
Machine Learning in Data Driven Medicine: how to not do it wrong

@kordinglab
UPenn

Shameless plug: Please read *10 simple rules for structuring papers*
AFAIK: Most tweeted scientific paper, ever