Machine Learning for Healthcare: Interpretability

David Sontag
Clinical Machine Learning Group
MIT
Many definitions of interpretability

• Papers use the words *interpretable, explainable, intelligible, transparent, and understandable*, both interchangeably (within papers) and inconsistently (across papers)

• One common thread, however, is that interpretability is something other than performance

(Slide credit: Zachary Lipton)
We want **good** models

\[
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_d \\
\end{pmatrix}
\rightarrow \text{Model} \rightarrow y^* \rightarrow \hat{y} \rightarrow \text{Evaluation Metric} \rightarrow \text{Output}
\]
We also want **interpretable** models

The human wants something the metric doesn’t. But, what?

(Slide credit: Zachary Lipton)
Two types of interpretability

• Global interpretability – understand model as a whole
  – Will it work prospectively as intended? (label leakage, label misspecification, dataset shift)
  – What data was most useful? (find more signal of a similar type, form causal hypotheses, figure out how to simplify for deployment purposes)

• Local interpretability – understand predictions for individual data points (i.e., patients)
  – Build trust in predictions; recognize errors due to model being poor, data point being an outlier, or engineering problems
  – Provide guidance to decision makers who may have additional information
Two types of interpretability

• Global interpretability – understand model as a whole
  – Will it work prospectively as intended? (label leakage, label misspecification, dataset shift)
  – What data was most useful? (find more signal of a similar type, form causal hypotheses, figure out how to simplify for deployment purposes)

• Local interpretability – understand predictions for individual data points (i.e., patients)
  – Build trust in predictions; recognize errors due to model being poor, data point being an outlier, or engineering problems
  – Provide guidance to decision makers who may have additional information
Transferability

- The idealized training setups often differ from the real world
  - E.g., label leakage, label misspecification

- Real problem may be non-stationary, noisier (or less noisy!), adversarial, etc.

- Want sanity-checks that the model doesn’t depend on weaknesses in setup

(Slide credit: Zachary Lipton)
Transferability: non-stationary

- Data created during health care is from a non-stationary process due to changes in:
  - Medical science
  - Incentives & regulations
  - Business processes

(Slide credit: Ken Jung)
Transferability: non-stationary

• Testing for covariate shift (wound healing):

Distinguish 2013 from pre-2013

Distinguish first 2/3 of 2013 from last 1/3 of 2013

(Slide credit: Ken Jung)
Transferability: non-stationary

Top 100 lab measurements over time

Time (in months, from 1/2005 up to 1/2014)
**Example 1:**  
Early detection of type 2 Diabetes

For gap=0, we saw that the first-line diabetic medication **Metformin** is predictive

---

**Table 3. Top predictive variables for type 2 diabetes onset within 2009–2010 (gap = 0), using patient records**

<table>
<thead>
<tr>
<th>Variable type</th>
<th>Variable evaluation period</th>
<th>Variable description</th>
<th>Number with diabetes</th>
<th>Number without diabetes</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory test</td>
<td>Past 2 years</td>
<td>Hemoglobin A1c/hemoglobin.total—high (LOINC-4548-4)</td>
<td>1845</td>
<td>8710</td>
<td>9.28 (8.81 9.78)</td>
</tr>
<tr>
<td></td>
<td>Past 2 years</td>
<td>Glucose—high (LOINC-2345-7)</td>
<td>5274</td>
<td>58,736</td>
<td>4.58 (4.43 4.73)</td>
</tr>
<tr>
<td></td>
<td>Past 2 years</td>
<td>Hemoglobin A1c/hemoglobin.total—request for test (LOINC-4548-4)</td>
<td>3908</td>
<td>45,519</td>
<td>4.06 (3.92 4.21)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Cholesterol.in HDL—low (LOINC-2085-9)</td>
<td>3233</td>
<td>49,524</td>
<td>2.94 (2.83 3.06)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Triglyceride—high (LOINC-2571-8)</td>
<td>6056</td>
<td>106,818</td>
<td>2.85 (2.77 2.94)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Cholesterol.total/cholesterol.in HDL—high (LOINC-9830-1)</td>
<td>3114</td>
<td>56,032</td>
<td>2.46 (2.37 2.56)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Alanine aminotransferase—high (LOINC-1742-6)</td>
<td>1208</td>
<td>22,205</td>
<td>2.26 (2.13 2.40)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Cholesterol.in VLDL—request for test (LOINC-13458-5)</td>
<td>3029</td>
<td>63,166</td>
<td>2.09 (2.01 2.18)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Cholesterol.total/cholesterol.in HDL—decreasing (LOINC-9830-1)</td>
<td>3277</td>
<td>75,701</td>
<td>1.89 (1.81 1.96)</td>
</tr>
<tr>
<td></td>
<td>Past 2 years</td>
<td>Carbon dioxide—request for test (LOINC-2028-9)</td>
<td>6044</td>
<td>158,472</td>
<td>1.77 (1.72 1.83)</td>
</tr>
<tr>
<td>ICD9 history</td>
<td>Entire history</td>
<td>Abnormal glucose (ICD9 790.29)</td>
<td>1198</td>
<td>10,099</td>
<td>5.00 (4.70 5.32)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Impaired fasting glucose (ICD9 790.21)</td>
<td>1285</td>
<td>11,521</td>
<td>4.72 (4.45 5.01)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Hypertension (ICD9 401)</td>
<td>12,175</td>
<td>227,759</td>
<td>4.09 (3.97 4.22)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Chronic liver disease (ICD9 571.8)</td>
<td>619</td>
<td>6845</td>
<td>3.71 (3.41 4.03)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Obesity (ICD9 278)</td>
<td>3104</td>
<td>48,000</td>
<td>2.90 (2.78 3.01)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Obstructive sleep apnea (ICD9 327.23)</td>
<td>1178</td>
<td>17,302</td>
<td>2.84 (2.67 3.02)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Hypersomnia with sleep apnea (ICD9 780.53)</td>
<td>1138</td>
<td>16,965</td>
<td>2.79 (2.63 2.97)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Abnormal blood chemistry (ICD9 790.6)</td>
<td>2388</td>
<td>38,726</td>
<td>2.68 (2.56 2.80)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Hyperlipidemia (ICD9 272.4)</td>
<td>8745</td>
<td>186,016</td>
<td>2.62 (2.54 2.69)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Anemia (ICD9 285.9)</td>
<td>3421</td>
<td>75,500</td>
<td>1.99 (1.92 2.07)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Hypothyroidism (ICD9 244.9)</td>
<td>3908</td>
<td>87,228</td>
<td>1.93 (1.86 2.00)</td>
</tr>
<tr>
<td>NDC medication history</td>
<td>Entire history</td>
<td>Acute bronchitis (ICD9 466.0)</td>
<td>3229</td>
<td>93,559</td>
<td>1.46 (1.41 1.52)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Medication group: Metformin</td>
<td>286</td>
<td>1142</td>
<td>10.17 (8.93 11.59)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Medication group: Antiarthritics</td>
<td>3055</td>
<td>88,506</td>
<td>1.46 (1.40 1.51)</td>
</tr>
<tr>
<td></td>
<td>Entire history</td>
<td>Medication group: nonsteroidal anti-inflammatory drugs</td>
<td>3216</td>
<td>94,531</td>
<td>1.44 (1.38 1.49)</td>
</tr>
<tr>
<td>Healthcare utilization</td>
<td>Past 2 years</td>
<td>Procedure group: ophthalmologic and otologic diagnosis and treatment</td>
<td>6681</td>
<td>247,300</td>
<td>1.13 (1.09 1.16)</td>
</tr>
</tbody>
</table>
Example 2: antibiotic resistance

- The moment we included bag of words features derived from clinical notes, we noticed that top predictors were ‘2010’, ‘2009’, ‘2014’, etc.

- We knew there was non-stationarity due to levels of resistance changing, but this was much more than we expected
Example 2: antibiotic resistance

What happened in 2006?

A new card was introduced to MIC testing with a lower range dilutions (more dynamic range)

As a result, cut points to decide difference between resistant/susceptible were moved down

This resulted in many more “positives” for pre-2006 years, but which were simply because these were the lowest possible values that could be recorded

[Figure from Helen Zhou]
Example 3: multiple myeloma

- Four months ago, submitted a paper using the MMRF’s IA9 data release. *Great results*

<table>
<thead>
<tr>
<th>Method</th>
<th>1 Yr Full</th>
<th>1 Yr ISS-FISH</th>
<th>2 Yr Full</th>
<th>2 Yr ISS-FISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.66 ± 0.1</td>
<td>0.62 ± 0.14</td>
<td>0.8 ± 0.08</td>
<td>0.69 ± 0.1</td>
</tr>
<tr>
<td>LR-B-PCA</td>
<td>0.66 ± 0.1</td>
<td>0.61 ± 0.13</td>
<td>0.79 ± 0.08</td>
<td>0.65 ± 0.11</td>
</tr>
<tr>
<td>LR-T-PCA</td>
<td>0.68 ± 0.1</td>
<td>0.61 ± 0.14</td>
<td>0.8 ± 0.08</td>
<td>0.65 ± 0.11</td>
</tr>
<tr>
<td>RF</td>
<td>0.65 ± 0.09</td>
<td>0.63 ± 0.12</td>
<td>0.82 ± 0.08</td>
<td>0.73 ± 0.09</td>
</tr>
<tr>
<td>RF-B-PCA</td>
<td>0.69 ± 0.11</td>
<td>0.63 ± 0.12</td>
<td>0.83 ± 0.08</td>
<td>0.73 ± 0.09</td>
</tr>
<tr>
<td>RF-T-PCA</td>
<td>0.72 ± 0.1</td>
<td>0.64 ± 0.12</td>
<td>0.85 ± 0.08</td>
<td>0.72 ± 0.09</td>
</tr>
</tbody>
</table>
Example 3: multiple myeloma

- Curious to see why "full" feature set so much better, so looked at a decision tree:

- Surprised to see cd319% at the top, but after discussing with clinical collaborator, concluded it is reasonable
Example 3: multiple myeloma

- 1 month ago, new release of data (IA11) is available and I ask students to reproduce results

### Old results (IA9):

<table>
<thead>
<tr>
<th>Method</th>
<th>1 Yr Full</th>
<th>1 Yr ISS-FISH</th>
<th>2 Yr Full</th>
<th>2 Yr ISS-FISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.66 ± 0.1</td>
<td>0.62 ± 0.14</td>
<td>0.8 ± 0.08</td>
<td>0.69 ± 0.1</td>
</tr>
<tr>
<td>LR-B-PCA</td>
<td>0.66 ± 0.1</td>
<td>0.61 ± 0.13</td>
<td>0.79 ± 0.08</td>
<td>0.65 ± 0.11</td>
</tr>
<tr>
<td>LR-T-PCA</td>
<td>0.68 ± 0.1</td>
<td>0.61 ± 0.14</td>
<td>0.8 ± 0.08</td>
<td>0.65 ± 0.11</td>
</tr>
<tr>
<td>RF</td>
<td>0.65 ± 0.09</td>
<td>0.63 ± 0.12</td>
<td>0.82 ± 0.08</td>
<td>0.73 ± 0.09</td>
</tr>
<tr>
<td>RF-B-PCA</td>
<td>0.69 ± 0.11</td>
<td>0.63 ± 0.12</td>
<td>0.83 ± 0.08</td>
<td>0.73 ± 0.09</td>
</tr>
<tr>
<td>RF-T-PCA</td>
<td>0.72 ± 0.1</td>
<td>0.64 ± 0.12</td>
<td>0.85 ± 0.08</td>
<td>0.72 ± 0.09</td>
</tr>
</tbody>
</table>

### New results (IA11):

<table>
<thead>
<tr>
<th>Models</th>
<th>1 Yr Full</th>
<th>1 Yr ISS-FISH</th>
<th>2 Yr Full</th>
<th>2 Yr ISS-FISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.68 ± 0.09</td>
<td>0.65 ± 0.14</td>
<td>0.76 ± 0.08</td>
<td>0.7 ± 0.09</td>
</tr>
<tr>
<td>LR-B-PCA</td>
<td>0.68 ± 0.1</td>
<td>0.65 ± 0.13</td>
<td>0.75 ± 0.08</td>
<td>0.67 ± 0.09</td>
</tr>
<tr>
<td>LR-T-PCA</td>
<td>0.69 ± 0.09</td>
<td>0.64 ± 0.13</td>
<td>0.77 ± 0.07</td>
<td>0.66 ± 0.09</td>
</tr>
<tr>
<td>RF</td>
<td>0.63 ± 0.1</td>
<td>0.63 ± 0.11</td>
<td>0.75 ± 0.08</td>
<td>0.73 ± 0.08</td>
</tr>
<tr>
<td>RF-B-PCA</td>
<td>0.66 ± 0.1</td>
<td>0.64 ± 0.11</td>
<td>0.76 ± 0.08</td>
<td>0.72 ± 0.08</td>
</tr>
<tr>
<td>RF-T-PCA</td>
<td>0.78 ± 0.08</td>
<td>0.64 ± 0.11</td>
<td>0.77 ± 0.08</td>
<td>0.72 ± 0.08</td>
</tr>
</tbody>
</table>

Big differences!
Example 3: multiple myeloma

- 1 month ago, new release of data (IA11) is available and I ask students to reproduce results

- Cd319% no longer shows up as a top predictor!
- What happened!?
Example 3: multiple myeloma

- After digging deeper, we realized that what was predictive originally was the feature Cd319% being missing, and moreover that this was correlated with the outcome (i.e. label leakage!)

[Figure from Rebecca Peyser]
Case study on transferability: Framingham CHD risk score

• Many ML models are trained in one place and deployed more broadly

• **Example:** Framingham coronary heart disease (CHD) risk score
  – Model based on 6 major risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

[Wilson et al., Circulation, 1998]
Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?

- C-statistic (=AUC on censored data) 0.705/0.742 (M/F)
  Pretty good, right out of the box!
- Re-fit using local data only slightly improves C-statistic (=AUC on censored data), to 0.736/0.759 (M/F)

[Liu et al., JAMA ‘04]
Could we say the same about more complex machine learning models?

What would we need to look at to get confidence that they would transfer as well?
Could an adversary fool the classifier?

- Small perturbations of image do not affect visual semantics, but do affect classifications using neural networks

Minimize $\|r\|_2$ subject to:
1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$
Discuss

What are examples where we might expect to see such attacks to ML used in healthcare?
Two types of interpretability

• Global interpretability – understand model as a whole
  – Will it work prospectively as intended? (label leakage, label misspecification, dataset shift)
  – What data was most useful? (find more signal of a similar type, form causal hypotheses, figure out how to simplify for deployment purposes)

• Local interpretability – understand predictions for individual data points (i.e., patients)
  – Build trust in predictions; recognize errors due to model being poor, data point being an outlier, or engineering problems
  – Provide guidance to decision makers who may have additional information
The future will be one of shared decision making

- Short term, humans will share in the reasoning because algorithms not very good
- Long term, humans likely to still be in the loop because managing health involves just as much sociology as it does medicine
Shared decision making: the problem

• Machine doesn't always know what the human knows, and vice-versa
  – Doctor may have access to more data about patient than the ML algorithm
  – Because of some data’s complexity, only the machine will analyze it

• We need:
  1. Natural language understanding to translate what human knows to machine knowledge
  2. Explanation methods to translate what machine knows so human can integrate into decision making
Discuss

What are examples where doctor may know something that the machine doesn’t?
My favorite methods for global interpretability

• Even if eventually will deploy a complex model, use a simpler model initially
• Look at features with large weights in linear model
• Learn a single decision tree and visualize it
Example of using a simpler model initially

- 769 variables have non-zero weight
- No time to look at all 769. Instead, we do a regression with much higher regularization *just for visualization purposes*

**Most predictive diagnosis codes**

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impaired Fasting Glucose</td>
<td>790.21</td>
</tr>
<tr>
<td>Abnormal Glucose NEC</td>
<td>790.29</td>
</tr>
<tr>
<td>Hypertension</td>
<td>401</td>
</tr>
<tr>
<td>Obstructive Sleep Apnea</td>
<td>327.23</td>
</tr>
<tr>
<td>Obesity</td>
<td>278</td>
</tr>
<tr>
<td>Abnormal Blood Chemistry</td>
<td>790.6</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>272.4</td>
</tr>
<tr>
<td><strong>Shortness Of Breath</strong></td>
<td>786.05</td>
</tr>
<tr>
<td><strong>Esophageal Reflux</strong></td>
<td>530.81</td>
</tr>
</tbody>
</table>

**Diabetes**

**1-year gap**
Recent work – neural attention

Motivation

• Complex (neural) models come at the cost of interpretability
• Applications often need interpretable justifications – rationales.

this beer **pours ridiculously clear with tons of carbonation** that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. **this is a real good lookin' beer**, unfortunately it gets worse from here ... first, **the aroma is kind of bubblegum-like and grainy**, next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter.

**Ratings**

*Look:* 5 stars
*Aroma:* 2 stars

(review with rationales)

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Recent work – neural attention

Motivation

• Complex (neural) models come at the cost of interpretability
• Applications often need interpretable justifications — rationally.

There is no evidence of extranodal extension.

BREAST (RIGHT), EXCISIONAL BIOPSY:

**INVASIVE DUCTAL CARCINOMA** (SEE TABLE #1). DUCTAL
CARCINOMA IN-SITU, GRADE 1. ATYPICAL DUCTAL
HYPERPLASIA. LOBULAR NEOPLASIA (ATYPICAL
LOBULAR HYPERPLASIA). TABLE OF PATHOLOGICAL
FINDINGS #1 INVASIVE CARCINOMA
... ...

prediction: high risk of recurring cancer

*Doctors won’t trust machines, unless evidence is provided*

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Recent work – neural attention

Evaluation: Parsing Pathology Report

<table>
<thead>
<tr>
<th>Category</th>
<th>Accession Number</th>
<th>Report Status</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDC</td>
<td>&lt;unk&gt;</td>
<td>Final</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>Surgical Pathology</td>
<td>Pathology Report:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>INVASIVE DUCTAL CARCINOMA poorly differentiated modified Bloom Richardson grade III III measuring at least 0 7cm in this limited specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCIS</td>
<td>Extensive LCIS DCIS Invasive carcinoma of left breast</td>
<td></td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>FINAL DIAGNOSIS BREAST LEFT LOBULAR CARCINOMA IN SITU PRESENT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ADJACENT TO PREVIOUS BIOPSY SITE SEE NOTE CHRONIC INFLAMMATION ORGANIZING HEMORRHAGE AND FAT NECROSIS BIOPSY SITE NOTE There is a second area of focal lobular carcinoma in situ noted with pagetoid spread into ducts No vascular invasion is seen The margins are free of tumor No tumor seen in 14 lymph nodes examined BREAST left breast is a &lt;unk&gt; gram 25 x 28 x 6cm left</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LVI</td>
<td>FINAL DIAGNOSIS BREAST RIGHT EXCISIONAL BIOPSY INVASIVE DUCTAL CARCINOMA DUCTAL CARCINOMA IN SITU SEE TABLE 1 MULTIPLE LEVELS EXAMINED TABLE OF PATHOLOGICAL FINDINGS 1 INVASIVE CARCINOMA Tumor size &lt;unk&gt; X &lt;unk&gt; X 1 3cm Grade 2 Lymphatic vessel invasion Present Blood vessel invasion Not identified Margin of invasive carcinoma Invasive carcinoma extends to less than 0 2cm from the inferior margin of the specimen in one focus Location of ductal carcinoma in situ</td>
<td></td>
<td>84%</td>
</tr>
</tbody>
</table>

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Recent work – LIME: *Local Interpretable Model-Agnostic Explanations*

1. Sample points around $x_i$
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to $x_i$
4. Learn new simple model on weighted samples
5. Use simple model to explain

(Slide credit: Marco Tulio Ribeiro)

[Ribeiro et al., KDD ‘16]
Summary

• Interpretability is essential for using machine learning in health care, particularly to
  – assess whether algorithm will work prospectively as intended
  – provide guidance to decision makers on whether to trust predictions

• For global interpretability, can get quite far by just using a simple model

• This is a very active area of research, particularly for interpretability with deep neural networks