Lecture 10: Interpretability of machine learning models

Prof. David Sontag
MIT EECS, CSAIL, IMES

(Thanks to Zack Lipton for many of the slides)
Outline of today’s class

1. The mythos of model interpretability in health care
2. Learning intelligible models
3. Post-hoc interpretability
What is interpretability?

• Many papers make axiomatic claims that some model is interpretable and therefore preferable

• But what interpretability is and precisely what desiderata it serves are seldom defined

(Slide credit: Zachary Lipton)
Inconsistent definitions

• 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

(Slide credit: Zachary Lipton)
We also want interpretable models

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

(Slide credit: Zachary Lipton)
Trust

• Does the model *know* when it’s uncertain?

• Does the model make same mistakes as human? (e.g., would we be happy delegating decision making authority?)

• Are we *comfortable* with the model?
Trust: can you 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. $x + r \in [0, 1]^m$
Causality

• We may want models to tell us something about the natural world

• Supervised models are trained simply to make predictions, but often used to take actions

• Caruana (2015) shows a mortality predictor (for use in triage) that assigns lower risk to asthma patients

• Naïve interpretations can be misleading

(Slide credit: Zachary Lipton)
Causality: reminder from Lecture 3

- Why one *might* interpret weights learned by linear model causally:
  \[ Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t \]
  \[ \mathbb{E}[\epsilon_t] = 0 \]

\[
\text{ATE} = \mathbb{E}[Y_1(x) - Y_0(x)] = \gamma
\]

- Here we care about \( \gamma \), not about \( Y_t(x) \)
- Identification, not prediction
- Danger: all bets are off with model misspecification
Causality: reminder from Lecture 3

• Suppose true data generating process, \( x \in \mathbb{R} \):
  \[
  Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2
  \]
  \[\text{ATE} = \mathbb{E}[Y_1 - Y_0] = \gamma\]

• Hypothesized linear model (misspecified):
  \[
  \hat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t
  \]

\[
\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt] \mathbb{E}[x^2] - \mathbb{E}[t^2] \mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2] \mathbb{E}[t^2]}
\]

The sign of the weight can flip from negative to positive (and vice-versa)!
Transferability

• The idealized training setups often differ from the real world
  • E.g., data leakage, errors in outcome definition from observational data

• Real problem may be non-stationary, noisier, 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
**Transferability: non-stationary**

- Testing for covariate shift (wound healing):
  - Fit a model to distinguish 2013 vs pre-2013 samples
  - 0.98 AUC on test set

0.00 0.25 0.50 0.75 1.00

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

- Train a model from first two-thirds of 2013 to predict on last third
- 29k train, 14k test (1/3 data)
- AUC of 0.863

0.00 0.25 0.50 0.75 1.00

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

- Distinguish 2013 from pre-2013
- Using just data from 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)
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]
CHD score sheet for men using TC or LDL-C categories.

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

Prediction of coronary heart disease using risk factor categories

Authors: Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, William B Kannel
Publication date: 1998/5/1
Journal: Circulation
Volume: 97
Issue: 18
Pages: 1837-1847
Publisher: Lippincott Williams & Wilkins

Description: The objective of this study was to examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk, to incorporate them into coronary prediction algorithms, and to compare the discrimination properties of this approach with other noncategorical prediction functions. Methods and Results—This work was designed as a prospective, single-center study in the setting of a community-based ...

Total citations: Cited by 8422
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) on Chinese population is 0.705/0.742 (M/F)

- What else should we look at?

[Liu et al., JAMA ‘04]
Case study on transferability: Framingham CHD risk score

**Example:** Framingham coronary heart disease (CHD) risk score (directly applied to Chinese population)

Figure 2. Ten-Year Prediction of CHD Events in CMCS Men and Women Using the Original Framingham Functions

[Figure 2 showing predicted vs. actual CHD events in men and women based on original Framingham functions.]
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)
- 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]
Case study on transferability: Framingham CHD risk score

**Example**: Framingham coronary heart disease (CHD) risk score (re-fit to Chinese population)

<table>
<thead>
<tr>
<th>Risk Factors</th>
<th>CMCS</th>
<th>Framingham*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>Age squared</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Blood pressure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal</td>
<td>-0.51</td>
<td>0.09</td>
</tr>
<tr>
<td>Normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High normal</td>
<td>0.21</td>
<td>0.42</td>
</tr>
<tr>
<td>Stage 1 hypertension</td>
<td>0.33</td>
<td>0.66</td>
</tr>
<tr>
<td>Stage 2-4 hypertension</td>
<td>0.77</td>
<td>0.90</td>
</tr>
<tr>
<td>TC, mg/dL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;160</td>
<td>-0.51</td>
<td>-0.38</td>
</tr>
<tr>
<td>160-199</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200-239</td>
<td>0.07</td>
<td>0.57</td>
</tr>
<tr>
<td>240-279</td>
<td>0.32</td>
<td>0.74</td>
</tr>
<tr>
<td>≥280</td>
<td>0.52</td>
<td>0.83</td>
</tr>
<tr>
<td>HDL-C, mg/dL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;35</td>
<td>-0.25</td>
<td>0.61</td>
</tr>
<tr>
<td>35-44</td>
<td>0.01</td>
<td>0.37</td>
</tr>
<tr>
<td>45-49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>-0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>≥60</td>
<td>-0.40</td>
<td>-0.46</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.09</td>
<td>0.53</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.62</td>
<td>0.73</td>
</tr>
</tbody>
</table>

[©2004 American Medical Association. All rights reserved.](#)
Case study on transferability: Framingham CHD risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (re-fit to Chinese population)

**Figure 1.** Ten-Year Prediction of CHD Events in CMCS Men and Women Using the CMCS Functions

CMCS indicates Chinese Multi-provincial Cohort Study. Coronary heart disease (CHD) events included coronary death and myocardial infarction.
KEY QUESTION TO THINK ABOUT

How robust are your models to changes in the data?
Informativeness

• We may train a model to make a decision

• But it’s real purpose is usually to aid a person in making a decision

• Thus an interpretation may be valuable for the extra bits it carries

  I.e., ability to integrate model output with human prior beliefs

(Slide credit: Zachary Lipton)
DISCUSS

What are examples where informativeness may be important for clinical decision making?
DISCUSS

Where does interpretability show up in your projects?
Outline of today’s class

1. The mythos of model interpretability in health care
2. Learning intelligible models
3. Post-hoc interpretability
Generalized additive models (GAMs)

- GAMs with pairwise interactions have the form:
  
  \[ g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i\neq j} f_{ij}(x_i, x_j) \]

- \( g \) is the link function (e.g. logistic, for binary data), and \( E[f] = 0 \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Pneumonia</th>
<th>Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.8432</td>
<td>0.7523</td>
</tr>
<tr>
<td>GAM</td>
<td>0.8542</td>
<td>0.7795</td>
</tr>
<tr>
<td>GA²M</td>
<td>0.8576</td>
<td>0.7833</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.8460</td>
<td>0.7671</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>0.8493</td>
<td>0.7835</td>
</tr>
</tbody>
</table>

[Caruana et al., KDD ‘15]
Falling rule lists

• Ordered list of if-then rules where:
  1. It is a decision list, i.e. order matters
  2. Probability of outcome decreases monotonically

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Probability</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF IrregularShape AND Age ≥ 60</td>
<td>THEN</td>
<td></td>
</tr>
<tr>
<td>ELSE IF SpiculatedMargin AND Age ≥ 45</td>
<td>THEN</td>
<td></td>
</tr>
<tr>
<td>ELSE IF IllDefinedMargin AND Age ≥ 60</td>
<td>THEN</td>
<td></td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>85.22%</td>
<td>230</td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>78.13%</td>
<td>64</td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>69.23%</td>
<td>39</td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>63.40%</td>
<td>153</td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>39.68%</td>
<td>63</td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>26.09%</td>
<td>46</td>
</tr>
<tr>
<td>ELSE IF Age ≥ 60</td>
<td>10.38%</td>
<td>366</td>
</tr>
<tr>
<td>Methods</td>
<td>Mean AUROC</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>(STD)</td>
<td></td>
</tr>
<tr>
<td>FRL</td>
<td>.80 (.02)</td>
<td></td>
</tr>
<tr>
<td>NF_FRL</td>
<td>.75 (.02)</td>
<td></td>
</tr>
<tr>
<td>NF_GRD</td>
<td>.75 (.02)</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>.79 (.03)</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>.62 (.06)</td>
<td></td>
</tr>
<tr>
<td>Logreg</td>
<td>.82 (.02)</td>
<td></td>
</tr>
<tr>
<td>Cart</td>
<td>.52 (.01)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: AUROC values for readmission data

[Wang & Rudin, AISTATS ‘15]
Supersparse linear integer models

- Learn **linear** model where:
  1. Coefficients are all integer
  2. As sparse as possible

Training objective: \[
\min_{\lambda} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left[ y_i \lambda^T x_i \leq 0 \right] + C_0 \|\lambda\|_0 + \epsilon \|\lambda\|_1
\]
s.t. \( \lambda \in \mathcal{L}. \)

**PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE \( \geq 1 \)**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( \text{age} \geq 60 )</td>
<td>4 points</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. ( \text{hypertension} )</td>
<td>4 points</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3. ( \text{body mass index} \geq 30 )</td>
<td>2 points</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>4. ( \text{body mass index} \geq 40 )</td>
<td>2 points</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>5. ( \text{female} )</td>
<td>-6 points</td>
<td></td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

ADD POINTS FROM ROWS 1 – 5 SCORE = ……

[Ustun & Rudin, ML ‘16]
Neural attention

Motivation

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

Ratings

Look: 5 stars
Aroma: 2 stars

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.

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Neural attention

Motivation

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

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
... ...

description: high risk of recurring cancer

Doctors won’t trust machines, unless evidence is provided

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Neural attention

Model Architecture

input x

Generator \( \text{gen}(x) \)

Encoder \( \text{enc}(z) \)

distribution over possible rationales \( P(z|x) \)

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 looking beer, unfortunately it gets worse from here...

0.8

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 looking beer, unfortunately it gets worse from here...

0.02

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 looking beer, unfortunately it gets worse from here...

0.1

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 looking beer, unfortunately it gets worse from here...

0.05

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 looking beer, unfortunately it gets worse from here...

0.01

\[ \star \star \star \]

generator specifies the distribution of rationales

(Slide credit: Tao Lei) [Lei et al., EMNLP '16]
Neural attention

Model Architecture

Encoder \text{enc}(z) \quad \text{Generator } \text{gen}(x)

distribution over possible rationales \( P(z|x) \)

encoder makes prediction given rationale

(Slide credit: Tao Lei) [Lei et al., EMNLP ’16]
Neural attention

Evaluation: Parsing Pathology Report

Dataset: patients’ pathology reports from hospitals such as MGH

Task: check if a disease/symptom is positive in text
binary classification for each category

Statistics: several thousand report for each category
pathology report is long (>1000 words) but
structured

Model: use CNNs fro $\text{gen}()$ and $\text{enc}()$

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Neural attention

Evaluation: Parsing Pathology Report

<table>
<thead>
<tr>
<th>Category</th>
<th>Accession Number</th>
<th>Report Status</th>
<th>Type</th>
<th>Pathology Report</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDC</td>
<td>&lt;unk&gt;</td>
<td>Final</td>
<td>Surgical Pathology</td>
<td>INVASIVEDUCTALCARCINOMA 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>98%</td>
</tr>
<tr>
<td>LCIS</td>
<td></td>
<td></td>
<td></td>
<td>Extensive LCIS DCIS Invasive carcinoma of left breast FINAL DIAGNOSIS BREAST LEFT LOBULAR CARCINOMA IN SITU PRESENT 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>97%</td>
</tr>
<tr>
<td>LVI</td>
<td></td>
<td></td>
<td></td>
<td>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>84%</td>
</tr>
</tbody>
</table>

(Slide credit: Tao Lei) [Lei et al., EMNLP ‘16]
Outline of today’s class

1. The mythos of model interpretability in health care
2. Learning intelligible models
3. Post-hoc interpretability
Compiling to a simpler model

- **Key idea**: use complex model (e.g. neural network) to train, then compile to a simpler model

[Che et al., arXiv:1512.03542, ‘15]
Compiling to a simpler model

- **Key idea**: use complex model (e.g. neural network) to train, then compile to a simpler model

<table>
<thead>
<tr>
<th>Method</th>
<th>Task</th>
<th>MOR</th>
<th>VFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AUC</td>
<td>AUC(std)</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>0.6431</td>
<td>0.059</td>
</tr>
<tr>
<td>LR</td>
<td></td>
<td>0.6888</td>
<td>0.068</td>
</tr>
<tr>
<td>DT</td>
<td></td>
<td>0.5965</td>
<td>0.081</td>
</tr>
<tr>
<td>GBT</td>
<td></td>
<td>0.7253</td>
<td>0.065</td>
</tr>
<tr>
<td><strong>NN-based</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td><strong>0.7726</strong></td>
<td>0.062</td>
</tr>
<tr>
<td>LR-DNN</td>
<td></td>
<td>0.7300</td>
<td>0.084</td>
</tr>
<tr>
<td>LR-SDA</td>
<td></td>
<td>0.7459</td>
<td>0.068</td>
</tr>
<tr>
<td>LR-LSTM</td>
<td></td>
<td>0.7658</td>
<td>0.063</td>
</tr>
<tr>
<td><strong>Mimic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBTmimic-DNN</td>
<td></td>
<td>0.7574</td>
<td>0.064</td>
</tr>
<tr>
<td>GBTmimic-SDA</td>
<td></td>
<td>0.7382</td>
<td>0.084</td>
</tr>
<tr>
<td>GBTmimic-LSTM</td>
<td></td>
<td><strong>0.7668</strong></td>
<td>0.059</td>
</tr>
<tr>
<td>GBTmimic-LR-DNN</td>
<td></td>
<td>0.7673</td>
<td>0.070</td>
</tr>
<tr>
<td>GBTmimic-LR-SDA</td>
<td></td>
<td>0.7793</td>
<td>0.066</td>
</tr>
<tr>
<td>GBTmimic-LR-LSTM</td>
<td></td>
<td>0.7555</td>
<td>0.067</td>
</tr>
</tbody>
</table>

[Che et al., arXiv:1512.03542, ‘15]
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)