Masterclass: Understanding Machine Learning

EXPERT SPEAKERS

DR. REUBEN BINNS
POSTDOCTORAL RESEARCHER IN COMPUTER SCIENCE, UNIVERSITY OF OXFORD; RESEARCH FELLOW, UK’S INFORMATION COMMISSIONER’S OFFICE

NICHOLAS SCHMIDT
PARTNER, BLDS LLC; HEAD OF THE AI/ML PRACTICE

DR. RICHARD TOMSETT
EMERGING TECHNOLOGY SPECIALIST, IBM
Introduction to Machine Learning

Reuben Binns

University of Oxford / Information Commissioner’s Office.
Reuben.binns@cs.ox.ac.uk @RDBinns

Future of Privacy Forum Masterclass on Machine Learning, CPDP 2020, 23rd January 2020, Brussels
What is AI vs Machine Learning?

Artificial Intelligence
- Symbolic AI (~1950-60s)
- Expert Systems (1980s)
- Search & Planning (~1990s)
- Combinatorial Optimisation
- Statistical Learning Theory
- Regression
- Decision Trees
- Natural Language Processing
- Facial recognition

Machine Learning (~2010s)
- Deep Learning
- Reinforcement Learning
- Statistical Learning Theory

Natural Language Processing
- Semantic Reasoning (~1990-2000s)

Deep Learning
- Strong / General AI: Terminator, HAL 900, etc. ...
What is AI vs Machine Learning

AI not a technology per se, and this is why there is no one definition of AI

AI is collection of computing techniques that can mimic behaviour that we would see as intelligent in humans. E.g. Learning from experience (Machine Learning), interpreting / classifying language (NLP), valid inferences (symbolic AI)

AI = Machine Learning

Machine Learning (ML) is a method of data analysis, which uses learning algorithms to automatically discover patterns in large data sets, generate models, and use them to make predictions. Deep learning is a type of ML.

Algorithms = AI models = AI systems

- An algorithm is a set of rules to be followed by a computer in calculation or other problem-solving operations.
- An ML model is created by feeding training data to the learning algorithm(s) which identify patterns in the data. Once trained, the model can be used on new data to generate predictions.
- An ML system may be formed by multiple models (ensemble) and embedded into software (e.g. a decision-support tool).
Machine learning

- *Supervised:*
  - Predict, classify, score unknown data points, based on models trained on **correctly labelled examples**

- *Unsupervised:*
  - Discover latent structure in data (with not particular purpose in mind)
Supervised learning
Data Model Learning Algorithm Test results
If $P(\text{default}) > \text{threshold}$, then deny credit
What can machine learning do differently?

- Include many more features
- Capture more complicated relationships between features and predictions
More complex relationships

Non-linear
Non-monotonic
Rule-based expert systems vs ML rules

### Data

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>Temperature</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>122</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td>108</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>138</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>121</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>125</td>
</tr>
<tr>
<td>6</td>
<td>59</td>
<td>136</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>144</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>142</td>
</tr>
<tr>
<td>9</td>
<td>67</td>
<td>149</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>161</td>
</tr>
<tr>
<td>11</td>
<td>72</td>
<td>167</td>
</tr>
<tr>
<td>12</td>
<td>76</td>
<td>168</td>
</tr>
<tr>
<td>13</td>
<td>75</td>
<td>162</td>
</tr>
<tr>
<td>14</td>
<td>76</td>
<td>171</td>
</tr>
<tr>
<td>15</td>
<td>76</td>
<td>175</td>
</tr>
<tr>
<td>16</td>
<td>80</td>
<td>182</td>
</tr>
<tr>
<td>17</td>
<td>82</td>
<td>180</td>
</tr>
<tr>
<td>18</td>
<td>83</td>
<td>183</td>
</tr>
<tr>
<td>19</td>
<td>87</td>
<td>188</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>200</td>
</tr>
<tr>
<td>21</td>
<td>93</td>
<td>194</td>
</tr>
<tr>
<td>22</td>
<td>94</td>
<td>296</td>
</tr>
<tr>
<td>23</td>
<td>95</td>
<td>207</td>
</tr>
<tr>
<td>24</td>
<td>97</td>
<td>210</td>
</tr>
<tr>
<td>25</td>
<td>100</td>
<td>219</td>
</tr>
</tbody>
</table>

### Learning Algorithm

### Model
How are data points represented in a machine learning model?
How are data points represented in a machine learning model?
‘Deep learning’

features: { 1,1 = black, 1,2 = brown, 1,3 = grey ...}

hidden layers: {?} 

...labrador?
What kinds of errors? What cost?
False Positive:  
the boy cried wolf... but no wolf

False Negative:  
The villagers thought 'no wolf’ ... but wolf!
$x =$ reoffend
$o =$ not reoffend
$x = \text{reoffend}$

$o = \text{not reoffend}$
\( x = \text{reoffend} \)
\( o = \text{not reoffend} \)

False positive rate = 3/14 = 21%

False negative rate = 6/24 = 25%
$x = \text{reoffend}$

$o = \text{not reoffend}$

False positive rate = $\frac{1}{10} = 10\%$

False negative rate = $\frac{9}{27} = 33\%$
Better that ten guilty persons escape than that one innocent suffer

— Sir William Blackstone (1765)
What tools do you need to build ML?
What tools do you need to build ML?

- Most ML depends on other software for development and deployment
- Frameworks include up to 137 external code dependencies

Outsourcing ML services

- AI-as-a-Service
- Send queries to model via cloud API
- Run model locally
Thanks for learning!
Explainable AI (XAI)

Dr Richard Tomsett

Emerging Technology
IBM Research UK
IBM Emerging Technology

ET Scope

First of a kind instantiation
Technology development
Concept demonstrators
Concept elaboration
Technology research
Fundamental research
Technology inception
Client funded research – DAIS ITA

Stanford

Purdue

UMass Amherst

Penn State

IBM Research

Army Research Laboratory

Imperial College

Cardiff

IBM UK

UCLA

Raytheon BBN

Yale

UCL

BAE Systems

Imperial College

UCL

Government (client)

Academic

Industrial

New DAIS member

Client funded research – DAIS ITA
Explainability

Can the system offer some level of explanation for its outputs?
Regulatory Challenges to AI Adoption

Consumers have the right to access “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”

– Articles 12-15, General Data Protection Regulation (EU)

If a bank denies credit, it is required to provide the applicant with the specific reasons why the applicant was turned down.

– Fair Credit Reporting Act (United States)

In its 2015 decision, the Supreme Court held that plaintiffs need only show that a policy had a discriminatory impact on a protected class, and not that the discrimination was intentional.

– Texas Department of Housing and Community Affairs v. Inclusive Communities Project (United States)
Why do machine learning models require additional explanation?
Statistical Modeling: The Two Cultures
Leo Breiman (2001)

Statistical modelling vs. Machine Learning

Goal: inference about relationship between x and y
Interpretability built-in

Goal: accurately predicting y from x
Interpretability not guaranteed
Some interpretable machine learning methods

- Classical statistical models
- Decision trees
- Rule mining
- Inductive logic programming

...  

if \( (age = 18 - 20) \) and \( (sex = male) \) then predict yes
else if \( (age = 21 - 23) \) and \( (priors = 2 - 3) \) then predict yes
else if \( (priors > 3) \) then predict yes
else predict no

Figure 1: An example rule list that predicts two-year recidivism for the ProPublica data set, found by CORELS.

Learning Certifiably Optimal Rule Lists for Categorical Data
Angelino et al. (2017)
A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems
Hall et al. 2019
But... how can we explain uninterpretable (black box) ML models?
A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems
Hall et al. 2019
Some “post-hoc” explanation methods
“LIME” (Locally Interpretable Model-agnostic Explanations)

Original Image
P(tree frog) = 0.54

Perturbed Instances | P(tree frog)
--- | ---
| | 0.85
| | 0.00001
| | 0.52

Why Should I Trust You? Explaining the Decisions of Any Classifier
Ribeiro et al. 2016

https://towardsdatascience.com/understanding-how-lime-explains-predictions-d404e5d1829c
(a) Husky classified as wolf

(b) Explanation

Why Should I Trust You? Explaining the Decisions of Any Classifier
Ribeiro et al. 2016
Why does the machine think this man isn’t smiling? 
Show a “contrastive” explanation:

Original -
Male: 1
Smile: 0
Young: 0

Explanation -
Male: 1
Smile: 1
Young: 0
Layer-wise relevance propagation

1. forward pass

2. conservative propagation

\[ R_j = \sum_k R_{j \leftarrow k} \]
\[ \sum_j R_{j \leftarrow k} = R_k \]
Spectral Relevance Analysis

Challenges & pitfalls
Developing the sensitivity of LIME for better machine learning explanation
Lee et al. 2019
Layer-wise relevance propagation variants...

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRP-0 [7]</td>
<td>$R_j = \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$</td>
<td>upper layers</td>
</tr>
<tr>
<td>LRP-$\epsilon$ [7]</td>
<td>$R_j = \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$</td>
<td>middle layers</td>
</tr>
<tr>
<td>LRP-$\gamma$</td>
<td>$R_j = \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$</td>
<td>lower layers</td>
</tr>
<tr>
<td>LRP-$\alpha\beta$ [7]</td>
<td>$R_j = \sum_k \left( \alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k$</td>
<td>lower layers</td>
</tr>
<tr>
<td>flat [30]</td>
<td>$R_j = \frac{1}{\sum_i 1} R_k$</td>
<td>lower layers</td>
</tr>
<tr>
<td>$w^2$-rule [36]</td>
<td>$R_j = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$</td>
<td>first layer ($\mathbb{R}^d$)</td>
</tr>
<tr>
<td>$z^B$-rule [36]</td>
<td>$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$</td>
<td>first layer (pixels)</td>
</tr>
</tbody>
</table>
Saliency map explanations for image classification outputs on CIFAR-10 dataset
How do we evaluate explanations?
Area Over Perturbation Curve

$$\text{AOPC} = \frac{1}{L+1} \left\langle \sum_{k=0}^{L} f(x^{(0)}_{\text{MoRF}}) - f(x^{(k)}_{\text{MoRF}}) \right\rangle_{p(x)}$$
Sanity Checks for Saliency Metrics
Tomsett et al. 2020 (to appear)
What makes a good explanation?

It depends on whom the explanation is for
"Interpretable to Whom” framework

Creators

Examiners

Machine learning system

Operators

Executors

Decision-subjects

Data-subjects

Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems
Tomsett et al. 2018
IBM Research Tools

AI Fairness 360
- Detect & mitigate bias for models in development

AI Explainability 360
- Generate explanations for model outputs

Watson OpenScale
- Business Users
- Fully Supported Enterprise Tool
- Measure, Improve and Explain Production AI
Any questions?

Dr Richard Tomsett

rtomsett@uk.ibm.com
@richardtomsett
Ethical Algorithms: Fixing Discriminatory Machine Learning and Biased AI

Nicholas Schmidt
BLDS, LLC & NexusLabs

Future of Privacy Forum
Masterclass: Understanding Machine Learning
CPDP 2020 Data Protection and Artificial Intelligence
January 23, 2020
Introduction

• Nicholas Schmidt
  • AI Practice Leader @ BLDS, LLC
  • CEO @ NexusLabs.ai

• BLDS, LLC
  • Fairness in algorithmic decisioning
  • We *Use AI to Fix AI*
    • Create and implement algorithms that find *fair and predictive models*
Algorithmic decisioning: It is not just math

• The average homeowner in the U.S. has 80 times the net worth of the average renter

• People in debt are 3 times more likely to have mental health problems than those not in debt

• Can AI make this better?
  • Is it well-designed and well-implemented?
  • Is it Fair?
What do I have to say?

- My response to, “We cannot even define fairness, so we cannot do anything about it.”
  - Dangerous
  - Dumb
  - Wrong

- Making AI fairer...
  - It is possible
  - We do not need to sacrifice predictive quality
  - The cost is not too high
Reasons why AI might discriminate

1. Problematic data
   • Underrepresentation
     • facial recognition
   • Inaccurate or missing data
     • credit invisibles
   • Differential patterns of causation and correlation
     • length of credit history
   • Incorporates past, present or potentially future discrimination
     • predictive policing

2. Less discriminatory models are available
Is AI Fairness Possible?

<table>
<thead>
<tr>
<th>The Myth</th>
<th>The Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. We do not know how to define fairness</td>
<td>1. Use the legal framework from the last 55 years</td>
</tr>
<tr>
<td>2. AI is too complicated</td>
<td>2. Use explainable AI</td>
</tr>
<tr>
<td>3. Making predictive AI fairer is not possible</td>
<td>3. BLDS (and many others) have proven this wrong</td>
</tr>
<tr>
<td>4. Finding fairer models takes too much time</td>
<td>4. Use the same computing power that creates AI to fix AI</td>
</tr>
</tbody>
</table>
Fairness Defined

“What you see is what you get”

Group Fairness
- Discrimination → lack of conditional parity across the entire group

Individual Fairness
- Discrimination → lack of conditional parity for similar individuals
- “Disparate Treatment”
Fairness Defined

“We are all equal”

• Discrimination → lack of parity
• “Disparate impact”
Disparate impact

The model predicts equally well for both men and women.
Differential predictions

Credit card spend is underestimated for all women and overestimated for all men with spend between $2,000 and $4,000.
How to balance competing views of fairness

- Remove anything that causes disparate treatment

- Find a model that follows the ideas put forward in the “burden shifting test:”
  
  1. Is the model fair?

  2. Does the model have a valid business justification?

  3. Are there alternative models that are fairer, but maintain reasonable predictive ability?
Fixing Discriminatory AI – Step 1

“A.I. software is only as smart as the data used to train it.”
...and maybe the person who builds it.

• Fixing AI, Step 1: Human Review
  • We are (for now) still smarter than our computers
  • It’s not as onerous as it seems
  • But it does require diversity in views and experience
Fixing Discriminatory AI – Step 2

• Fixing AI, Step 2: Algorithmic Options
  
  • The challenge: Find a fairer model that meets business necessity
  
  • AI makes this easier because of the “Multiplicity of Good Models”
  
  • With many choices, optimize on a second metric - *fairness*
Using AI to Fix AI: the Pareto Frontier
Using AI to Fix AI: the Pareto Frontier in action
Using *AI to Fix AI*: methods

- Feature Selection
- Model Tuning
- Algorithm Selection
- Regularization
- Adversarial Modeling
- Data Preprocessing

See IBM’s “AI Fairness 360” for implementations of many of these methods.
Using AI to Fix AI: feature selection
Using AI to Fix AI: dual-objective optimization

Set Requirements for business validity
(acceptable drop in quality)

Algorithm tests model quality and DI tradeoff at each iteration

Model splits are determined based on minimizing loss while maximizing accuracy
\[ \mathcal{L}(F) = \mathcal{L}(C) - \lambda \cdot \mathcal{L}(DI) \]
Using AI to Fix AI: dual-objective optimization

- Credit Score < 640
  - Model Quality (AUC): 0.65
  - Disparate Impact (AIR): 85%
- Credit Score ≥ 640
  - All Observations
- Credit Score < 660
- Credit Score ≥ 660
  - Model Quality (AUC): 0.70
  - Disparate Impact (AIR): 70%
- Credit Score < 700
- Credit Score ≥ 700
  - Model Quality (AUC): 0.72
  - Disparate Impact (AIR): 60%
Parting thoughts

• AI can be discriminatory – even when the model builder has no ill intent
  • Review data closely – get a diverse perspective
  • Aim towards having a more causal model

• It is possible to fix discriminatory AI

• Making fairer AI can be relatively low-cost

• Please reach out to me if you would like to discuss this more!
Thank You

Nicholas Schmidt
BLDS
LLC
nschmidt@bldsllc.com
https://www.linkedin.com/in/nickpschmidt