Model evaluation in the land of deep learning

May 2, 2018
Host

Pramit Choudhary

@MaverickPramit

www.linkedin.com/in/pramitc/

pramit@datascience.com

- Lead data scientist at DataScience.com.
- Currently, exploring better ways to extract, evaluate, and explain the learned decision policies of predictive models. Recently started an open source project, Skater, to improve the process of model interpretation to enable better model evaluation and model security.
- Before joining DataScience.com, used machine learning algorithms to find love for eHarmony customers.
Agenda

- Understand the problem of model opacity
- Define the “what” and “why” of model interpretation
- Define the scope of model interpretation
- Introduce Skater
- How to interpreting Deep Neural Networks (DNNs) model to improve Model Performance with Skater?
- Demo
- Ability to generate and prevent adversarial attacks with Skater
- Q&A
- References
The Problem of Model Opacity

“By 2018, half of business ethics violations will occur through improper use of big data analytics.” — Gartner

Source: https://www.gartner.com/newsroom/id/3144217

Why I am getting weird predictions? Was my model biased?

I am not 100% sure what’s in the box; I didn’t build the model.
A collection of visual and/or interactive artifacts that provide a user with sufficient description of the model behavior to accurately perform tasks like evaluation, trusting, predicting, or improving the model.

Assistant Professor Sameer Singh, University of California, Irvine
What is Model Interpretation?

- Ability to explain and present a model in a way that is **understandable to humans**.

- A Model’s result is **self descriptive** and **needs no further explanation**; expressed in terms of input and output.
Why is Model Interpretation Important?

“Explain the model.”

Producer / Model Maker:
- Data scientist/analyst building a model
- Consultants helping clients

Consumer / Model Breaker:
- Business owners or data engineers
- Risk/security assessment managers
- Humans being affected by the model
Ideas collapse.
While model interpretation is a hard problem, it's within the role of the data scientist to guide the other stakeholders through different levels of interpretation, recognize the caveats, highlight ambiguities, etc

Paco Nathan,
Director Learning Group, O’Reilly
# Motives for Model Interpretation

<table>
<thead>
<tr>
<th><strong>Producer/Model Maker</strong></th>
<th><strong>Consumer/Model Breaker</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data Scientist</td>
<td>• Data Science Manager</td>
</tr>
<tr>
<td>• Machine Learning Engineer</td>
<td>• Business Owner</td>
</tr>
<tr>
<td>• Data Analyst</td>
<td>• Data Engineer</td>
</tr>
<tr>
<td>• Statistician</td>
<td>• Auditors/Risk Managers</td>
</tr>
<tr>
<td>1. <strong>Debugging and improving</strong> an ML system.</td>
<td>1. Explain the model/algorithm.</td>
</tr>
<tr>
<td>2. <strong>Exploring and discovering latent or hidden feature interactions</strong> (useful for feature engineering/selection and resolving preconceptions).</td>
<td>2. Explain the key features driving the KPI.</td>
</tr>
<tr>
<td>3. Understanding <strong>model variability</strong>.</td>
<td>3. <strong>Verify and validate the accountability</strong> of ML learning systems, e.g., causes for false positives in credit scoring, insurance claim frauds.</td>
</tr>
<tr>
<td>4. Helps in <strong>model comparison</strong>.</td>
<td>4. Identify <strong>blind spots to</strong> prevent adversarial attacks or fix dataset errors.</td>
</tr>
<tr>
<td>5. Building <strong>domain knowledge</strong> about a particular use case.</td>
<td>5. <strong>Ability to share</strong> the explanations to consumers of the predictive model.</td>
</tr>
<tr>
<td>6. Bring <strong>transparency</strong> to decision making to enable <strong>trust</strong>.</td>
<td>6. Comply with <strong>data protection regulations</strong>, e.g., EU’s GDPR.</td>
</tr>
</tbody>
</table>
What Do We Want to Achieve?

With model interpretation, we want to answer the following questions:

- **Why** did the model behave in a certain way?

- **What** was the reason for false positives? What are the relevant variables driving a model’s outcome, e.g., customer lifetime value, fraud detection, image classification, spam detection?

- **How** can we trust the predictions of a “black box” model? Is the predictive model biased? **How** can we guarantee model’s security against adversarial attacks?
Focusing on supervised learning problems.
Scope of Interpretation

Global Interpretation
Being able to explain the conditional interaction between dependent (response) variables and independent (predictor or explanatory) variables based on the complete dataset

Local Interpretation
Being able to explain the conditional interaction between dependent (response) variables and independent (predictor or explanatory) variables with respect to a single prediction
How Do We Enable Model Interpretation?

Visual question answering (VQA): Is this a case of distracted driving?

Relevance: Insurance fraud

*Top predictions include $p($seat-belt$)=0.75$, $p($limousine$)=0.051$, and $p($golf cart$) = 0.017*

Introducing Skater

GitHub: https://github.com/datascienceinc/Skater

Gitter Channel (join us here): https://gitter.im/datascienceinc-skater/Lobby

SKATER

A unified framework to enable Model Interpretation both globally (on the basis of a complete data set) and locally (in-regards to an individual prediction)
Machine Learning Workflow

1. Define Hypothesis
   - Use relevant key performance indicators

2. Handle Data
   - Handle Missing Data
   - Data Partitioning

3. Engineer and Select Features
   - Transform data
   - Select relevant features

4. Build Model
   - Build a predictive model

5. Deploy Model
   - Operationalize analytics as scalable REST APIs

6. Test and Monitor Model
   - 1. Log and track behavior
   - 2. Evaluate
   - 3. Conduct A/B or multi-armed bandit testing

Model Interpretation: In-Memory Models
- Model assessment
- Explain model at a global and local level
- Publish insights, make collaborative and informed decisions

Model Interpretation: Deployed Models
- Explore and explain model behavior
- Debug and discover errors to improve performance

Improve existing hypothesis or generate a new one
An Interpretable Machine Learning System

- humans can verify decisions using post-hoc evaluation
- "Human in the loop" verification & validation

Dataset/Features ➔ ML System ➔ Interpretability with Rule Extraction ➔ SKATER

- validate and verify in-memory model
- validate and verify when deployed
- ability to explain predictions using "Skater"
- predictions

Ability to interpret leads to better hypothesis improvement
Skater’s goal

Interpretability/
Degree of
Opacity

Model Performance (Accuracy)

Enable trust and faithfulness by ability to infer, debug and understand

Note: The purpose of the chart is not to mirror any benchmark on model performance, but to articulate the opacity of predictive models.
Deep Neural Networks

- Helps build expressive and flexible models by learning arbitrary non-linear and non-convex functions easily.

- Can be expressed in different architectures; optimizing for accuracy and computation efficiency often leads to complex designs.

- Need for manual feature engineering is less; lower layers can extract complex features.

- With advancement in software - Keras/Tensorflow/MXNet and hardware-better integration with GPU, it’s easier to train DNN’s over billions of data points optimizing over large number of parameters.

- But, models are often perceived as black boxes because of lack of tools to infer them.
Modern DNNs with complex designs

Optimizing on accuracy, has made Modern DNN architectures complex and often difficult to interpret and understand as humans.

Layer-wise Relevance Propagation (LRP)

- Decomposes the predictions of a deep neural network to pixel level relevance scores using first order approximation

- Initially proposed by Bach S. et. al (2015) [https://doi.org/10.1371/journal.pone.0130140](https://doi.org/10.1371/journal.pone.0130140)

- Computed as a backward pass using a modified gradient from the output layer to the input layer

- Skater supports e-LRP (a version of LRP) as proposed by Ancona M., Ceolini E., Cengiz Ö., Gross M. (2018) in Towards better understanding of gradient-based attribution methods for Deep Neural Networks using chain rule with a modified gradient

\[
    r_i^{(l)} = \sum_j \frac{z_{ji}}{\sum_{i'} (z_{ji'} + b_j) + \epsilon \cdot \text{sign} \left( \sum_{i'} (z_{ji'} + b_j) \right) r_j^{(l+1)}}
\]

\[
x_i \frac{\partial g}{\partial x_i}, \quad g = \frac{f(z)}{z}
\]

- **Scope of Interpretation**: Local Interpretation

- Framework supported by Skater: 🟢 🔄
Integrated Gradient

- Computes relevance score for Deep Networks for Image and Text using first order approximation
- Proposed by Sundararajan, Mukund, Taly, Ankur & Yan, Qibi (2017) in Axiomatic Attribution for Deep Networks
- Implementation adopted as suggested by Ancona M., Ceolini E., Cengiz Ö., Gross M. (2018) in Towards better understanding of gradient-based attribution methods for Deep Neural Networks
- Determines relevance (contribution) of an input $X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$ relative to baseline input $X'$
- Compute the average gradient while the input varies along a linear path from a baseline $x'$ to $x$

$$IG(x) = (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial}{\partial x_i} F(x' + \frac{k}{m} \times (x - x')) \times \frac{1}{m}$$

- Baseline $x'$: for Image: ; for Text: zero embedding vector
- Satisfies sensitivity and implementation Invariance
- **Scope of Interpretation**: Local Interpretation
- Framework supported by Skater: 

Learn more at datascience.com | Empower Your Data Scientists
Demo
Evaluating Deep Neural Networks with Skater

1. CNN on MNIST dataset
2. Imagenet with pre-trained Inception-V3
3. CNN/LSTM sentiment analysis with IMDB dataset
Evaluate Model Stability

Figure: An MNIST experiment with CNN model with 98.8% train and 98.6 test ‘Accuracy’. Interpretation CNN model with ‘ReLU activation using e-LRP. Image-6 on the left is in-correctly classified as 0. Skater provides the ability to infer the cause of mis-classification ( Pixels colored in Red have a positive influence and Blue negative influence ). Images share a semantic properties globally. In the above example we can see 6 and 0 sharing semantic properties around the lower curvy round ‘O’, probably the reason for misclassification.
Identifying blind spots

\[ X \in \mathbb{R}^m, \text{ where } \mathbb{R} \text{ is a image vector, mapped to a discrete label set, } L \in \{1 \ldots k\}. \text{ Interpreting CNN model on MNIST dataset.} \]

\[ X + r \in [0, 1]^m \]

Relevant Image pixels are retained and correctly identified.
Creating Adversarial Examples

Class Labels | Scores
---|---
0 giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca | 0.934644
1 soccer ball | 0.001220
2 space shuttle | 0.000602
Conditional adversarial tested against pre-trained Inception-V3

<table>
<thead>
<tr>
<th>Class Labels</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>giant panda, panda, panda bear, coon bear, Aluropoda melanoleuca</td>
<td>0.034644</td>
</tr>
<tr>
<td>soccer ball</td>
<td>0.001220</td>
</tr>
<tr>
<td>space shuttle</td>
<td>0.000602</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class Labels</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>indri, indris, Indri indri, Indri brevicaudatus</td>
<td>0.266485</td>
</tr>
<tr>
<td>guenon, guenon monkey</td>
<td>0.037744</td>
</tr>
<tr>
<td>colobus, colobus monkey</td>
<td>0.035104</td>
</tr>
</tbody>
</table>
More Examples

- **sports car:** 0.54%
- **grille:** 0.21%
- **washbasin:** 0.43%
- **ping-pong ball:** 0.14%
What about text classification?

For a trained DNN classifier model $F$, let's consider the following:
- Craft an adversarial attack by adding perturbation $\Delta X$
  $$X^* = X + \Delta X$$
- $\Delta X = \text{<Insert, delete, replace>}$
- $F(X) \neq F(X^*)$

Trained Model: CNN for sentiment analysis
Dataset: IMDB

X = excellent tale of two boys that do whatever they can to get away from there abusive father lord of the rings star elijah wood is outstanding in this unforgettable role this movie is one of the main reasons i haven't touched a single beer and never will as long as i live that might make me sound like a nerd but that's what i have to say it is a wonder why this isn't as a classic american tale
Replace [“outstanding”, “excellent”, “classic”, “unforgettable”, “touched”, “never”] with “<UNK>”

<UNK> tale of two boys that do whatever they can to get away from there abusive father lord of the rings star elijah wood is in this <UNK> role this movie is one of the main reasons i haven’t <UNK> a single beer and <UNK> will as long as i live that might make me sound like a nerd but that’s what i have to say it is a wonder why this isn’t as a <UNK> american tale
WITHOUT INTERPRETATION ...

Data → \( f(Y|X) \) → Evaluate

Black box model

WITH SKATER ...

Data → Evaluate

Unboxed model

R or Python model (linear, nonlinear, ensemble, neural networks)
Scikit-learn, caret and rpart packages for CRAN
H20.ai, Algorithmia, etc.

How do I understand my models?

Partial dependence plot
Relative variable importance
Local Interpretable Model Explanation (LIME)
Bayesian rule list (BRL)
DNNs - Integrated Gradient
DNNs - e-Layerwise Relevance Propagation (e-LRP)
Special Thanks

- Marco Ancona, Researcher at ETH Zurich-Department of Computer Science, for helping us in enabling the journey of supporting interpretation for DNNs in Skater

- O’Reilly and AI Conference for allowing us to share our thoughts with all you guys
Future Work and Improvement

- Add better support for generating adversarial examples

- Define and support ways to quantitatively evaluate Model Interpretation

- Implement feature perturbation using Occlusion.

- Add support for other frameworks - PyTorch/MXNet/CNTK


- More experimentation and notebook examples on inferring DNNs
References
