SINTEF

EXPLAINABLE AI

WHY? WHEN? HOW?

Volker Hoffmann, PhD <volker.hoffmann@sintef.no>

NFKF Seminar, Oslo, 09 April 2019

Contents

1. Motivation & Recap ML & What's Our Goal?
2. Interpretable Models vs. Model-Agnostic Approaches
3. SHapley Additive exPlanations (SHAP)
4. Deep Learning & So What?
5. Conclusion
6. (Further Reading)

Motivation

https://old.reddit.com/r/DarkFuturology/comments/b1scj6/indepthlook_at_camera_for_entry_facial/

https://twitter.com/mbrennanchina/status/1109741811310837760

SINTEF

Motivation

SINTEF

Motivation

Articles

European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"

Bryce Goodman, Seth Flaxman

■ *We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.*

In April 2016, for the first time in more than two decades, the European Parliament adopted a set of comprehensive regulations for the collection, storage, and use of personal information, the General Data Protection Regulation (GDPR)[1] (European Union, Parliament and Council 2016). The new regulation has been described as a "Copernican Revolution" in data-protection law, "seeking to shift its focus away from paper-based, bureaucratic requirements and towards compliance in practice, harmonization of the law, and individual empowerment" (Kuner 2012). Much in the regulations is clearly aimed at perceived gaps and inconsistencies in the European Union's (EU) current approach to data protection. This includes, for example, the codification of the "right to be forgotten" (Article 17), and

 Copyright © 2017, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

Articles

… for the first time in more than two decades, the European Parliament adopted a set of comprehensive regulations for the collection, storage, and use of personal information

© pepifoto, iStock

regulations for foreign companies collecting data from European citizens (Article 44).

However, while the bulk of language deals with how data is collected and stored, the regulation contains Article 22: Automated individual decision making, including profiling (see figure 1) potentially prohibiting a wide swath of algorithms currently in use in recommendation systems, credit and insurance risk assessments, computational advertising, and social networks, for example. This prohibition raises important issues that are of particular concern to the machine-learning community. In its current form, the GDPR's requirements could require a complete overhaul of standard and widely used algorithmic techniques. The GDPR's policy on the right of citizens to receive an explanation for algorithmic decisions highlights the pressing importance of human interpretability in algorithm design. If, as expected, the GDPR takes effect in its current form in mid-2018, there will be a pressing need for effective algorithms that can operate within this new legal framework.

Background

The General Data Protection Regulation is slated to go into effect in April 2018 and will replace the EU's 1995 Data Protection Directive (DPD). On the surface, the GDPR merely reaffirms the DPD's right to explanation and restrictions on automated decision making. However, this reading ignores a number of critical differences between the two pieces of legislation (Goodman 2016a, 2016b).

First, it is important to note the difference between a *directive* and a *regulation*. While a directive "set[s] out general rules to be transferred into national law by each country as they deem appropriate," a regulation is "similar to a national law with the difference that it is applicable in all EU countries" (European Documentation Centre 2016). In other words, the 1995 directive was subject to national interpretation and was only ever indirectly implemented through subsequent laws passed within individual member states (Fromholz 2000). The GDPR, however, requires no enabling legislation to take effect. It does not direct

"[…] When put into practice, the law may also effectively ***create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision*** *that significantly affects them. […]"*

https://doi.org/10.1609/aimag.v38i3.2741

SINTEF

Recap of AI/ML

TRAIN ON THIS

PREDICT THIS

Income 2017	ZIP Code	#/Kids	Loan Decision
650’000	0123	3	Accept
750’000	0123	1	Accept
850’000	1231	2	Reject
…			
…			
710’000	0123	1	**(PREDICT THIS)**
1’110’000	7123	2	**(PREDICT THIS)**

FEATURES

LABELS

Goals?

- Explain (relative) importance of features
- Explain individual predictions

Interpretable Models vs. Agnostic Approaches

- Blabla on interpretable models (lin/log reg, trees, etc)
- Blabla on agnostic approaches (entropy, shuffling, etc)

Let’s Poke

Feature 1	Feature 2	Prediction
101	5.34	0.5
104	4.14	0.7
105	7.24	0.5

Feature 1	Feature 2	Prediction
301	5.34	0.9
304	4.14	1.0
505	7.24	0.8

Explaining Single Predictions

BASE VALUE (0.5 FOR 50/50 DATASET)

FINAL PREDICTION

SINTEF

Explaining Overall Predictions

SINTEF

Deep Learning?

https://github.com/slundberg/shap

Deep Reinforcement Learning?

Silver et al. 2016 (AlphaGo)
Silver et al. 2017 (AlphaGoZero)

SINTEF

So What?

PERSPECTIVE | ARTIFICIAL INTELLIGENCE

In defense of the black box

Elizabeth A. Holm

+ See all authors and affiliations

Science 05 Apr 2019:
Vol. 364, Issue 6435, pp. 26-27
DOI: 10.1126/science.aax0162

Article | Figures & Data | Info & Metrics | eLetters | PDF

Download PDF

INSIGHTS | *PERSPECTIVES*

tions ongoing in their cores, which can last for hundreds of millions to many billions of years, depending on their mass. When such a star exhausts the hydrogen fuel in its core, it expands enormously, shatters any close-enough planet, and becomes a white dwarf. Thereafter, what remains of the planetary system may move close enough to the star to become subject to collisions and to strong tidal forces, grinding the remaining planetary cores (*3*). This leaves behind a shroud of rocky debris of various sizes, ranging from micrometer-sized dust parti-

existence of numerous polluted white dwarfs indicates that planetesimals indeed orbit around these stars. However, planetesimals orbiting white dwarfs have been directly found in just one case using the Kepler space telescope and the transit method (*12*), despite the large number of polluted white dwarfs discovered to date, the fact that white dwarfs are the descendant of almost all planet hosts known to date, and that their small size facilitates the detection of transiting bodies.

The method of Manser *et al.* has revealed the presence of planetesimals without the

ARTIFICIAL INTELLIGENCE

In defense of the black box

Black box algorithms can be useful in science and engineering

Science

Vol 364, Issue 6435
05 April 2019

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

ARTICLE TOOLS

Email | Download Powerpoint
Print | Save to my folders
Alerts | Request Permissions
Citation tools | Share

SIMILAR ARTICLES IN:

- PubMed
- Google Scholar

NAVIGATE THIS ARTICLE

Holm (2019) // http://science.sciencemag.org/content/364/6435/26

SINTEF

So What?

PERSPECTIVE | ARTIFICIAL INTELLIGENCE

In defense of the black box

Elizabeth A. Holm
+ See all authors and affiliations

Science 05 Apr 2019:
Vol. 364, Issue 6435, pp. 26-27
DOI: 10.1126/science.aax0162

Article | Figures & Data | Info & Metrics | eLette

Download PDF

INSIGHTS | *PERSPECTIVES*

tions ongoing in their cores, which can last for hundreds of millions to many billions of years, depending on their mass. When such a star exhausts the hydrogen fuel in its core, it expands enormously, shatters any close-enough planet, and becomes a white dwarf. Thereafter, what remains of the planetary system may move close enough to the star to become subject to collisions and to strong tidal forces, grinding the remaining planetary cores (*3*). This leaves behind a shroud of rocky debris of various sizes, ranging from micrometer-sized dust parti-

existence of numerous polluted white dwarfs indicates that planetesimals indeed orbit around these stars. However, planetesimals orbiting white dwarfs have been directly found in just one case using the Kepler space telescope and the transit method (*12*), despite the large number of polluted white dwarfs discovered to date, the fact that white dwarfs are the descendant of almost all planet hosts known to date, and that their small size facilitates the detection of transiting bodies.

The method of Manser *et al.* has revealed the presence of planetesimals without the

ARTIFICIAL INTEL

In defe
the bla

Black box algo
can be useful i
and engineeri

Pedro Domingos
@pmddomingos

Follow

Three good reasons to use black-box AI systems:

In defense of the black box
The science fiction writer Douglas Adams imagined the greatest computer ever built, Deep Thought, programmed to answer the deepest question ever asked: the Great Question of Life, the U...
science.sciencemag.org

6:18 PM - 5 Apr 2019

14 | 198 | 570

Tweet your reply

Zhipeng Wang @PKUWZP · Apr 6
Replying to @pmddomingos
I think the question remains tricky: if the model works well, it might be OK; but if the model does not work well, how we can understand where the problem is and fix it accordingly? In that sense being black box is a dangerous scenario.

2 | 13

1 more reply

Grady Booch @Grady_Booch · Apr 6
Replying to @pmddomingos

Holm (2019) // http://science.sciencemag.org/content/364/6435/26

https://twitter.com/pmddomingos/status/1114200712471293952

Conclusion

- You want to understand your own models (biases!)
- Customers may have a “Right to Explain” (muddy waters, ask your lawyer)

- Many approaches XAI, SHAP values are one
- Output: “Probability of 90% assigned because X_1 > 0.1 and X_2 < 0.7”

- Works for some Deep Learning, tricky for others
- Sometimes, maybe you don’t care

(Further Reading)

- SHAP Paper w/ Reviews:
  - http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
  - http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips30/reviews/2493.html
- SHAP Python Package / Cool Notebook for High-Order Interactions:
  - https://github.com/slundberg/shap
  - https://slundberg.github.io/shap/notebooks/NHANES%20I%20Survival%20Model.html
- Interpretable ML Book
  - https://christophm.github.io/interpretable-ml-book/
- Kaggle Tutorial
  - https://www.kaggle.com/ml-for-insights-signup

SINTEF

Teknologi for et bedre samfunn