Right, but Why?

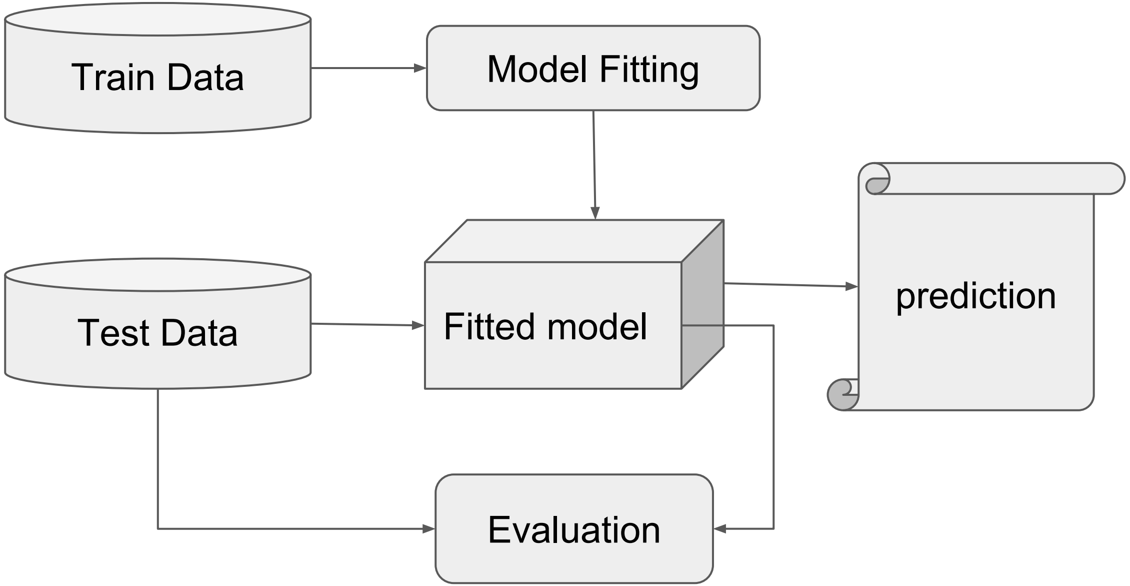
Explaining a Model Decision

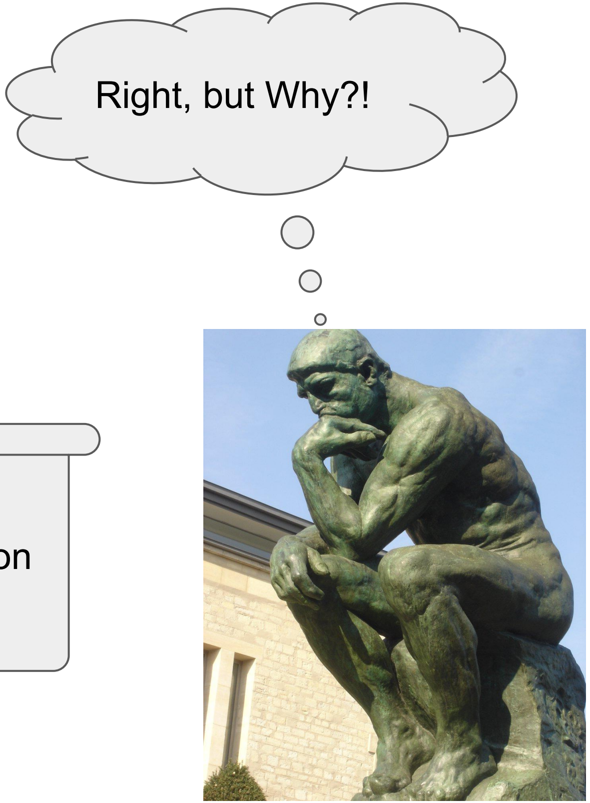
Data Science Summit 2018

Written by Dr. Hanan Shteingart

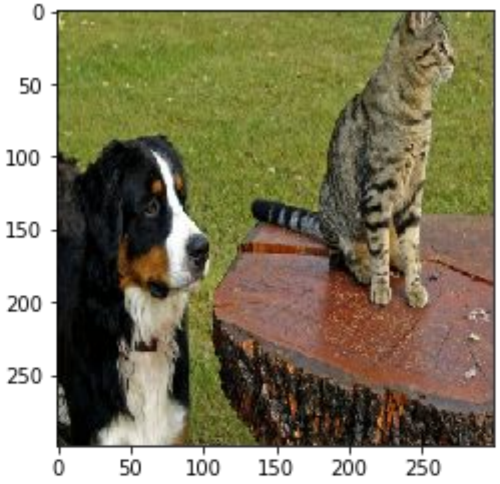
Why?

- Learn how to interpret a model's prediction





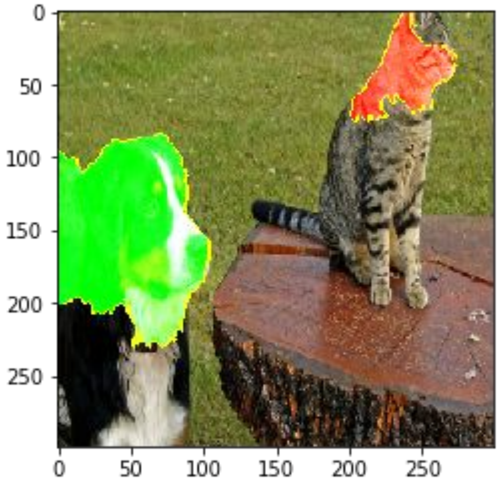
Why? Example



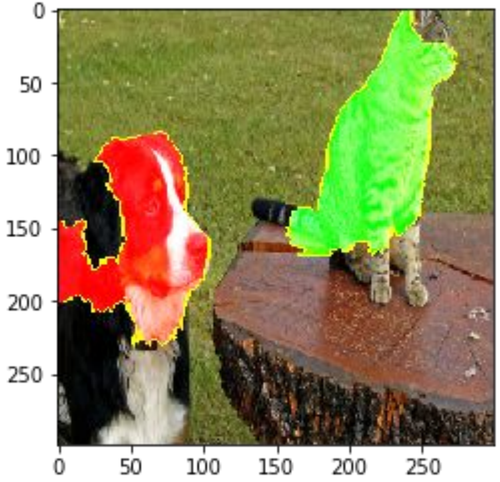
Prediction:

- Bernese mountain dog 0.83
- Egyptian cat 0.0009

Explain dog:



Explain cat:

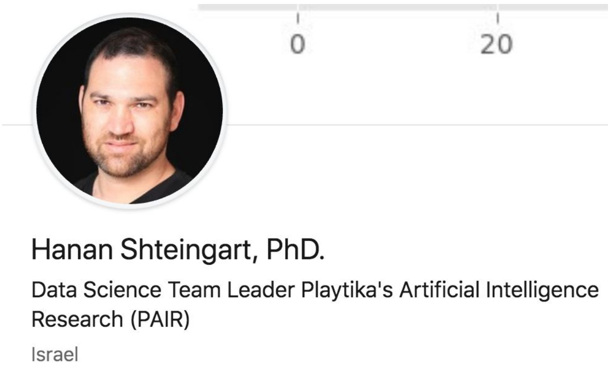


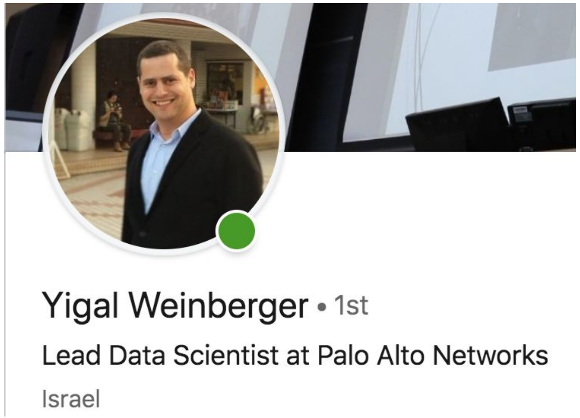
How?

- Quick on Theory
- Example Driven
- Hands-on Exercises
- Duration: 3 hours
  - Two sessions, 1:30 each

What? Plan for Today

- Session 1: **Introduction & Interpretable Models** by Hanan Shteingart, PhD.
  - Introduction and motivation
  - Linear models
  - Naive Bayes
  - Tree Ensembles
- Session 2: **Black Box Approach using LIME** by Yigal Weinberger





3 commercial slides…

from the workshop's sponsors

PLAYTIKA OVERVIEW

Founded in
2010

1800+
Employees

Playtika's AI Research Lab - Problems and Scale

- Problem Spaces
  - Reinforcement Learning
  - User Behavior Modeling
  - Ad-Tech
  - Optimization
  - Recommendation Systems
- Scale
  - Rate of 3.5 Terra byte a day
  - Daily up to Real-time solutions

Goal
collaborate a cross-company while aiming at gaining a competitive marketplace advantage and reaching better business results

We Recruit Talents https://www.playtika.com/careers

Talented ML/DS/SW send your cv to hanans@playtika.com

Introduction

Model Interpretability

- Two popular notions of interpretability:
  - **Understandability** - grasp how the models work.
  - **Post-hoc Interpretations** - explain predictions without elucidating the mechanisms by which models work
- The <u>latter</u> is the focus of this workshop

Motivation

- When objectives are difficult to encode in ML framework

"**Trust**" - subjective judgment of the model, e.g. racial bias

"**Informativeness**" - the supervised model is used instead to provide information to human decision makers

"**Causality**" - hope of inferring properties or generating hypotheses about the natural world.

"**Transferability**" - can we use this model outside of its comfort zone?

"**Fair and Ethical Decision-Making**" - purpose of assessing whether decisions provided automatically by algorithms conform to ethical standards

Naive Bayes

Goto naive_bayes notebook

Tree Ensemble

Goto random_forest notebook

Linear Model

Goto linear_model notebook

Deep Learning Note

- One popular approach for deep neural nets is to compute a saliency map.
- Typically, take the gradient of the output corresponding to the correct class with respect to a given input vector.

Class activation maps for one object class

http://cnnlocalization.csail.mit.edu/

References

1. Zachary C. Lipton. “The mythos of model interpretability.” arXiv preprint arXiv:1606.03490 (2016).
2. Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. “Examples are not enough, learn to criticize! criticism for interpretability.” Advances in Neural Information Processing Systems. 2016.
3. Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. “Interpretable decision sets: A joint framework for description and prediction.” Proc. 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM, 2016.
4. Robnik-Šikonja, Marko, and Igor Kononenko. “Explaining classifications for individual instances.” IEEE Transactions on Knowledge and Data Engineering 20.5 (2008): 589-600.
5. Baehrens, David, et al. “How to explain individual classification decisions.” Journal of Machine Learning Research 11.Jun (2010): 1803-1831.
6. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?: Explaining the predictions of any classifier.” Proc. 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM, 2016.
7. Yiming Yang, and Jan O. Pedersen. “A comparative study on feature selection in text categorization.”Intl. Conf. on Machine Learning. Vol. 97. 1997.
8. Isabelle Guyon, and André Elisseeff. “An introduction to variable and feature selection.” Journal of Machine Learning Research 3.Mar (2003): 1157-1182.
9. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples.” ICLR 2015.
10. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. “Intriguing properties of neural networks.” Intl. Conf. on Learning Representations (2014)
11. Nguyen, Anh, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.” Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2015.
12. Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model.” The Annals of Applied Statistics 9, No. 3 (2015): 1350-1371