AI System Evaluation

Week 10: AI Interpretability

Aug 23 - Week 1: 7-10	Introduction	
Aug 30 - Week 2: 7-10	AI Robustness	Exercise 1
Sep 06 - Week 3: 7-10	Improving AI Robustness	Exercise 2
Sep 13 - Week 4: 7-10	AI Backdoors	Exercise 3
Sep 20 - Week 5: 7-10	Mitigating AI Backdoors	Exercise 4; Project Proposal
Sep 27 - Week 6: 7-10	AI Fairness	Exercise 5
Oct 11 - Week 7: 7-10	Improving AI Fairness	Exercise 6
Oct 18 - Week 8: 7-10	AI Privacy	Exercise 7
Oct 25 - Week 9: 7-10	Improving AI Privacy	Exercise 8
Nov 01 - Week 10: 7-10	AI Interpretability	Project Due
Nov 08 - Week 11: 1-3	End-of-Term Exam	

Outline

When/why do we need interpretability

Ways of achieving interpretability

Do We Need Interpretability?

Arguments “FOR” Interpretability

“*We need to be able to understand how neural network makes decisions before we can deploy them.*” (Anonymous)

Not so much

“*As soon as you have a complicated enough machine, it becomes almost impossible to completely explain what it does.*" (Yoshua Bengio, 2016)

“*We build amazing models. But we don't quite understand them. Every year this gap is going to get a little bit larger.*" (Paul Voosen, 2017)

We don’t actually need to understand a fridge before using it.

Do We Need Interpretability?

Arguments "FOR" Interpretability

"*The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.*" (GDPR)

Not so much

"*Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.*" (Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 2017)

Reasons for Interpretability

Practical Reasons

Trust: How do we trust it if we can’t understand it?

Fair and Ethical Decision Making: How do we know it’s fair if we don’t understand it?

Accountability: How do we know who to blame if we don’t know the decision is made?

Domains demand interpretability

Medical Domain/Health-Care: Why is the diagnosis cancer?

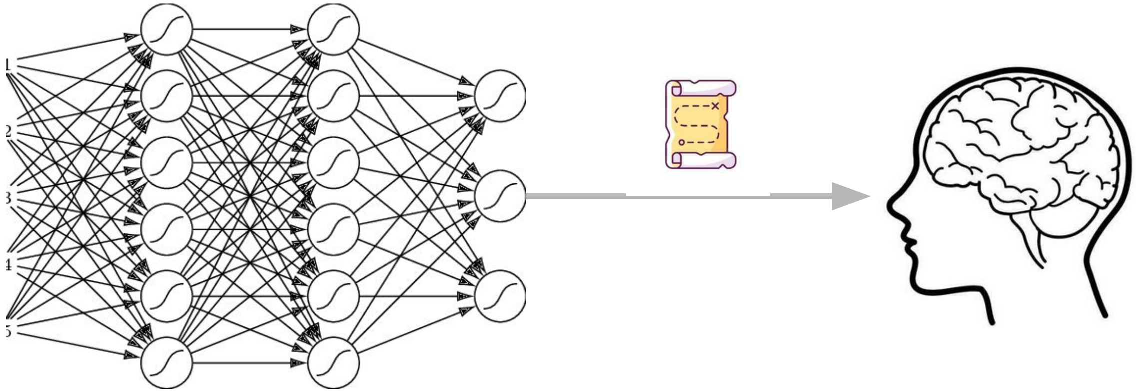
Judicial System: Why is it guilt?

Banking/Financial Domain: Why am I denied a credit?

Automobile Industry: Why did Tesla crash?

Deep learning is imperfect (e.g., robustness, backdoor, fairness, and privacy), and thus we need to understand how a neural network works so that we can improve it further.

What Is Interpretability?



That makes sense.

Human-Simulability: A function is interpretable if it is human simulatable. The trouble is human brains understand only fairly simple things.

Achieving Interpretability

Ante-hoc Interpretability

Interpretability is built-in from the beginning of the model creation,

- either by adopting models that are considered interpretable *naturally;*
- or training neural network models that encourage interpretability in the *design* of the training.

Post-hoc Interpretability

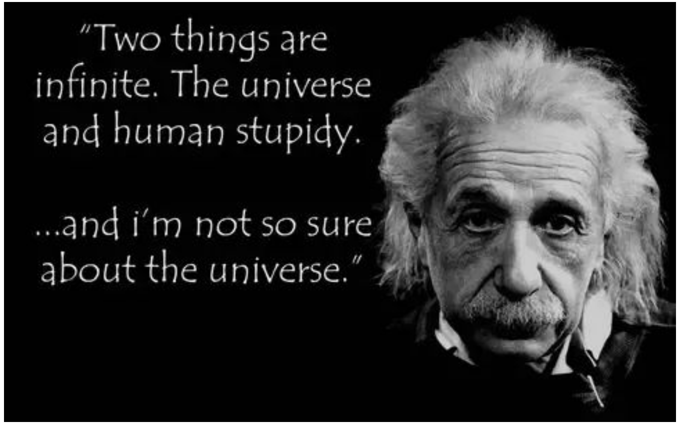
Interpretability is created after model creation for

- either explaining how a model works globally (i.e. the explanation works for almost all samples)
- or locally (i.e. the explanation works for a single data instance and its close vicinity).

Interpretable by Nature

Some classic machine learning algorithms shows that they provide models that can already be interpreted without forcing interpretability on them.

- Linear models
- Decision trees
- Rule-based models
- Naive Bayesian classification
- Markov Chain
- ...


"Two things are infinite. The universe and human stupidy.

...and i'm not so sure about the universe."

Interpretable by Nature

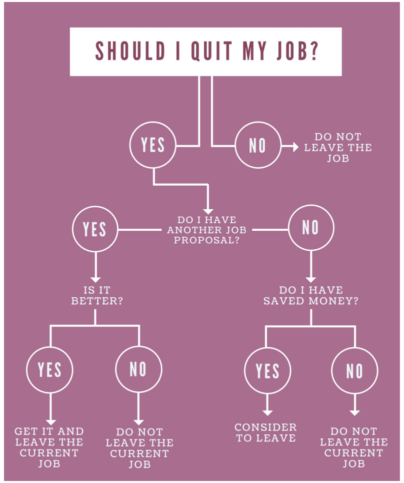
Linear Models: Example

Sales (K)	Adv (K)
368	1.7
340	1.5
665	2.8
954	5
331	1.3
556	2.2
376	1.3

Sales = 125.8 + 171.5*Adv

Q: What is the Sales if Adv is 4?

Decision Trees: Example


SHOULD I QUIT MY JOB?
YES
NO
DO NOT LEAVE THE JOB
YES
DO I HAVE ANOTHER JOB PROPOSAL?
NO
IS IT BETTER?
DO I HAVE SAVED MONEY?
YES
NO
YES
NO
GET IT AND LEAVE THE CURRENT JOB
DO NOT LEAVE THE CURRENT JOB
CONSIDER TO LEAVE
DO NOT LEAVE THE CURRENT JOB

Q: If I win SGD 10M lottery, should I quit?

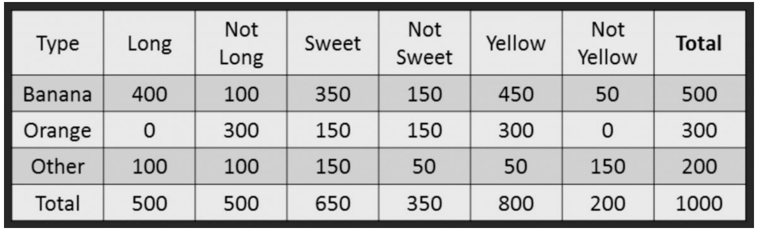
Interpretable by Nature

Rule-based models: Example

If a review contains the word “happy”, “wonderful”, “amazing” or “lucky” and there is no “not” or “little” or “hardly”, the review is positive.

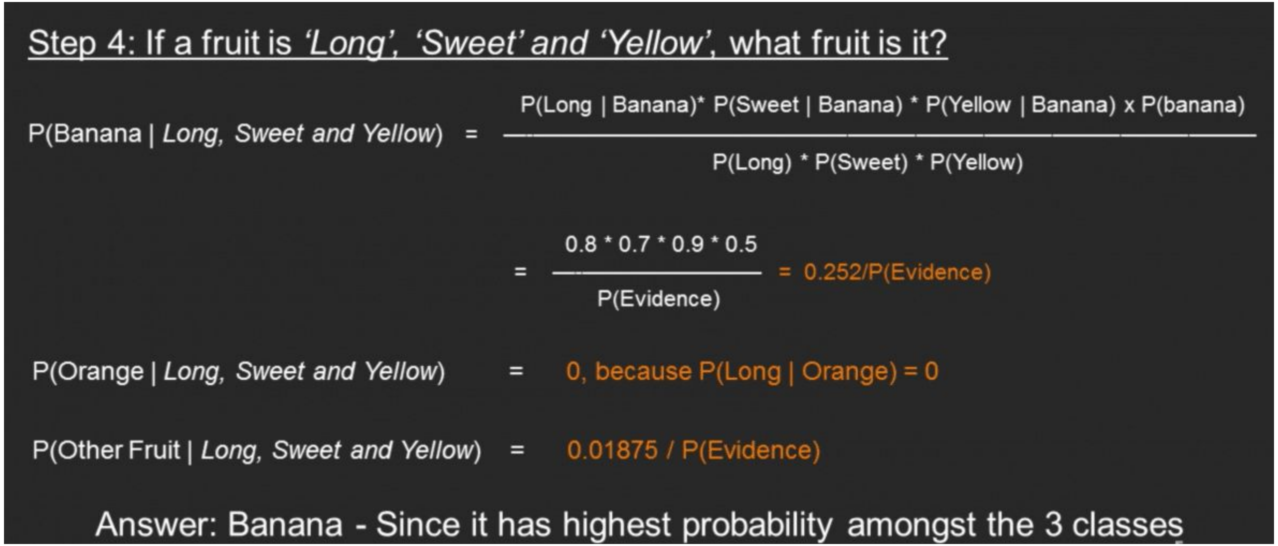
Question: Is “I was extremely lucky to get the chance to see this film upon its first day release, before entering the cinema, my expectations were already high, ...” positive?

Naive Bayes: Example



Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Question: If a fruit is long, yellow and sweet, what fruit is it?



Step 4: If a fruit is *'Long', 'Sweet' and 'Yellow'*, what fruit is it?

$$\text{P(Banana | }\textit{Long, Sweet and Yellow}\text{)} = \frac{\text{P(Long | Banana)* P(Sweet | Banana) * P(Yellow | Banana) x P(banana)}}{\text{P(Long) * P(Sweet) * P(Yellow)}}$$

$$= \frac{0.8 * 0.7 * 0.9 * 0.5}{\text{P(Evidence)}} = \text{0.252/P(Evidence)}$$

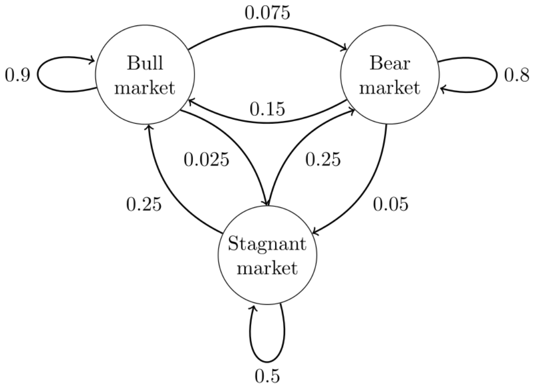
P(Orange | *Long, Sweet and Yellow*) = 0, because P(Long | Orange) = 0

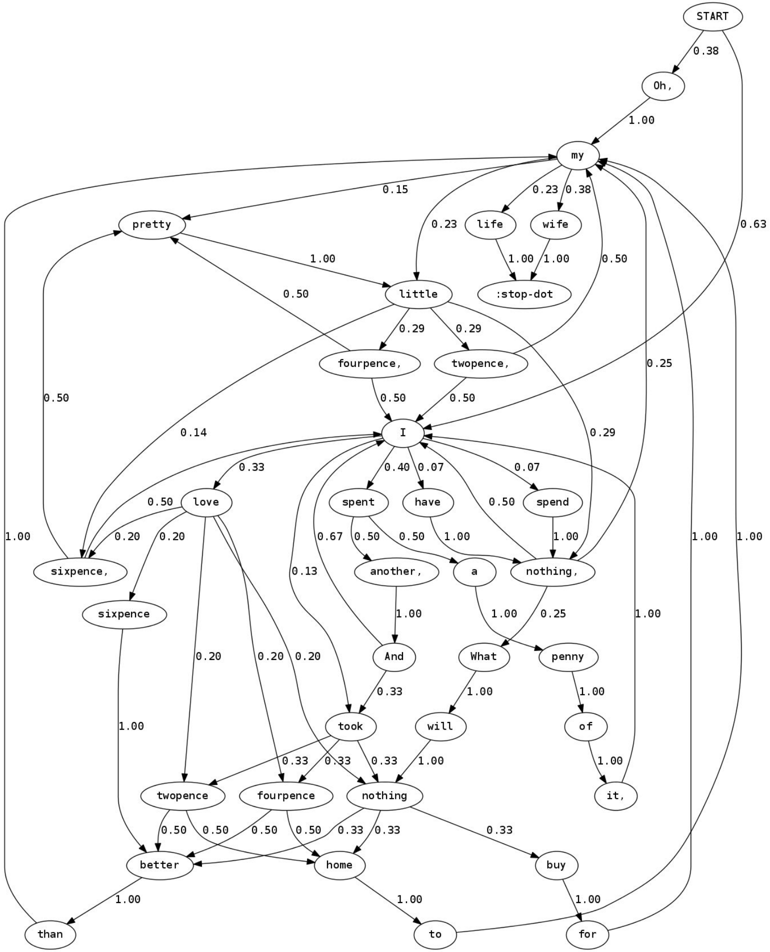
P(Other Fruit | *Long, Sweet and Yellow*) = 0.01875 / P(Evidence)

Answer: Banana - Since it has highest probability amongst the 3 classes

Discussion

Do you consider Markov Chains are interpretable by nature? (Recall Exercise 4 of Week 6)





Interpretable by Design

High-level ideas

Naturally interpretable models often have limited expressiveness.

An alternative approach is to use expressive models such as neural networks but to include interpretability in the design or in the training.

Approaches

During training, add a regularizer to encourage training more interpretable deep models

- Tree regularization of deep models
- Rule-based regularization

Tree regularization of deep models

Ideal Approach*

Train with the following objective

$$\min L_{CE}(\theta, x_i, y_i) + \lambda\Psi(\theta, x_i, y_i)$$

where λ is a constant weight and $\Psi(\theta, x_i, y_i)$ is the regularizer.

Where

$\Psi(\theta, x_i, y_i)$ is a complexity measurement of a decision tree classifier which is consistent with the neural network.

1. Train a decision tree according to the current model;
2. Measure the average path length (i.e., the number of nodes encountered before making a decision) of the training set.

**Beyond sparsity: Tree regularization of deep models for interpretability, AAAI 2018*

Tree regularization of deep models

Actual Approach*

Train with the following objective

$$\min L_{CE}(\theta, x_i, y_i) + \lambda\Psi(\theta, x_i, y_i)$$

where λ is a constant weight and $\Psi(\theta, x_i, y_i)$ is the regularizer.

Where

$\Psi(\theta, x_i, y_i)$ is approximated using another pre-trained neural network which predicts the average path length given a neural network.

(This is because $\Psi(\theta, x_i, y_i)$ is not differentiable or at least very expensive to compute in general).

**Beyond sparsity: Tree regularization of deep models for interpretability, AAAI 2018*

Tree regularization of deep models

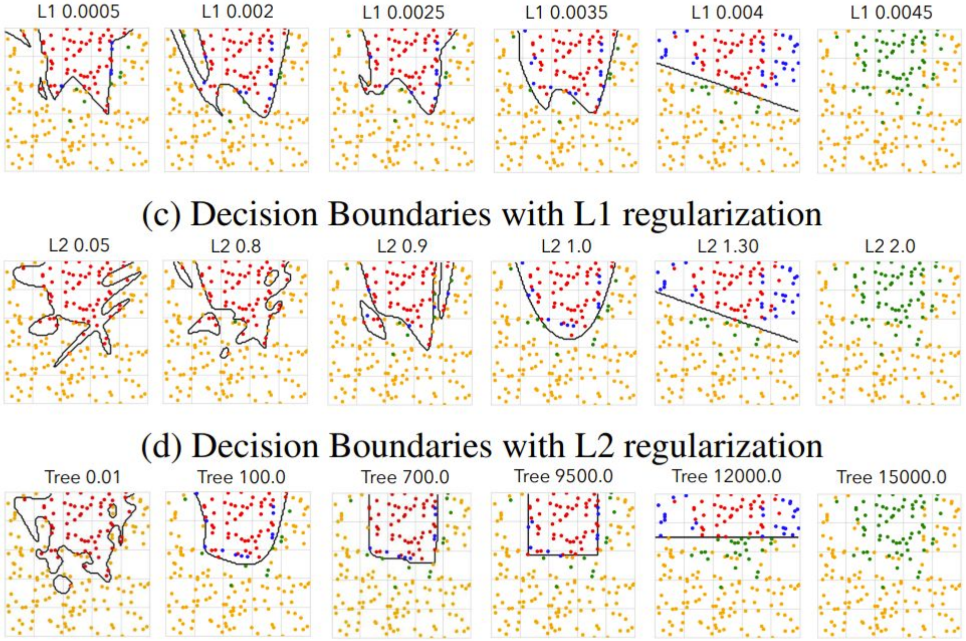
Experiment*

The true model

$$y = 5 * (x - 0.5)^2 + 0.4.$$

**Beyond sparsity: Tree regularization of deep models for interpretability, AAAI 2018*

It is not clear how this would work on realistic neural networks.



(c) Decision Boundaries with L1 regularization

(d) Decision Boundaries with L2 regularization

(e) Decision Boundaries Tree regularization

Discussion

“Interpretable by design” essentially penalties complex models. What are the pros and cons of such approaches?

Post-hoc Local Direct Explanation

Intuition

They are independent of certain model predictions and try to reveal certain properties of the black box model.

Local Direct Explanation

- Grad-CAM
- Counterfactual Explanation

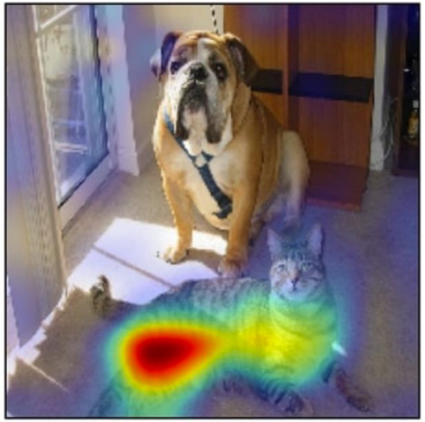
Gradient-Weighted Class Activation Mapping

Grad-CAM*

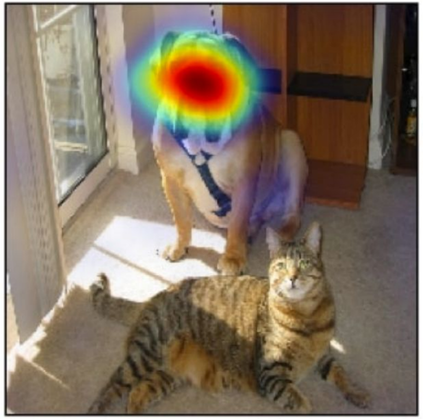
Given a sample x with class y, start with the logit layer, and compute the gradient of the input feature with respects to the class y. Highlight only those pixels which make a positive contribution to class y.

**Gradcam: Why did you say that? visual explanations from deep networks via gradient-based localization, IJCV 2016.*

Example



(c) Grad-CAM ‘Cat’



(i) Grad-CAM ‘Dog’

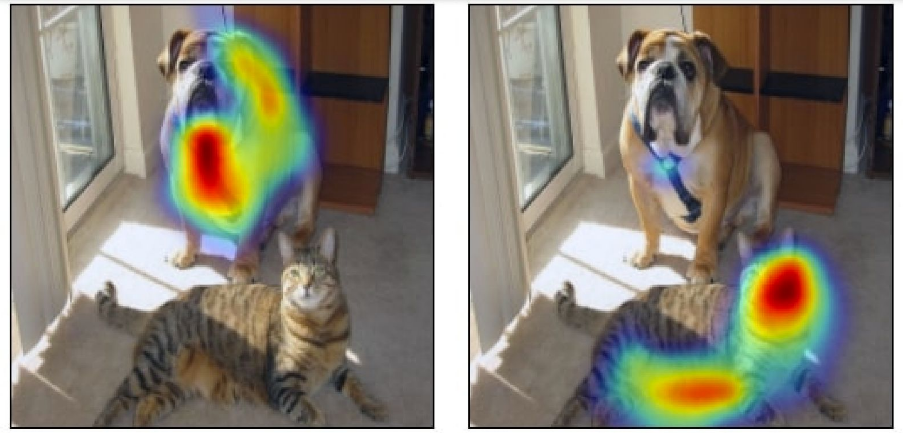
Gradient-Weighted Class Activation Mapping

Counterfactual Grad-CAM*

Given a sample x with class y, start with the *negate* of the score of class y at the logit layer, and compute the gradient of the input feature with respects to the class y.

Highlight those pixels which would make a negative contribution to class y.

Example



(b) Cat Counterfactual exp (c) Dog Counterfactual exp

Counterfactual Explanation

Counterfactual Explanation*

Given a sample x with predication y, the idea is to generate some x’ such that the predication changes.

Examples

- Person 1: If your LSAT was 34.0, you would have an average predicted score (0).
- Person 2: If your LSAT was 32.4, you would have an average predicted score (0).
- Person 3: If your LSAT was 33.5, and you were ‘white’, you would have an average predicted score (0).
- …

**Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, 2017.*

Post-hoc + Local Interpretability

Local Surrogate Models

This technique is applied whenever the model is not interpretable by itself, i.e., whenever it is a black box. An interpretable model is built on top of the black box.

The local surrogate is valid only for a specific data instance and its close vicinity.

Approaches

Local surrogate linear models

- LIME
- SHAP

Local surrogate decision trees

- Anchor
- MUSE
- LORE

LIME

High-level idea*

The aim is to produce a simple explanation model to explain how decisions are made around an instance.

**"Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD 2016*

Approach

Learn a simple model g in the form of a decision tree or linear model using the following objective.

$$\text{argmin}_g\ L(N, g, \pi_x) + \Omega(g)$$

where π_x is the proximity of an instance x and $\Omega(g)$ is a complexity measure of g (i.e., the height of a decision tree or the number of non-zero weights of a linear model).

LIME

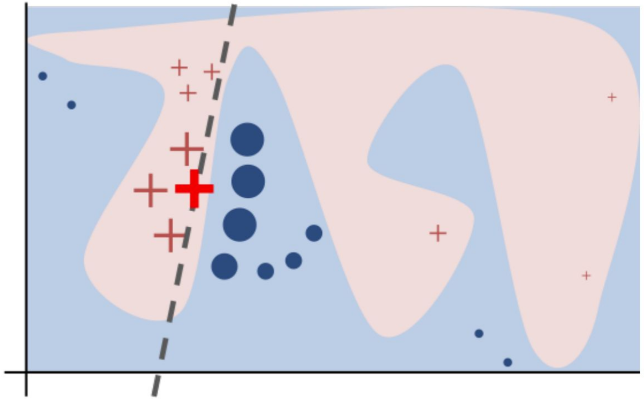
Algorithm*

Draw samples X uniformly around x (i.e., those in π_x).

Weight samples in X by their proximity to x.

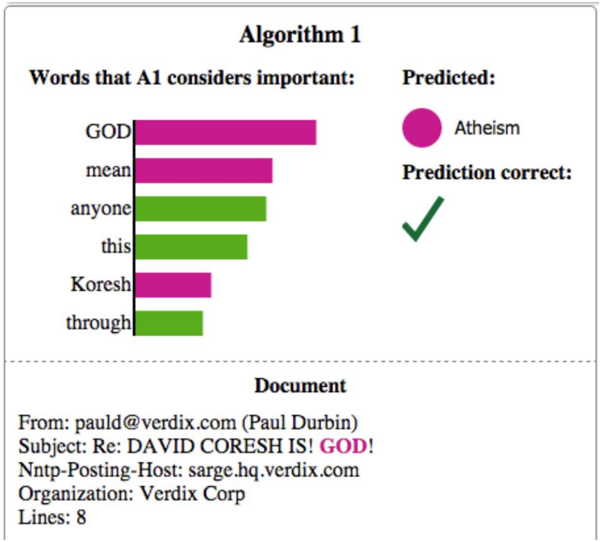
Obtain the labels of X using N.

Train a decision tree or linear model using the objective. The features in the decision tree or linear model are typically whether a word or a superpixel (a contiguous patch of similar pixels) is present or not. The loss is typically the square loss.



LIME

Example 1: “Christianity” or “Atheism”, Color indicates which class the word contributes to (green/purple for “Christianity”/“Atheism”).



Algorithm 1

Words that A1 considers important:

Predicted:

● Atheism

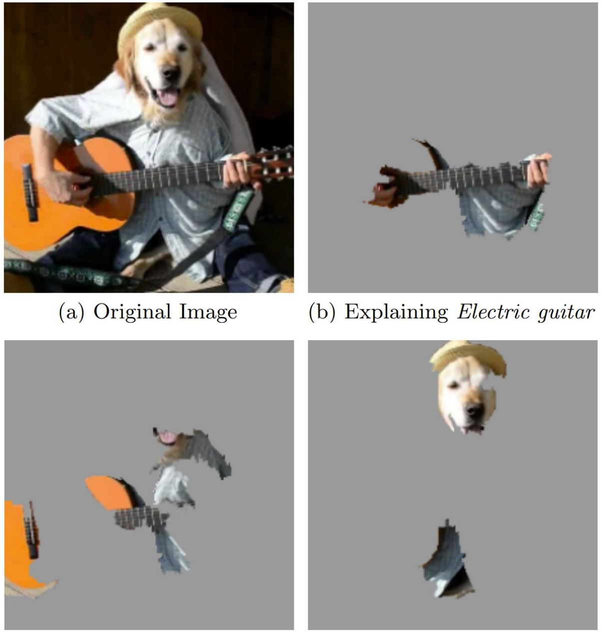
Prediction correct:

✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD**!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

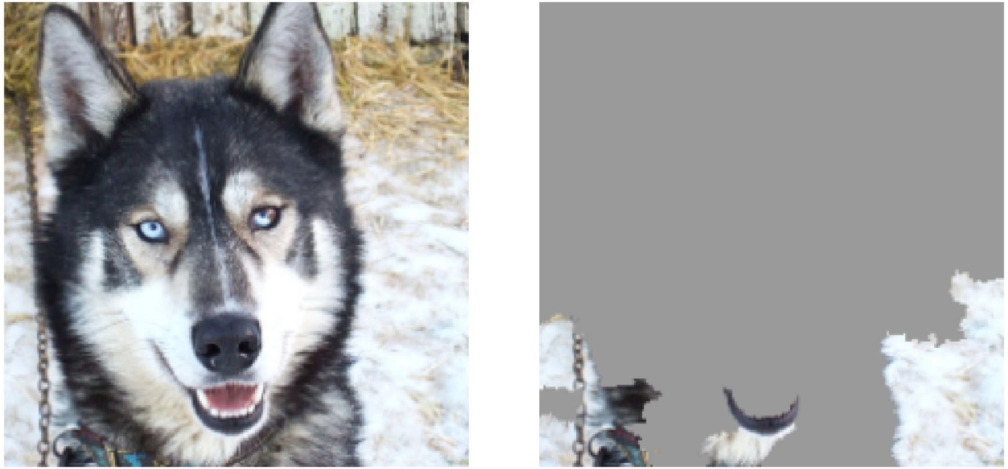
Example 2: Google image classification



(a) Original Image (b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

LIME



(a) Husky classified as wolf

(b) Explanation

What do you think? Is AI justified in this case?

SHAP

How do we compute these numbers?

SHapley Additive exPlanations*

Explain how a model works in a local region using a linear additive model.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

where z’ is any instance close to the input z; ϕ_i is a binary simplified feature.

**A unified approach to interpreting model predictions, NIPS 2017*

Example: Why is my loan rejected?

The base risk of repayment problem is 20%.

Being a day trader makes the risk 35%.

Being only 20 makes the risk 65%.

Having only one account makes the risk 75%.

Having a lot of capital gain makes the risk 55%.

Overall, the risk is 55%. Thus, it is rejected.

Shapley Value

The problem

A group of players cooperates, and obtains certain overall gain. How important is each player, and what payoff can he or she reasonably expect?

Formula

The contribution of i is as follows.

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\;(n-|S|-1)!}{n!}(v(S \cup \{i\}) - v(S))$$

Intuitively, it is the average added contribution of player i considering all combinations of players.

It is an NP-hard problem.

Shapley Value

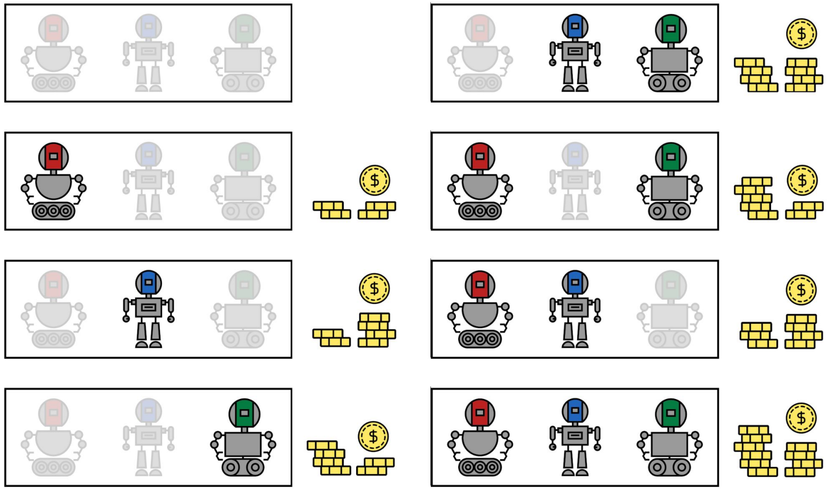
Example

You are in a 3-member group who score 50 on a project. How do you calculate your contribution?

(Alice,Bob,You) = 50
(Alice,You) = 40
(Bob,You) = 30
(You) = 20

(Alice,Bob) = 45
(Alice) = 40
(Bob) = 35
() = 0

Your contribution = (5-5+20)/4 = 5.



SHAP

SHapley Additive exPlanations*

Approximate the Shapley value of each input features (e.g., a super-pixel in an image or a word in a sentence) through sampling or Shapley kernel.

Example: Why is my loan rejected?

On average, the risk of repayment problem is 20%, i.e., **E**(N(x)) = 20%.

Being a day trader adds the risk by 15%, i.e., **E**(N(x) | being a trader) = 35%.

Being only 20 adds the risk by 30%, i.e.,

E(N(x) | being a trader, being 20) = 65%.

……

**A unified approach to interpreting model predictions, NIPS 2017*

Anchor

High-level Idea*

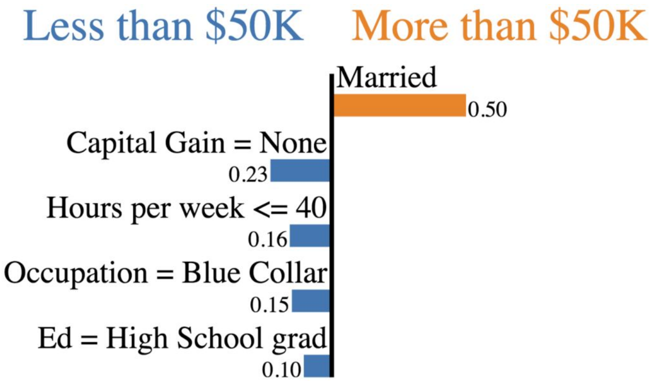
A problem of LIME is that it is not clear how far its local explanation extends to in the input space.

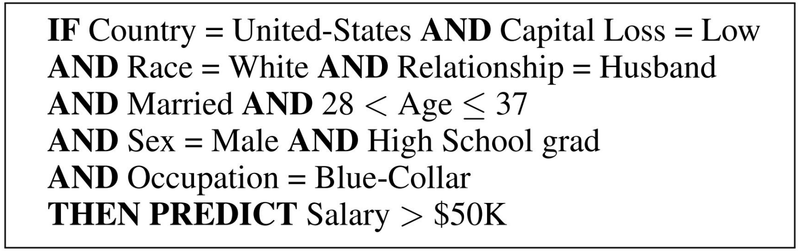
Anchor aims to generate explanations in the form of if-then rules, which specifies when it applies.

**"Anchors: High-precision model-agnostic explanations, AAAI 2018*

Example

LIME (top) vs Anchor (bottom)




IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** $28 < \text{Age} \leq 37$
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary $> \$50\text{K}$

Anchor

Approach*

An anchor A is one if-then rule which is a sufficient condition to explain why the prediction for an input x is N(x).

$$\mathbb{E}_{\mathcal{D}(z|A)}[\mathbb{1}_{f(x)=f(z)}] \geq \tau, A(x) = 1.$$

where D(z|A) D denotes the conditional distribution when the rule A applies; and τ is a threshold.

**"Anchors: High-precision model-agnostic explanations, AAAI 2018*

Example: Sentiment analysis

x is: “The movie is not bad”.

A is: if “not bad” => positive.

D(z|A) is: “This novel is not bad”, “The weather is not bad”, and so on.

If among D(z|A), A is sufficiently likely (e.g., with an accuracy of 70%), A is an anchor.

Anchor

A Greedy Algorithm*

Start with an empty rule A;

Extend A with a feature predicate (e.g., containing “not bad”) which is the best among all the predicates concerning the features.

Repeat until A is qualified as an anchor.

**"Anchors: High-precision model-agnostic explanations, AAAI 2018*

Example: Sentiment analysis

x is: “The movie is not bad”.

A is empty.

Candidate feature predicates: Containing “not bad”, Containing “bad”, and Containing “not”.

Sample many sentences containing “not bad”, 75% of them is positive; sample many sentences containing “bad”, 10% of them is positive; ...

A becomes “not bad” => positive.

LORE

Local Rule-Based Explanations (LORE)*

The model is assumed to be a black box.

The goal is to explain why certain decision is made on sample x by querying the model many times.

**"Local Rule-Based Explanations of Black Box Decision Systems, 2018*

Explanation

The explanation is in the form of

$$(\varphi \Rightarrow y, \{\varphi_0 \Rightarrow y_0, \varphi_1 \Rightarrow y_1, \ldots\})$$

which intuitively reads as “the decision is y because φ is satisfied; should φ_0 be the case, y_0 will be decision; should φ_1 be the case, y_1 will be the decision; …”

The second part are the counterfactual explanations.

LORE

Example 1: Loan Application

Sample: x = (age = 22, job = none, amount = 10k, car = no)

Decision: deny

Explanation: age ≤ 25 && job = none && amount > 5k => deny

Φ = {age > 25 && amount ≤ 5k => grant),
job = clerk && car = yes => grant}

LORE

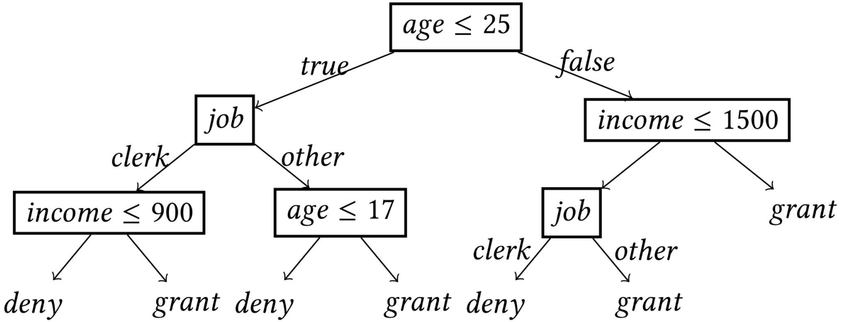
Approach

Given a sample x and its prediction N(x) = y,

1. Query N with many samples around x
2. Learn a decision tree with the queried samples
3. Output x's path in the decision tree as the explanation, and alternative paths to a different outcome as the counterfactual explanations.

Example

(age = 22, job = none, amount = 10k, car = no)



Post-hoc + Global Interpretability

Global Surrogate Models

This technique is applied whenever the model is not interpretable by itself, i.e., whenever it is a black box. An interpretable model is build on top of the black box to explain all model predictions globally.

Approaches

Global Surrogate Linear Models

- SP-Lime

Global Surrogate Decision Trees

- Decision Tree Extraction
- Soft Decision Tree
- Markov Chain Extraction

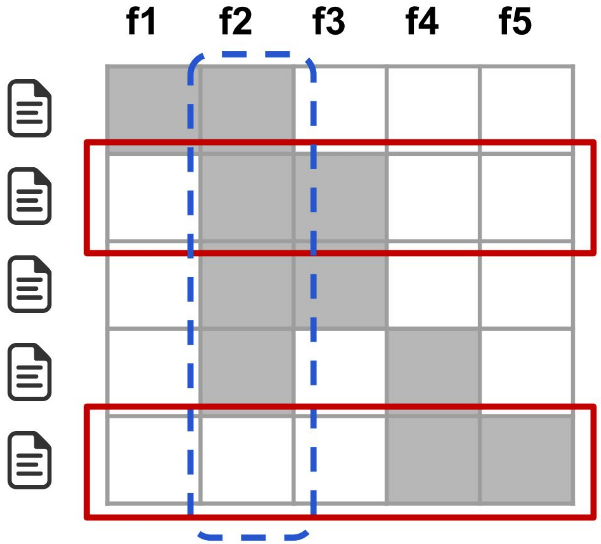
Global Surrogate Linear Models

Submodular Pick LIME*

LIME explains one sample. To have a global view, SP-LIME randomly chooses a set of samples X; picks multiple samples from X; and generates LIME explanations for the picked samples as the global explanation.

Samples from X are picked to show the relevant of different features.

**"Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD 2016*



Global Surrogate Decision Trees

Decision Tree Extraction*

Given a neural network N, learn a decision tree T to mimic N.

Note the optimization goal is to maximize the fidelity between N and T.

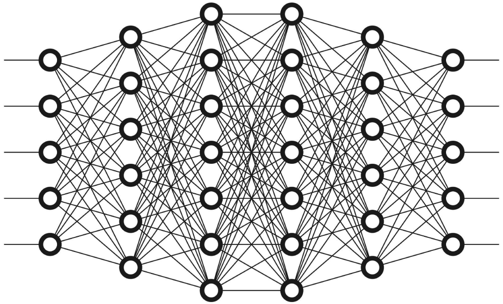
**Interpreting Blackbox Models via Model Extraction, CoRR abs/1705.08504, 2017*

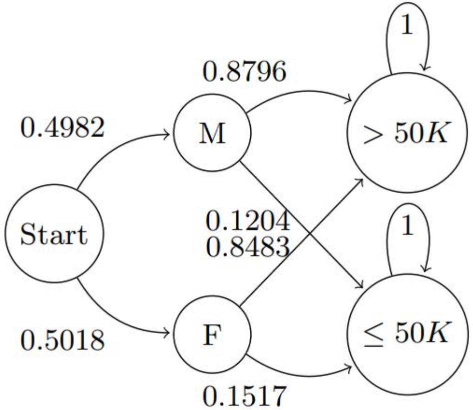
Soft Decision Tree Extraction*

Given a neural network N, learn a soft decision tree T (i.e., a decision tree with probabilities) to mimic N.

**Distilling a Neural Network Into a Soft Decision Tree, CEx@AI*IA 2017*

Global Surrogate Markov Chain





Probabilistic analysis

Week 6: slide 48

Conclusion

General interpretability is useful only to certain limited extent.

The need of interpretability is there only because deep learning is not working properly yet (e.g., problems with regard to robustness, backdoor, fairness and privacy).

The real problem is to solve is thus to improve deep learning techniques.

Aug 23 - Week 1: 7-10	Introduction	
Aug 30 - Week 2: 7-10	AI Robustness	Exercise 1
Sep 06 - Week 3: 7-10	Improving AI Robustness	Exercise 2
Sep 13 - Week 4: 7-10	AI Backdoors	Exercise 3
Sep 20 - Week 5: 7-10	Mitigating AI Backdoors	Exercise 4; Project Proposal
Sep 27 - Week 6: 7-10	AI Fairness	Exercise 5
Oct 11 - Week 7: 7-10	Improving AI Fairness	Exercise 6
Oct 18 - Week 8: 7-10	AI Privacy	Exercise 7
Oct 25 - Week 9: 7-10	Improving AI Privacy	Exercise 8
Nov 01 - Week 10: 7-10	AI Interpretability	Project Due
Nov 08 - Week 11: 1-3	End-of-Term Exam	

Remaining Challenges

Robustness

How to guarantee the neural network is robust without compromising its accuracy?

Backdoor-freeness

How to guarantee the neural network is free of all kinds of backdoors without compromising its accuracy?

Fairness

How to guarantee the neural network is fair without compromising its accuracy?

Privacy

How to guarantee the neural network preserves privacy without compromising its accuracy?