Proxy Methods for Post Hoc Interpretability

Apr 17, 2020

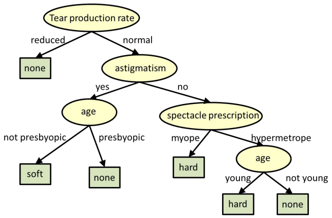
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

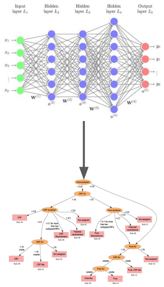
Stanford University

Recap

Previous Lecture: Intrinsic Methods for Interpretability



This Lecture: Post Hoc Methods for Interpretability



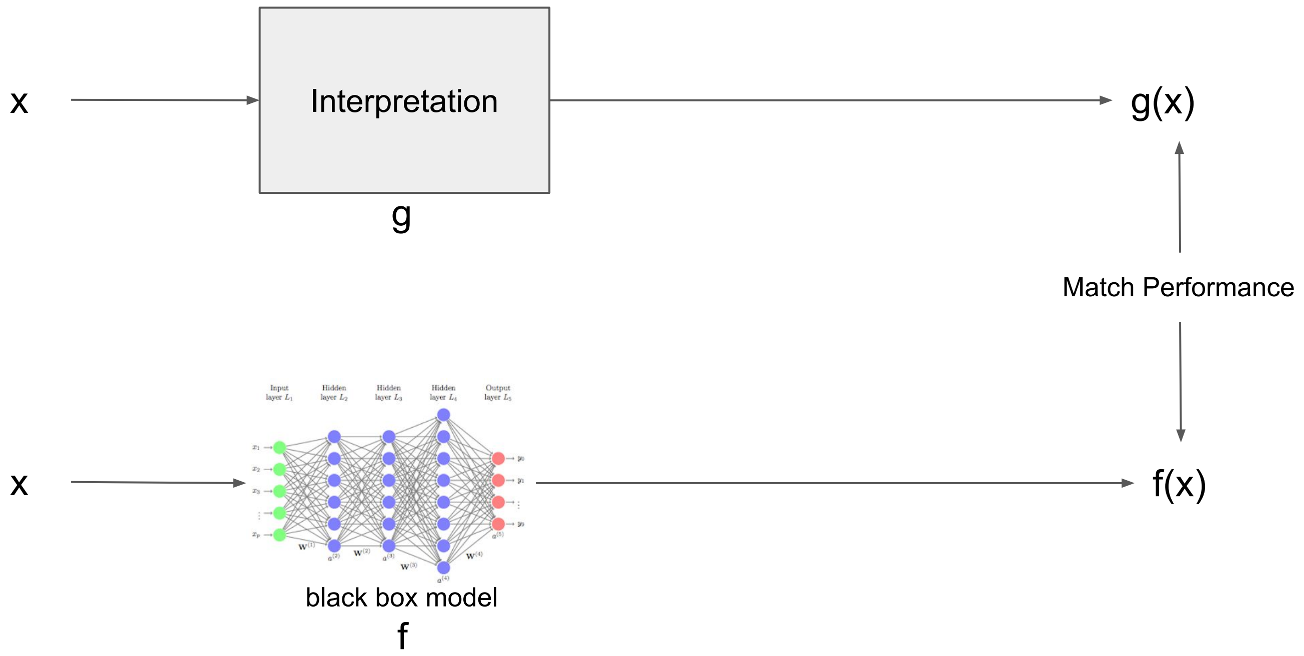
Outline

- Post Hoc Interpretability
  - Proxy Models
- Local Surrogate Methods
  - LIME
- Rule Based Learner
  - Anchors

Post Hoc Interpretability

- Model Agnostic
  - Can be applied across many different black box models
  - Multiple techniques can be applied at the same time
- Availability
  - Do not require training data
  - Do not require model training/fine-tuning
- No Performance Degeneration
  - Will not alter the black box model

Proxy Models for Post Hoc Interpretability



Outline

- Post Hoc Interpretability
  - Proxy Models
- Local Surrogate Methods
  - LIME
- Rule Based Learner
  - Anchors

Local Surrogate Methods

- Local surrogate methods aim at finding explanation g to approximate f around x based on Model Fidelity

neighbors of x

$$F(f, g, N_x) := \mathbb{E}_{x' \sim N_x}[(g(x') - f(x'))^2]$$

explanation

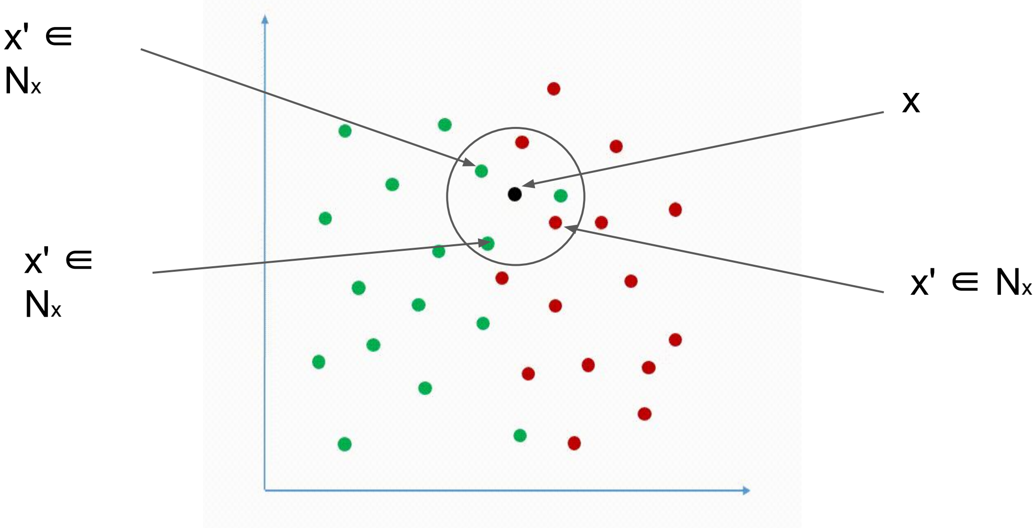
blackbox

[Plumb el al, 2018]

Local Surrogate Methods

- Local surrogate methods aim at finding explanation g to approximate f around x based on Model Fidelity

$$F(f, g, N_x) := \mathbb{E}_{x' \sim N_x}[(g(x') - f(x'))^2]$$

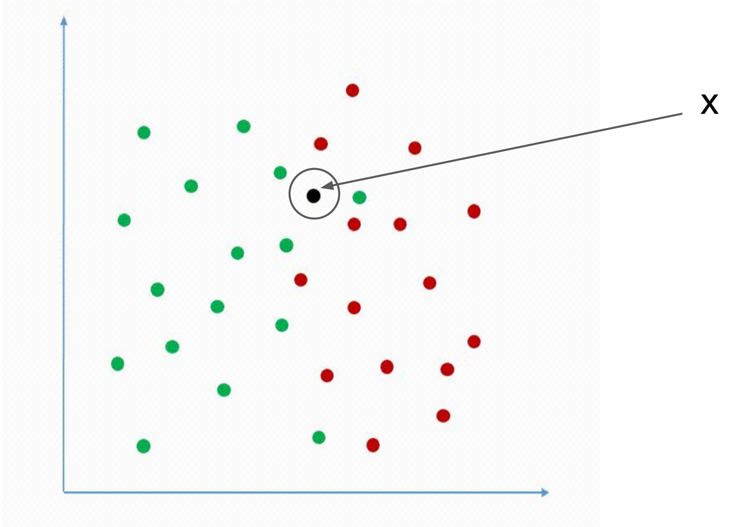


Plumb el al, 2018

Local Surrogate Methods

- Local surrogate methods aim at finding explanation g to approximate f around x based on Model Fidelity

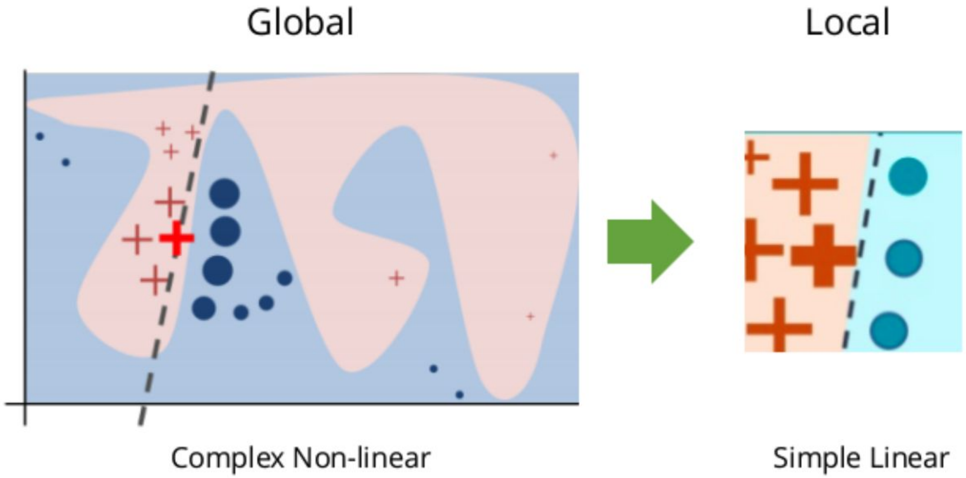
$$F(f, g, N_x) := \mathbb{E}_{x' \sim N_x}[(g(x') - f(x'))^2]$$



[Plumb el al, 2018]

Local Interpretable Model-agnostic Explanations (LIME)

- Deep learning models are usually too complex for global interpretation
  - Instead, we seek for local interpretability using simple interpretable models (e.g. linear models)



[Ribeiro el al, 2016]

LIME

- LIME generates an explainable model that optimizes both model fidelity and explanation

Black Box Model f

Explanation g

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Local Surrogate Loss

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z') \right)^2$$

Proximity Measure

Complexity Penalty

Linear Explainable Model

Local Surrogate Loss $\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$

- Linear Explainable Model
  - We use a linear model for explanation $g(z') = w_g \cdot z'$, z' ∈ {0, 1}$^{d'}$

Linear Explainable Model

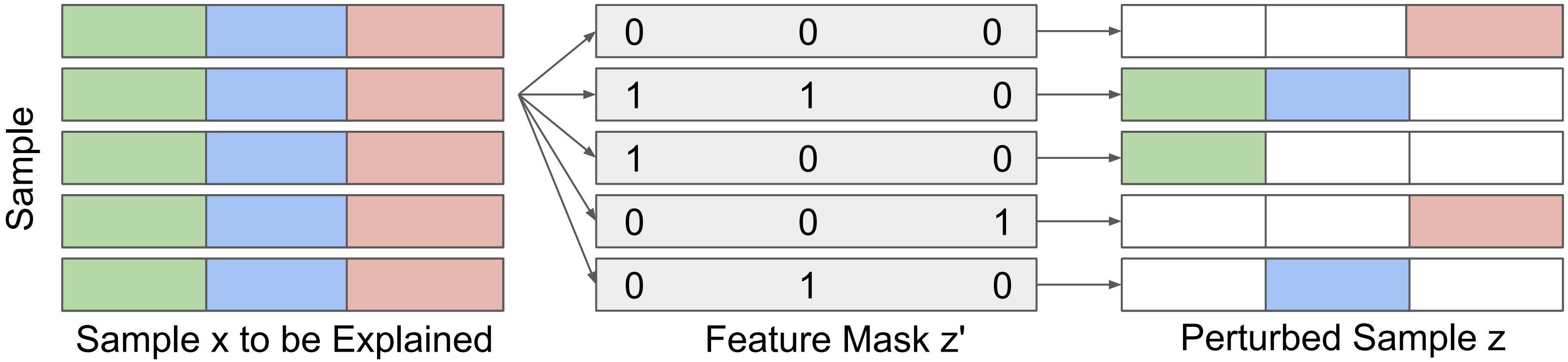
Local Surrogate Loss $\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$

- Linear Explainable Model
  - We use a linear model for explanation $g(z') = w_g \cdot z'$, z' $\in \{0, 1\}^{d'}$
  - z' is a feature mask indicating whether a specific input will be included in the explanation
  - A perturbed sample z can be recovered from mask z', $z = h_x(z')$

Linear Explainable Model

Local Surrogate Loss $\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$

- Linear Explainable Model
  - We use a linear model for explanation $g(z') = w_g \cdot z'$, z' ∈ {0, 1}$^{d'}$
  - z' is a feature mask indicating whether a specific input will be included in the explanation
  - A perturbed sample z can be recovered from mask z', z = $h_x(z')$



Training Objective for LIME

- Loss Function
  - Match predictions of the explanation model g with that of the black box model f around x
  - We use an exponentially scaled function to measure proximity
    - D = cosine distance for text
    - D = L2 distance for images

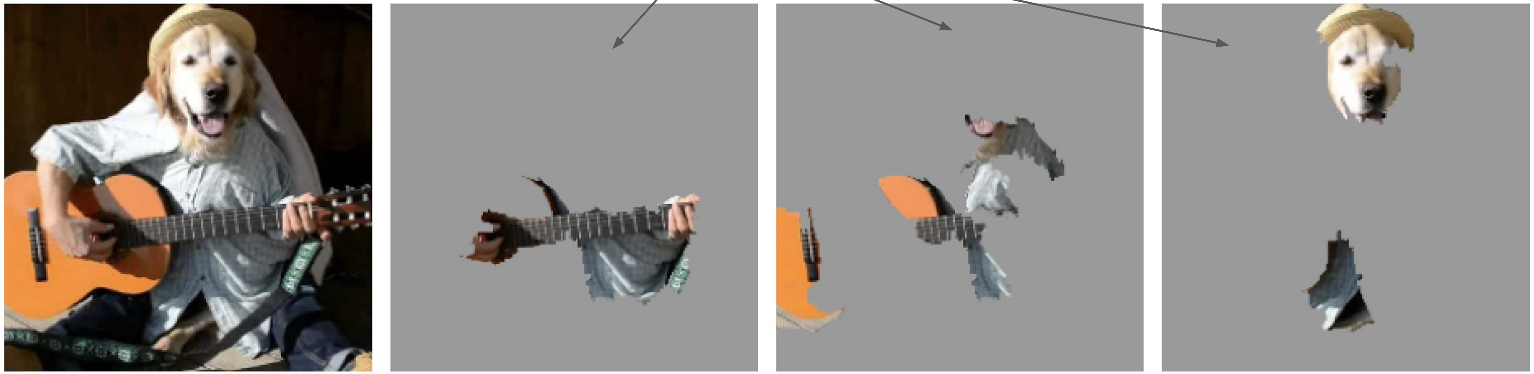
$$\pi_x(z) = exp(-D(x,z)^2/\sigma^2)$$

$$\epsilon(x) = \arg\min_{g \in G} \underbrace{\sum_{z,z'} \pi_x(z)(f(z) - g(z'))^2}_{\text{Local Surrogate Loss}} + \underbrace{\infty \cdot \mathbb{1}_{||w_g||_0 > K}}_{\text{Complexity Penalty}}$$

Ribeiro el al, 2016

Explaining Google InceptionNet

Complexity Penalty



(a) Original Image | (b) Explaining *Electric guitar* $p = 0.32$ | (c) Explaining *Acoustic guitar* $p = 0.24$ | (d) Explaining *Labrador* $p = 0.21$

x | $x \otimes w_g$

$$\epsilon(x) = \arg\min_{g \in G} \sum_{z,z'} \pi_x(z)(f(z) - g(z'))^2 + \infty \cdot \mathbb{1}_{||w_g||_0 > K}$$

Ribeiro el al, 2016

Example for Bad ML Predictions

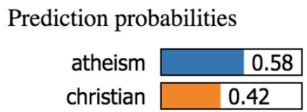
- Explanations on a model that misclassified Husky as Wolf

(a) Husky classified as wolf (b) Explanation

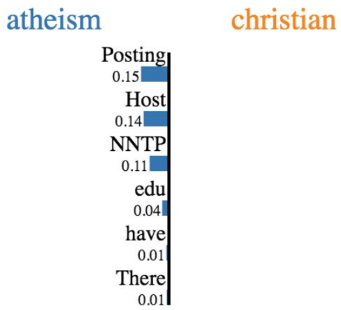
Ribeiro el al, 2016

Explaining Text Classifiers

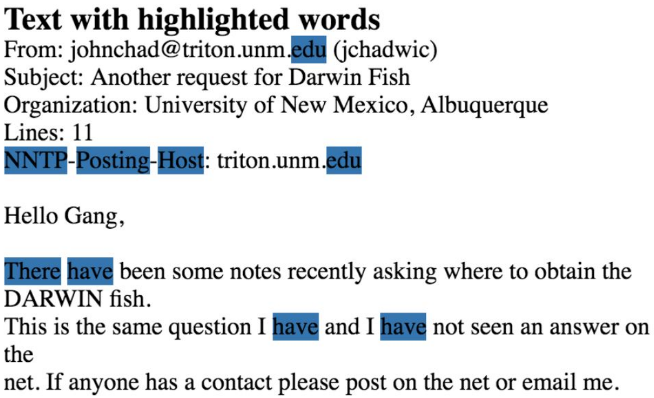
- Explanations for a SVM classifier with 94% accuracy
  - Predictions are made for arbitrary reasons
  - The word “Posting” appears in 22% of examples in the training set
    - 99% of which are samples attribute to class “Atheism”



$f(x)$



w_g



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

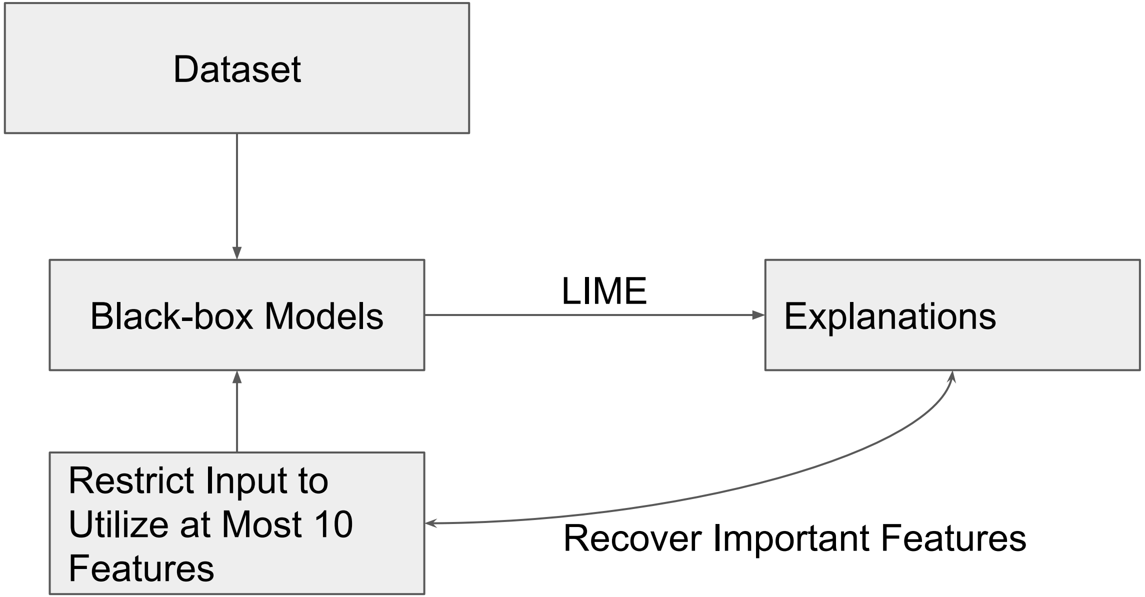
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

$x \otimes w_g$

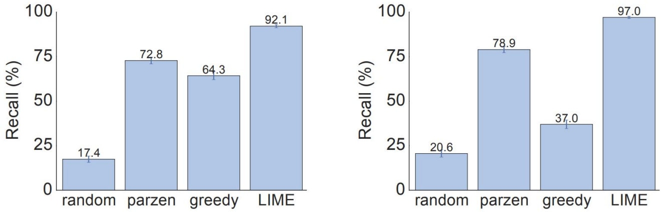
Ribeiro el al, 2016

Faithfulness of Explanations



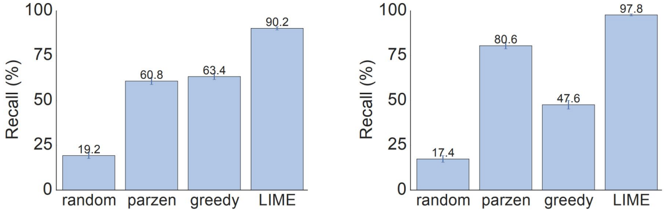
Faithfulness of Explanations

- LIME Achieves Good Faithfulness
- Sentiments classification tasks
  - Books, DVDs
- Classifiers
  - logistic regression with L2 reg. (Sparse LR)
  - decision tree



(a) Sparse LR (b) Decision Tree

Books Dataset



(a) Sparse LR (b) Decision Tree

DVDs Dataset

parzen - Baehrens et al, 2010
random - randomly pick K features
greedy - remove features contribute most to the classifiers

Ribeiro el al, 2016

Trustworthiness for ML Models

- Human discredits certain features in the learning tasks
- Classifiers that use those features will be considered not trustable.

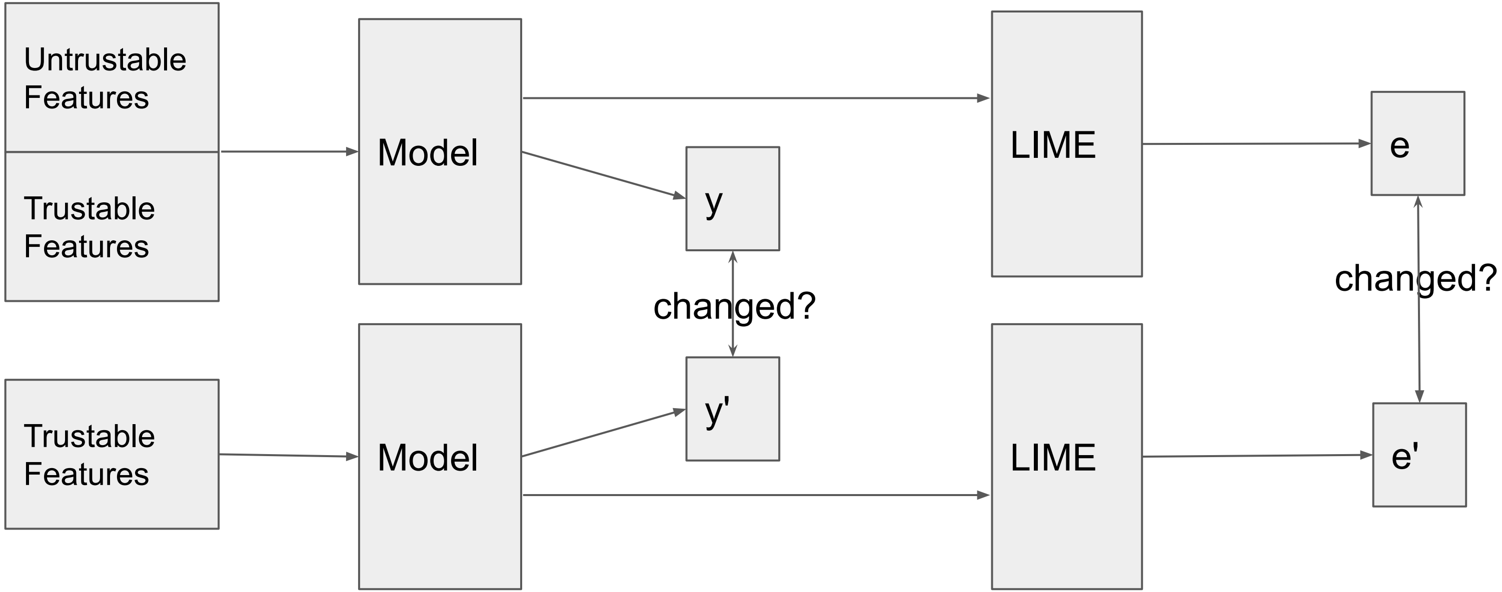
Predict the need for ICU

heart beat	temperature	~~salary~~
120 BPM	101 F	$20,000
80 BPM	104.4 F	$40,000
140 BPM	99 F	$800,000
110 BPM	100 F	$30,000

Need for ICU?
N
Y
Y
N

Ribeiro el al, 2016

Trustworthiness for Explanations



Trustworthiness of Predictions

- Untrustable Features
  - 25% of features are "untrustable features"

- Trustworthiness of Predictions
  - Compares changes of model predictions and the changes of model explanations when unstable features are removed

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

Trustworthiness of LIME with different ML models:

- Logistic Regression with L2 regularization (LR)
- Nearest Neighbors (NN)
- Random Forests (RF)
- Support Vector Machines (SVM)

Ribeiro el al, 2016

Explaining Multiple Samples

- Explain a set of samples to get a complete picture of the model
  - Each sample $x_i \in X$ will have its interpretation

$$g_{x_i}(z) = w_i \cdot z = \sum_j w_{i,j} \cdot z_j$$

- How do we select samples?
  - Select samples to cover the maximum information about the model

$$I_j = \sqrt{\sum_{x_i \in X} |w_{i,j}|}$$

Ribeiro el al, 2016

Explaining Multiple Samples

- How do we select samples?
  - Select samples to cover the maximum information about the model

$$I_j = \sqrt{\sum_{x_i \in X} |w_{i,j}|}$$

  - Set function

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$$

  - We want to get a set of samples V up to B elements that maximize c

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}} \, c(V, \mathcal{W}, I)$$

Ribeiro el al, 2016

Explaining Multiple Samples

- How do we select samples?
  - We want to get a set of samples V up to B elements such to maximize c

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$$

- Optimization
  - Searching for the global optimal set of V is NP-Hard (Feige, 1998)
  - We turn to greedy algorithm as an approximation method

Ribeiro el al, 2016

Greedy Algorithm for Sample Selection

- Pick a subset of samples up to B elements from X to maximize c

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}}\, c(V, \mathcal{W}, I)$$

- Start with an empty set $V_0 = \varnothing$,
- For the i^{th} step
  - Pick the next element $x_i \in X \setminus V_i$, such that x_i maximizes $c(V_i \cup \{x_i\}, W, I)$
  - repeat until $|V_i| \geq B$

Ribeiro el al, 2016

Theoretical Guarantees on Performance

- A function defined on a set is submodular if
  - for every $V_A \subseteq V_B$

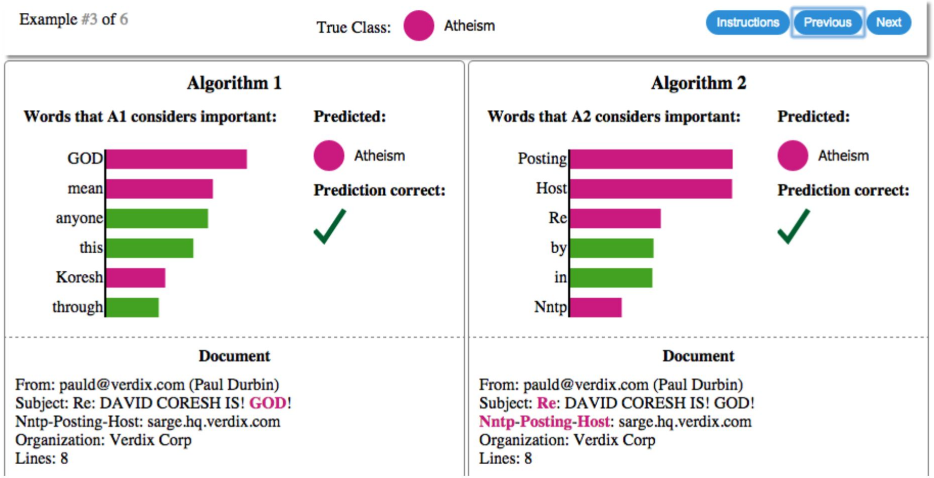
$$c(V_A \cup \{x\}) - c(V_A) \geq c(V_B \cup \{x\}) - c(V_B)$$

- Properties of Submodular functions
  - The performance of a greedy algorithm is at least 1-1/e (~63%) to the optimum
  - $c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$ is submodular
  - The performance of a greedy algorithm on c is guaranteed with a lower bond

Ribeiro el al, 2016

Human Experiments

- Ask Human to Select the Best Classifier
  - Annotators are shown the explanations
  - Annotators have no knowledge in machine learning

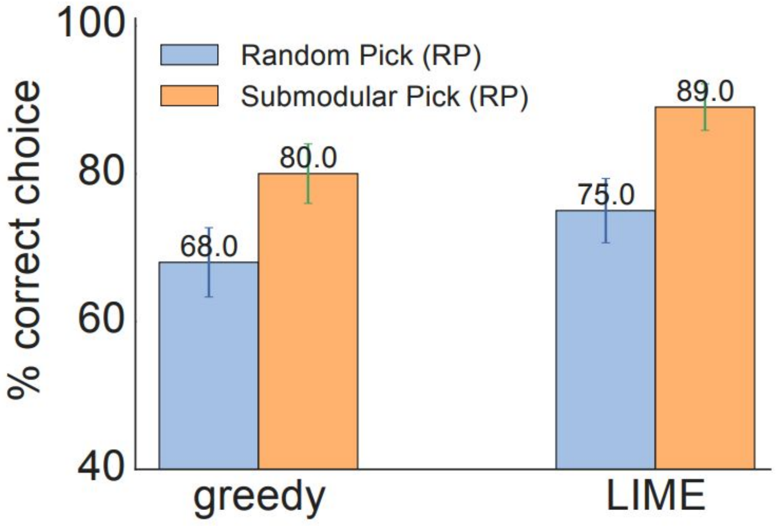


Classification of Atheism/Christian in the 20 newsgroups dataset

[Ribeiro el al, 2016]

Human Experiments - Select the Best Classifier

- Original model: SVM trained on the dataset with original features
- Cleaned model: SVM trained on the dataset with "cleaned features"



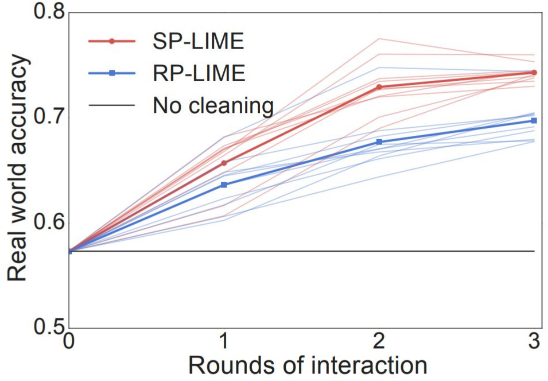
results on 20 newsgroups dataset

greedy - remove features contribute most to the classifiers

Ribeiro el al, 2016

Improving Models Through ML Interpretability

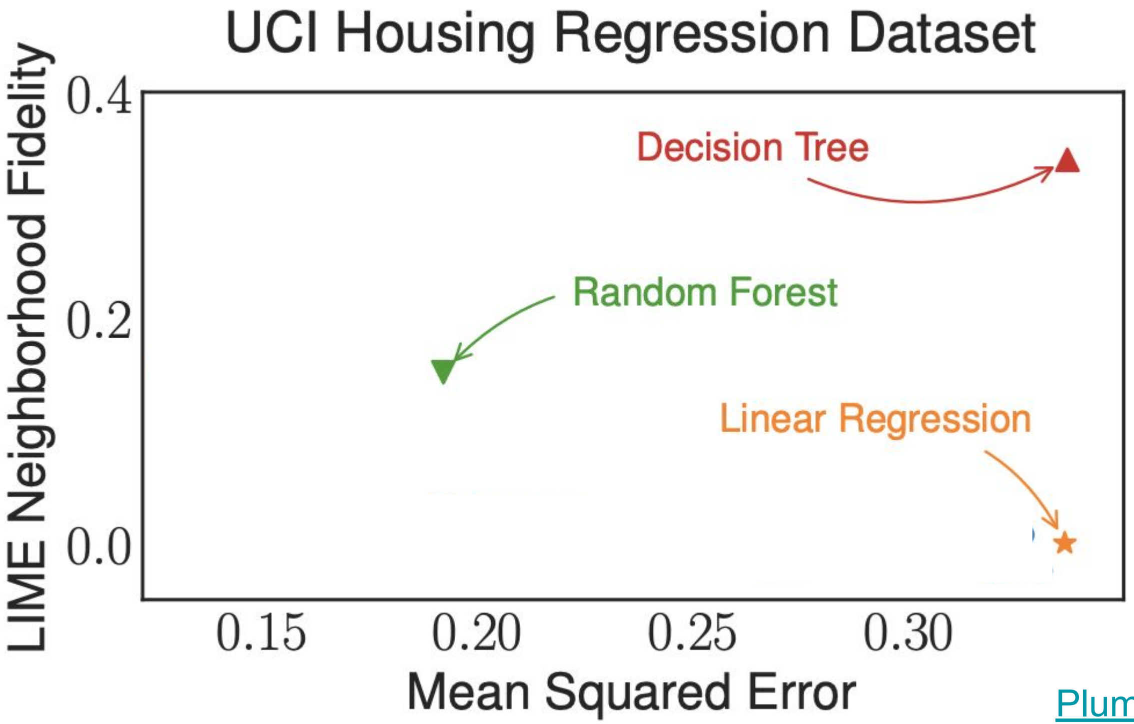
- Improving ML Models
  - Human raters are shown model interpretability
  - They are asked to improve the model by masking out unnecessary features
    - Which words from the explanations should be removed from subsequent training
  - SP - select samples by random
  - RP - select samples by greedy algorithm



results on 20 newsgroups dataset

Ribeiro el al, 2016

Faithfulness of Model Explanations



Plumb et al, 2019

Outline

- Post Hoc Interpretability
    - Proxy Models
- Local Surrogate Methods
    - LIME
- Rule Based Explainers
    - Anchors

Rule Based Explainers

- Explain the Predictions of Deep Learning Models Using Rules
  - How do we find the set of rules for a particular predictor?

	If	Predict
adult	No capital gain or loss, never married	$\leq 50K$
	Country is US, married, work hours > 45	$> 50K$
rcdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score ≤ 649	Bad Loan
	$649 \leq$ FICO score ≤ 699 and $\$5,400 \leq$ loan amount $\leq \$10,000$	Good Loan

Ribeiro el al, 2018

Anchors

- Generate A Set of Feature Predicates Known as Anchors A (i.e., rules)
  - Using anchors to explain the performance of deep model f
    - mimic the decisions of deep models on x, f(x)
    - explain a wide range of similar decisions in the dataset

	If	Predict
adult	No capital gain or loss, never married	$\leq 50K$
adult	Country is US, married, work hours > 45	$> 50K$
rcdv	No priors, no prison violations and crime not against property	Not rearrested
rcdv	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score ≤ 649	Bad Loan
lending	$649 \leq$ FICO score ≤ 699 and $\$5,400 \leq$ loan amount $\leq \$10,000$	Good Loan

Anchors found in adult income dataset

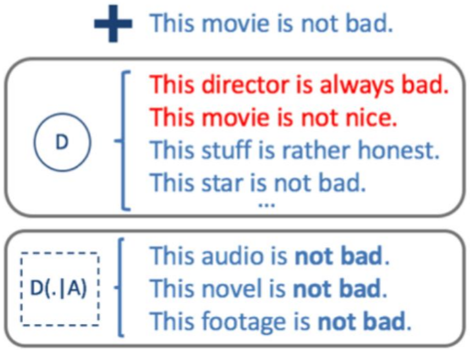
Ribeiro el al, 2018

Anchors

- An Anchor is a set of feature predicates applied to the feature space

A = {“not”, “bad”}

  - Any text sample x containing both "not" and "bad" will be selected by the anchor
    A(x) = 1
  - An anchor can be applied to a dataset D to generate a subset D(.|A)



A = {“not”, “bad”}

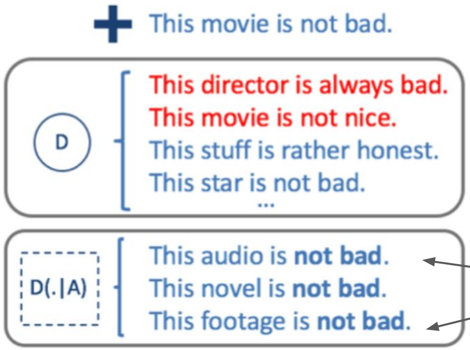
Samples in a data set D

Samples selected by Anchor A

Ribeiro el al, 2018

Formal Definitions of Anchors

- Preconditions of Anchors
  - Applies to the sample x being interpreted
  - Precisions
    - Samples covered by the same anchor A need to have the similar predictions
    - i.e., f(x)=f(z) for z~D(.|A)
  - Coverage
    - A significant portion of the data needs to be covered by Anchor A.



Anchor generated from
x= "This movie is not bad. "

same predictions
f(x)=f(z)

[Ribeiro el al, 2018](#)

Formal Definitions of Anchors

- Preconditions of Anchors
  - Applies to the sample x being interpreted

$$A(x) = 1$$

  - Precision
    - Samples covered by the same anchor A need to have similar predictions

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

  - Coverage
    - A significant amount of data needs to be covered by one anchor A.

$$\mathbb{E}_{\mathcal{D}(z)} A(z) \geq c$$

Ribeiro el al, 2018

Anchors for Part of Speech Tagging

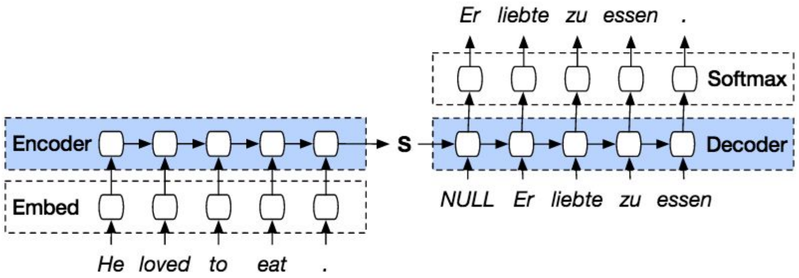
Instance	If	Predict
I want to play(V) ball.	previous word is PARTICLE	play is VERB.
I went to a play(N) yesterday.	previous word is DETERMINER	play is NOUN.
I play(V) ball on Mondays.	previous word is PRONOUN	play is VERB.

Ribeiro el al, 2018

Anchors for Machine Translation

- Group Predictions of Words with Similar Meanings
    - "esta" (feminine of word "this")
    - "este" (masculine of word "this")
    - "isso" (if its referent is not in the sentence)

English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar
This is the **problem** we must address	Este é o problema que temos que enfrentar
This is what we must address	É isso que temos de enfrentar



[Ribeiro el al, 2018](https://)

Anchors for Image Classification (InceptionV3)

original image

Anchors for "beagle"

Ribeiro el al, 2018

Anchors for Visual Question Answering (VQA)

What animal is featured in this picture ?	dog
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

Anchor for predicting "dog"

Where is the dog?	on the floor
What color is the wall?	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

Other Anchors

Ribeiro el al, 2018

Generating Anchors

- Preconditions
  - Precision $$\text{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right]$$
  - Coverage $$\text{cov}(A) = \mathbb{E}_{\mathcal{D}(z)}[A(z)]$$

- Challenges in Generating the Optimal A
  - Calculating precision and coverage is computationally intensive
    - will need to iterate through the predictions of f over the entire dataset
  - Usually difficult to apply white box optimization techniques (e.g., gradient descent)

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A)$$

Ribeiro el al, 2018

Generating Anchors

- Optimization Target

$$\max_{A \text{ s.t. } P(\mathrm{prec}(A) \geq \tau) \geq 1-\delta} \mathrm{cov}(A)$$

- Searching for the Optimal A
  - for each step t,
    - 1) Construct a set of candidate solutions with the best coverage
      - Candidate solutions need to satisfy cov(A) ≥ c
    - 2) Pick top-k candidates with the best precision
      - Candidates need to have prec(A) ≥ τ with confidence at least 1-δ
    - 3) Update the optimal Anchor A*

Ribeiro el al, 2018

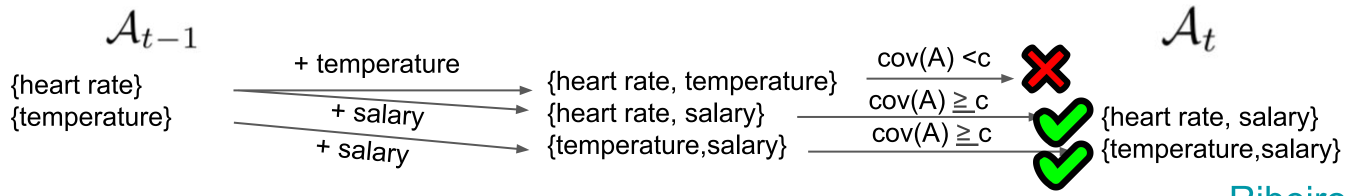
Generating Anchors - Optimizing Coverage

- Searching for the Optimal A
  - 1) Optimizing coverage with cov(A) ≥ c
  - 2) Optimizing precision with prec(A) ≥ τ and confidence at least 1-δ
  - 3) Update the optimal solution A*

$$\mathrm{cov}(A) = \mathbb{E}_{\mathcal{D}(z)}[A(z)]$$

- Optimizing Coverage
  - Start with $\mathcal{A}_0$ = ∅
  - Expand $\mathcal{A}_{t-1}$ by one element to get $\mathcal{A}_t$

heart beat	temperature	salary
120 BPM	101 F	$20,000
80 BPM	104.4 F	$40,000
140 BPM	99 F	$800,000



Ribeiro el al, 2018

Generating Anchors - Optimizing Precisions

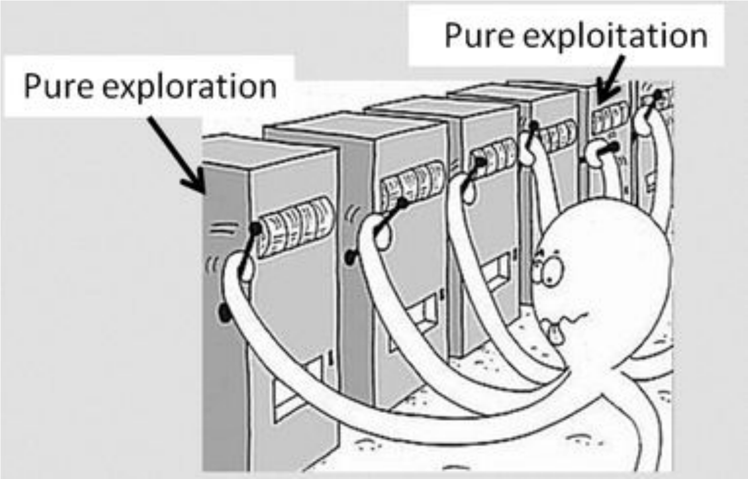
- Searching for the Optimal A
    - 1) Optimizing coverage with cov(A) ≥ c
    - 2) Optimizing precision with prec(A) ≥ τ and confidence at least 1-δ
    - 3) Update the optimal solution A*

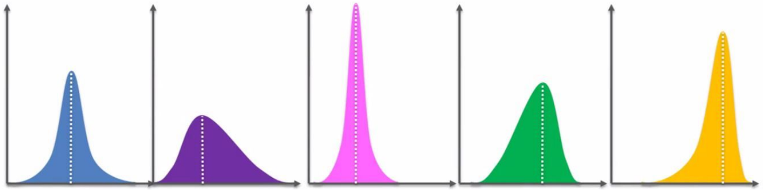
$$\operatorname{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right]$$

- Optimizing Precisions
    - Formulate it as a Multi-armed bandit optimization problem

Ribeiro el al, 2018

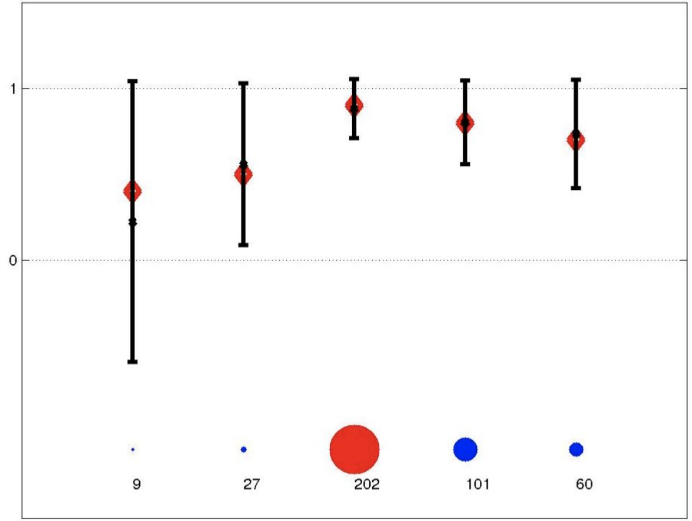
Multi-Armed Bandit Problem

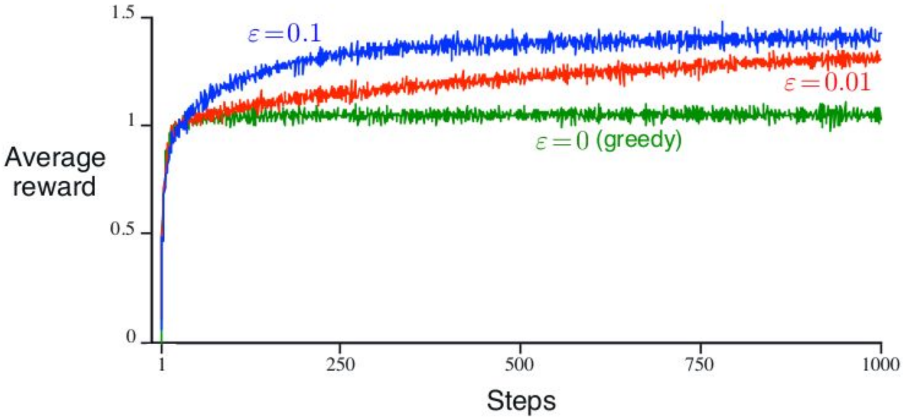




Reward Distribution of Each Arm

Exploration and Exploitation Trade-offs





Generating Anchors - Optimizing Precisions

- Searching for the Optimal A
  - 1) Optimizing coverage with cov(A) ≥ c
  - 2) Optimizing precision with prec(A) ≥ τ and confidence at least 1-δ
  - 3) Update the optimal solution A*

$$\text{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right]$$

- Optimizing Precisions
  - Formulate it as a Multi-armed bandit optimization problem
    - Find out candidates with Prec(A) ≥ τ
    - Using minimal costs (number of pulls of the arms)
    - Each candidate solution A is an arm
    - Prec(A) of a single sample is the latent reward

Ribeiro el al, 2018

Generating Anchors - Optimizing Precisions

- Searching for the Optimal A
  - 1) Optimizing coverage with cov(A) ≥ c
  - 2) Optimizing precision with prec(A) ≥ τ and confidence at least 1-δ
  - 3) Update the optimal solution A*

$$\mathrm{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right]$$

- Optimizing Precisions
  - Formulate it as a Multi-armed bandit optimization problem
    - Find out candidates with Prec(A) ≥ τ
    - Using minimal costs (number of pulls of the arms)
    - Each candidate solution A is an arm
    - Prec(A) of a single sample is the latent reward
  - Return the top K arms (i.e., A) with the highest reward (Prec(A)) that satisfies conditions
    - Prec(A) ≥ τ, P(Prec(A) ≥ τ) ≥ 1-δ

Ribeiro el al, 2018

Generating Anchors - Update Optimal A*

- Searching for the Optimal A
  - 1) Optimizing coverage with cov(A) ≥ c
  - 2) Optimizing precision with prec(A) ≥ τ and confidence at least 1-δ
  - 3) Update the optimal solution A*

- Update A*
  - For the top-k A returned in step 2)
    - Find the best A* based on the Coverage criteria

$$\textbf{if } \operatorname{cov}(A) > \operatorname{cov}(A^*) \textbf{ then } A^* \leftarrow A$$

    - Loop into the next step

Ribeiro el al, 2018

Precision and Coverage

- Precision

$$\text{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right]$$

- Coverage

$$\text{cov}(A) = \mathbb{E}_{\mathcal{D}(z)}[A(z)]$$

- Limes
  - lime-n - Naive LIME algorithm
  - lime-t - Make predictions only when its predictive probability is above a threshold

		Precision		Coverage	
		anchor	**lime-n**	**anchor**	**lime-t**
adult	logistic	95.6	81.0	10.7	21.6
	gbt	96.2	81.0	9.7	20.2
	nn	95.6	79.6	7.6	17.3
rcdv	logistic	95.8	76.6	6.8	17.3
	gbt	94.8	71.7	4.8	2.6
	nn	93.4	65.7	1.1	1.5
lending	logistic	99.7	80.2	28.6	12.2
	gbt	99.3	79.9	28.4	9.1
	nn	96.7	77.0	16.6	5.4

logistic: logistic regression, gbt: gradient boosted trees
nn: two layers of 50 units MLP

Ribeiro el al, 2018

User Study

- Ask Users to Guess the Outcomes of A ML Model After Explanations
  - Human annotators are 26 students who took a machine learning course
  - Calculate precision and coverage of the users' performance
  - Human mark "I don't know" when they are not certain, which makes coverage the perceived one.

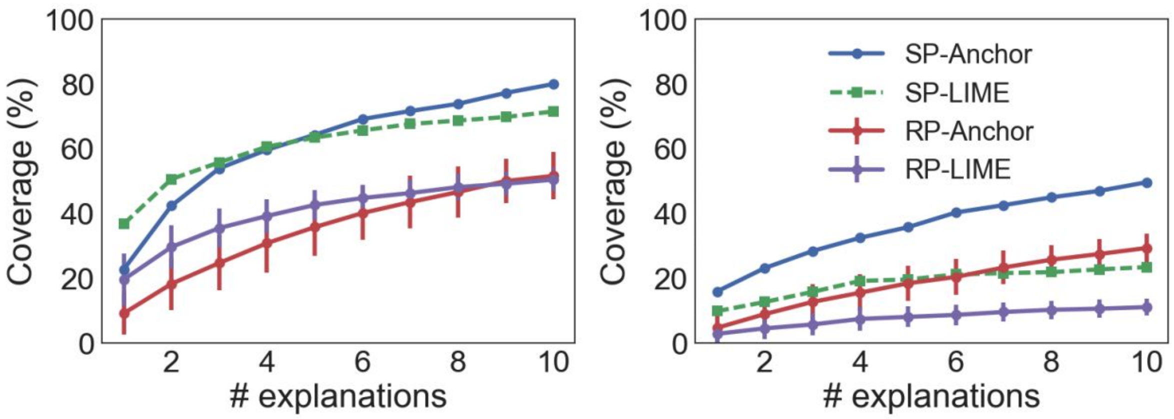
Method	Precision				Coverage (perceived)				Time/pred (seconds)			
	adult	rcdv	vqa1	vqa2	adult	rcdv	vqa1	vqa2	adult	rcdv	vqa1	vqa2
No expls	54.8	83.1	61.5	68.4	79.6	63.5	39.8	30.8	29.8 ±14	35.7±26	18.7±20	13.9±20
LIME(1)	68.3	98.1	57.5	76.3	89.2	55.4	71.5	54.2	28.5±10	24.6±6	8.6±3	11.1±8
Anchor(1)	100.0	97.8	93.0	98.9	43.1	24.6	31.9	27.3	13.0±4	14.4±5	5.4±2	3.7±1
LIME(2)	89.9	72.9	-	-	78.5	63.1	-	-	37.8±20	24.4±7	-	-
Anchor(2)	87.4	95.8	-	-	62.3	45.4	-	-	10.5±3	19.2±10	-	-

LIME(n): results after n LIME explanations
Anchor(n): results after n Anchor explanations

Ribeiro el al, 2018

User Study Results

- Coverage change with number of explanations seen by the same user.
  - gradient boosted trees(gb)
  - SP - Submodular Pick
  - RP - Random PIck



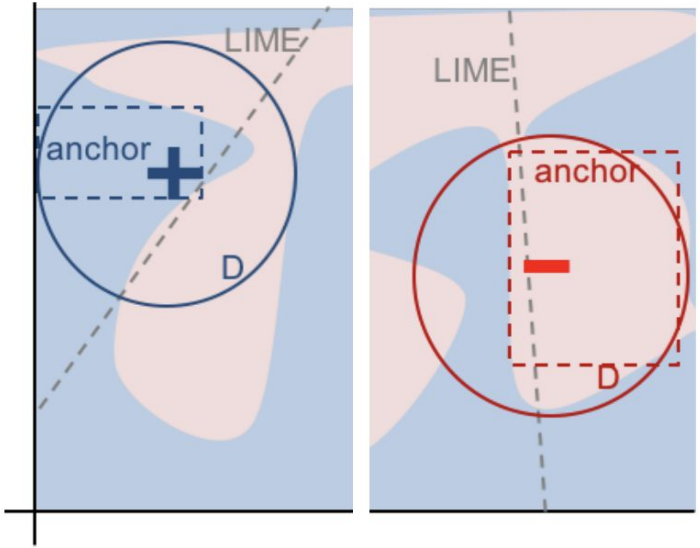
(a) *adult* dataset (b) *rcdv* dataset

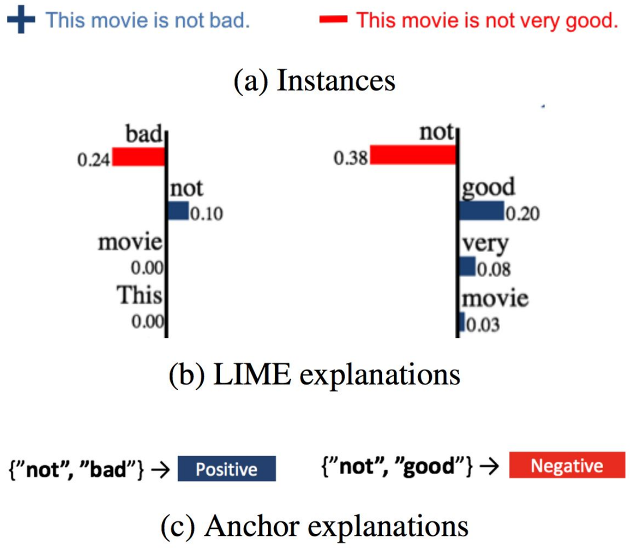
[Ribeiro el al, 2018](#)

Comparisons to LIME

	LIME	Anchors
Explanations	$g(z') = w_g \cdot z'$	Anchors A
Optimization Target	$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$	$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A)$

Comparisons to LIME



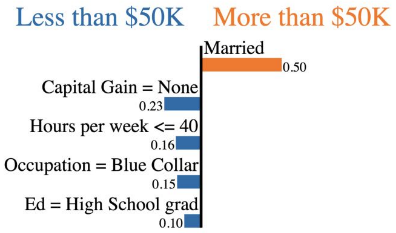


Ribeiro el al, 2018

Overly Specific Anchors

$28 < \text{Age} \leq 37$
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week ≤ 40.00
Country = United-States
$P(\textbf{Salary} > \$50\textbf{K}) = 0.57$

(a) Instance and prediction



(b) LIME explanation

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 > $50K

(c) An *anchor* explanation

Ribeiro el al, 2018

Project Review

- Project Proposal Due Apr 22
  - Up to 1.5 pages
  - The problem you are solving
  - Datasets
  - Metrics
  - Baselines

- Use the Slack Channel to Find Partners
  - https://cs335-2020sp.slack.com/archives/C0120BNJJHW

- Google Cloud Credits

Required Reading

Molnar: Ch 5.7, Ch 5.8

Reading Assignments (Pick One)

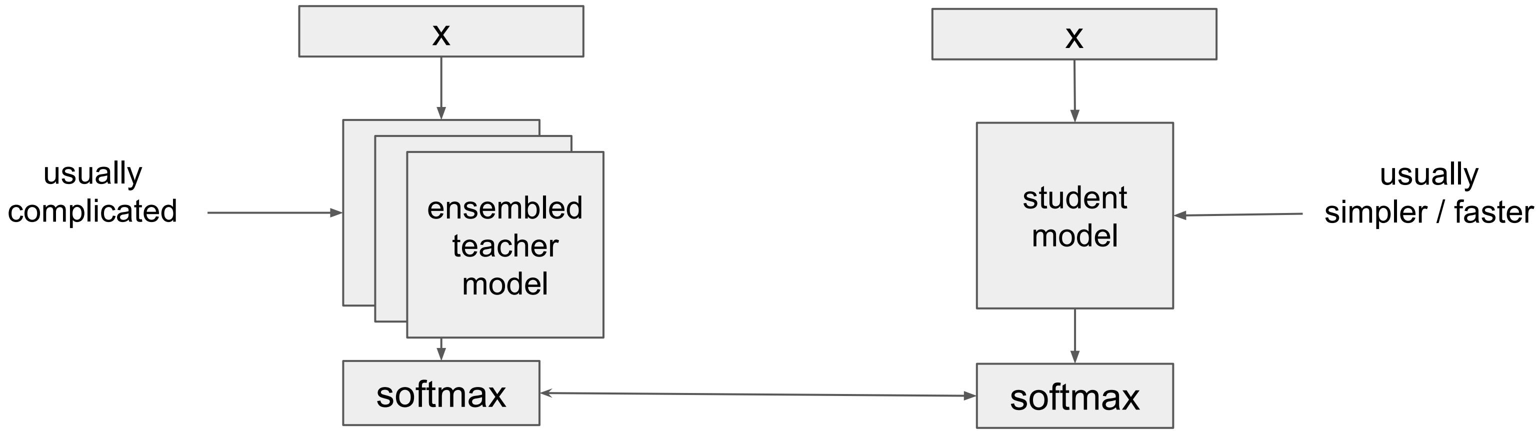
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier, SIGKDD 2016
- Lakkaraju, Himabindu, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction, SIGKDD 2016
- Che, Z., Purushotham, S., Khemani, R., & Liu, Y. Distilling knowledge from deep networks with applications to healthcare domain, Arxiv 2015
- Plumb, Gregory, Denali Molitor, and Ameet S. Talwalkar. Model agnostic supervised local explanations, NeurIPS 2018
- Robnik-Šikonja, M., & Kononenko, I. Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering, 2008

Next Lecture

Feature Interactions for Interpretability

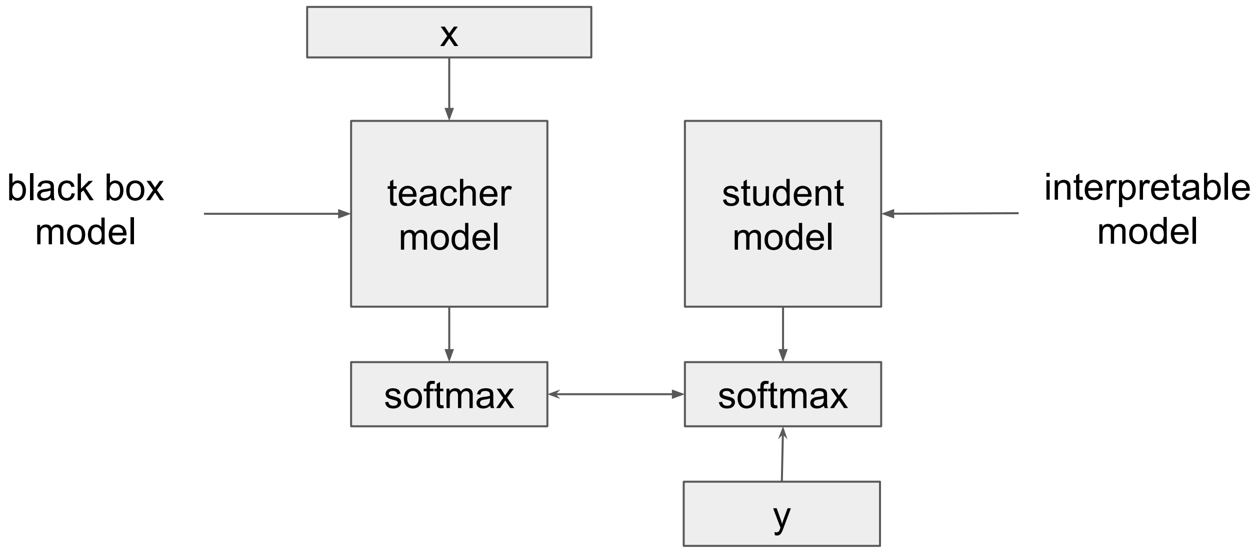
Knowledge Distillation (Hinton et al, 2015)

- Distillation of Neural Networks
  - use a simple network to approximate the more complicated ones
  - applications: improve performance (matching ensemble models), improve latency



DarkSight (Xu et al, 2018)

- Teacher - Student Architecture
  - match the softmax output between the teacher model and the student model
  - $P_T(k \mid x) \sim P_S(k \mid y, \Theta)$



DarkSight (Xu et al, 2018)

- Optimization
  - match the distribution of the softmax layer
  - D is a divergence measure

$$L(Y,\theta) = \frac{1}{N}\sum_{i=1}^{N} D(P_T(\cdot|x_i), P_S(\cdot|y_i;\theta))$$

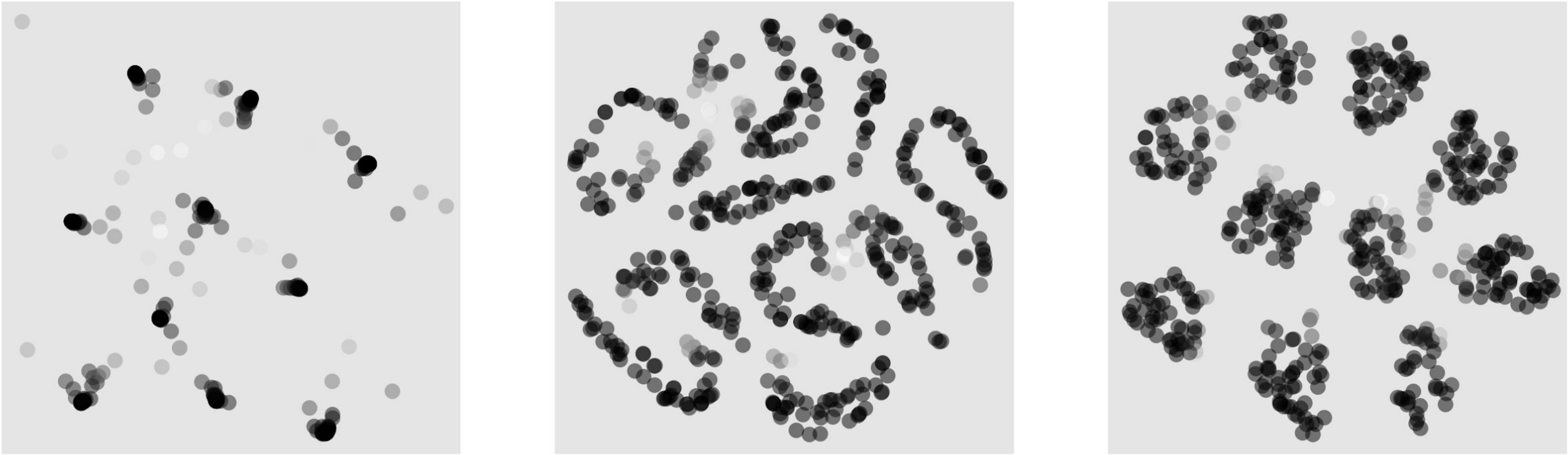
  - One such choice can be the symmetric KL

$$KL_{sym}(P,Q) = \frac{1}{2}(KL(P,Q) + KL(Q,P))$$

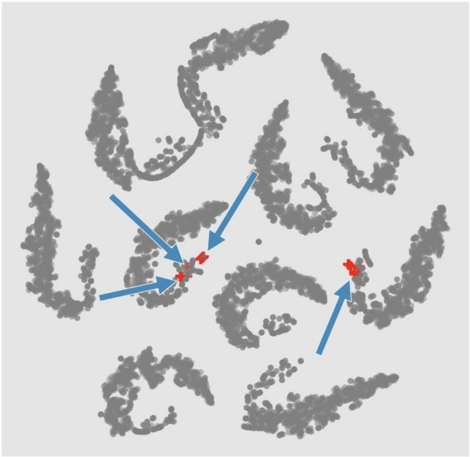
DarkSight (Xu et al, 2018)

- Interpretable Model
  - Naive Bayes Classifier

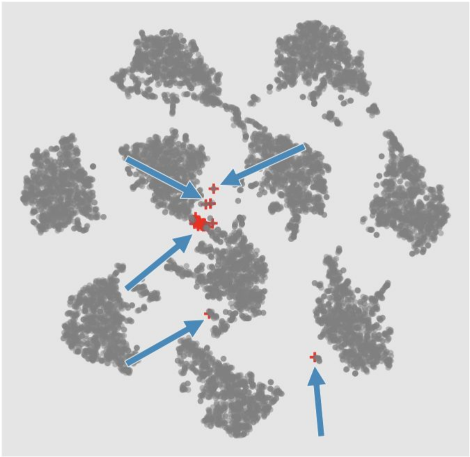
$$P_S(c_i = k | y_i; \theta) = \frac{P(y_i | c_i = k; \theta_c) P(c_i = k; \theta_p)}{P(y_i | \theta)}$$



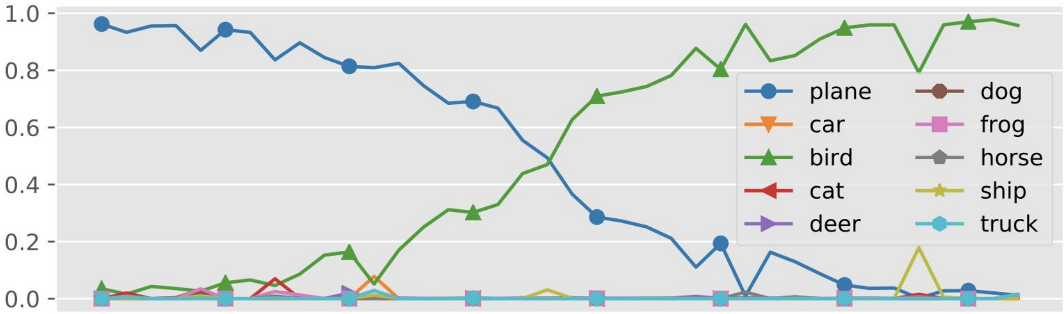
(a) DarkSight (b) t-SNE prob (c) t-SNE logit



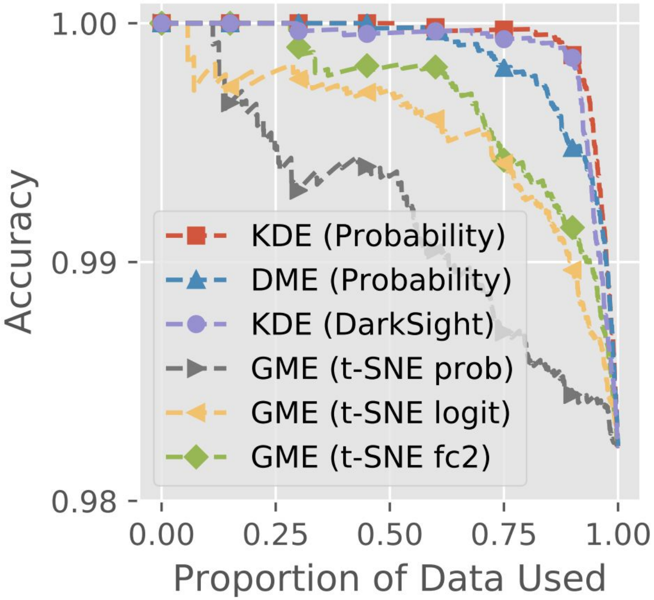
(a) t-SNE prob



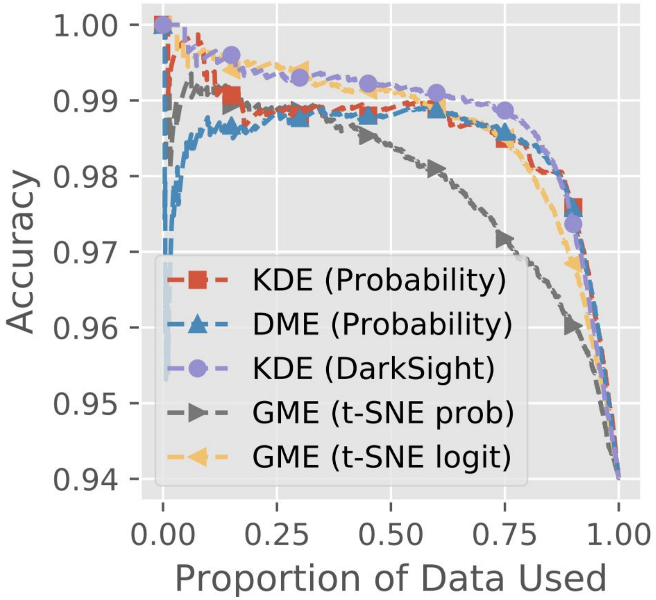
(b) t-SNE logit



(c) Predictive probabilities of points in the black box

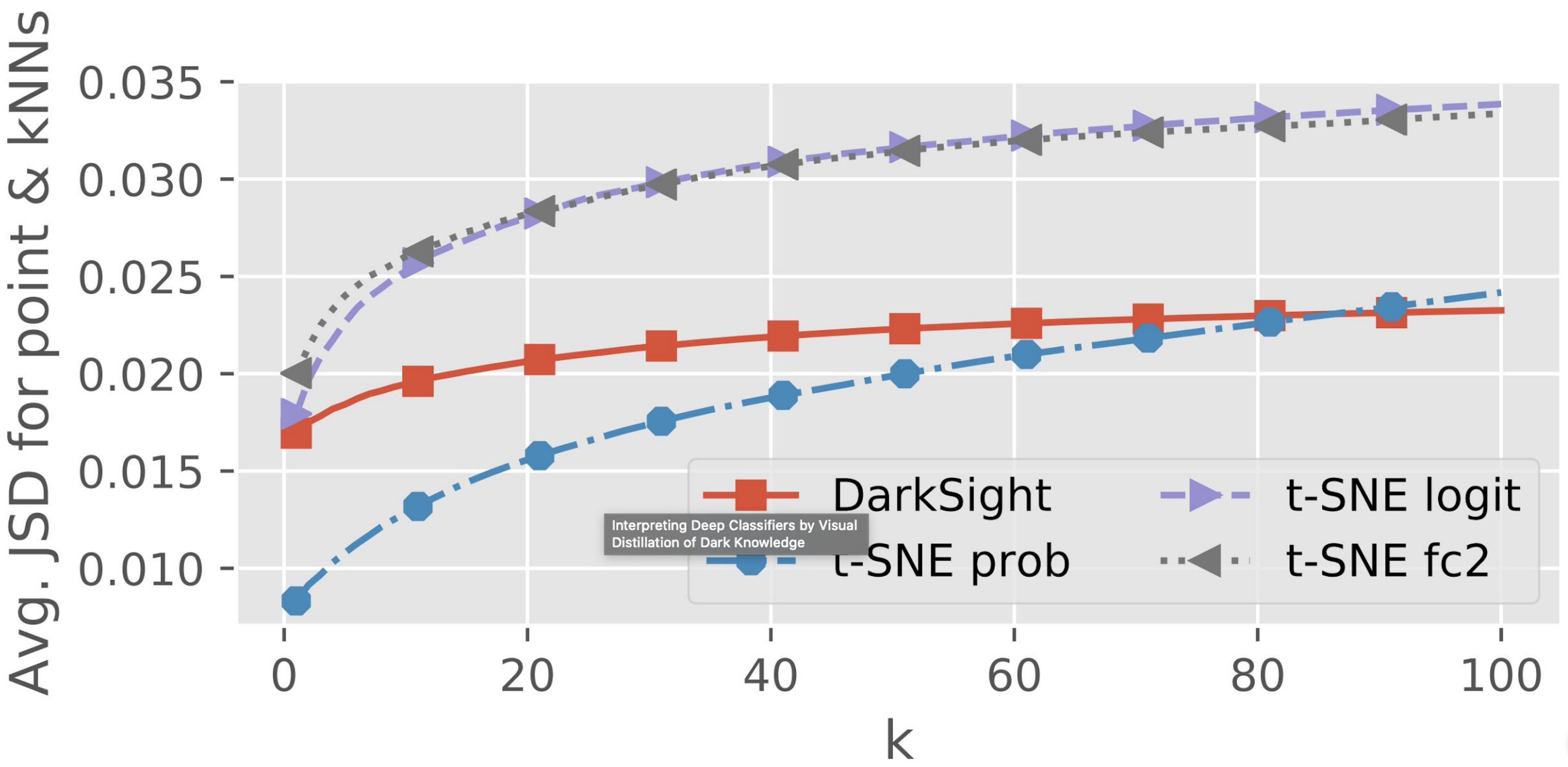


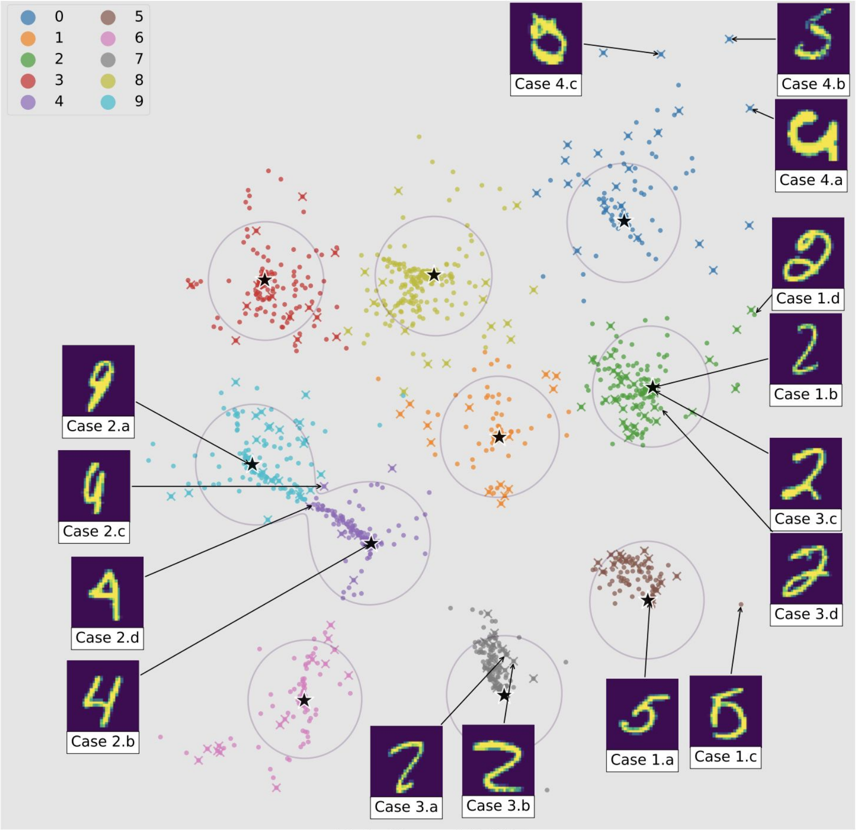
(a) MNIST



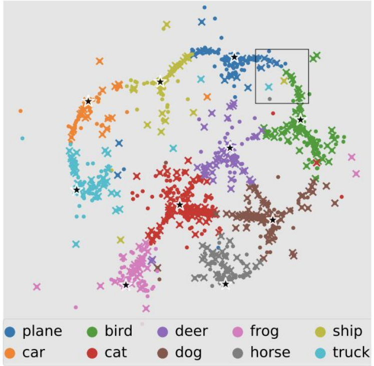
(b) Cifar10

$$M_k(Y) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{k} \sum_{j \in \mathrm{NN}_k(y_i)} JSD(p_i, p_j)$$

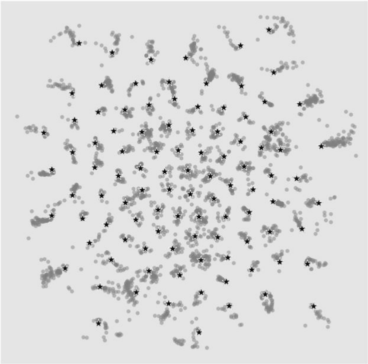

Avg. JSD for point & kNNs
0.035
0.030
0.025
0.020
0.015
0.010
0
20
40
60
80
100
k
DarkSight
t-SNE logit
t-SNE prob
t-SNE fc2
Interpreting Deep Classifiers by Visual Distillation of Dark Knowledge



(a) LeNet on MNIST



(b) VGG16 on Cifar10



(c) Wide-ResNet on Cifar100