Interpretable Machine Learning

2021/01/26

Ping-Han Hsieh

Disclaimer

- Black box machine learning model help us develop rule-based theories
- Rule-based theories help us refine black box machine learning model

Outline

- What is interpretable machine learning
  - Examples to solve high dimensional regression problem.
    - Principal Component Analysis
    - Factor Analysis
  - Do problems affect interpretability
- Properties of Interpretable Machine Learning
  - Transparency
  - Post-hoc Interpretability

What is Interpretable Machine Learning (1)

- What should be interpretable or explainable in modeling
  - Relation between input and output

$$\mathbf{y} = w_1\mathbf{x_1} + w_2\mathbf{x_2}$$

  - How the model is optimized

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

What is Interpretable Machine Learning (2)

- What should be interpretable or explainable in modeling
  - Relation between input and output

$$\mathbf{y} = w_1\mathbf{x_1} + w_2\mathbf{x_2} + \cdots + w_{512}\mathbf{x}_{512} \quad (p \gg n)$$

  - How the model is optimized

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (failed)$$

  - Principal component analysis
  - Factor analysis

Principal Component Analysis

- What should be interpretable or explainable in modeling
  - How the model is optimized (PCA)

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$\begin{aligned}\mathbf{\Sigma} &= \frac{1}{n}\mathbf{X^T}\mathbf{X} \\ &= \frac{1}{n}\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T \\ &= \frac{1}{n}\mathbf{V}\mathbf{S}^2\mathbf{V}^T\end{aligned}$$

$$\mathbf{P} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{S}$$

$$\text{(take the first K PCs based on variance): } \mathbf{P}^*$$

  - How the model is optimized (Regression)

$$\mathbf{w} = (\mathbf{P}^{*T}\mathbf{P}^*)^{-1}\mathbf{P}^{*T}\mathbf{y}$$

  - Relation between input and output

$$\mathbf{y} = w_1\mathbf{p_1} + w_2\mathbf{p_2}$$

What is Interpretable Machine Learning (3)

- What should be interpretable or explainable in modeling
  - Relation between input and output

    $$\mathbf{y} = w_1\mathbf{x_1} + w_2\mathbf{x_2} + \cdots + w_{512}\mathbf{x}_{512} \quad (p \gg n)$$

  - How the model is optimized

    $$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (failed)$$

  - Principal component analysis
  - Factor analysis

Factor Analysis (1)

- What should be interpretable or explainable in modeling
  - How the model is optimized (FA)

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$\mathbf{X} = \mathbf{F}\mathbf{w} + \epsilon \quad \text{(assumption)}$$

$$\begin{aligned}
\Sigma &= \frac{1}{n}\mathbf{X}^T\mathbf{X} \\
&= \frac{1}{n}(\mathbf{F}\mathbf{w} + \epsilon)^T(\mathbf{F}\mathbf{w} + \epsilon) \\
&= \frac{1}{n}(\mathbf{w}^T\mathbf{F}^T + \epsilon^T)(\mathbf{F}\mathbf{w} + \epsilon) \\
&= \frac{1}{n}\{\mathbf{w}^T\mathbf{F}^T\mathbf{F}\mathbf{w} + \mathbf{w}^T\mathbf{F}^T\epsilon + \epsilon^T\mathbf{F}\mathbf{w} + \epsilon^T\epsilon\} \\
&= \frac{1}{n}\{\mathbf{w}^T\mathbf{w} + 0 + 0 + \epsilon^T\epsilon\} \\
&= \frac{1}{n}\mathbf{w}^T\mathbf{w} + \mathbf{\Psi}
\end{aligned}$$

$$\mathbf{U} = \mathbf{\Sigma} - \mathbf{\Psi} \quad \text{(positive symmetric)}$$

$$\begin{aligned}
\mathbf{U} &= \mathbf{C}\mathbf{D}\mathbf{C}^T \quad \text{(take a few eigenvectors)} \\
&\sim \mathbf{C}^*\mathbf{D}^*\mathbf{C}^{*T} \\
&= \mathbf{C}^*\mathbf{D}^{*1/2}\mathbf{D}^{*1/2}\mathbf{C}^{*T} \\
&= (\mathbf{C}^*\mathbf{D}^{*1/2})(\mathbf{C}^*\mathbf{D}^{*1/2})^T
\end{aligned}$$

$$\mathbf{w}^T\mathbf{w} \sim (\mathbf{C}^*\mathbf{D}^{*1/2})(\mathbf{C}^*\mathbf{D}^{*1/2}) \qquad (\mathbf{Q}\mathbf{w})^T(\mathbf{Q}\mathbf{w}) \text{ if } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$$

solve $\mathbf{F}$ with $\mathbf{w}$

but how could we get ϵ (or Ψ) in the first place

What is Interpretable Machine Learning

Factor Analysis (2)

- What should be interpretable or explainable in modeling
  - How the model is optimized (FA)

$$\mathbf{X} = \mathbf{USV}^T$$

$$\mathbf{X} = \mathbf{Fw} + \epsilon \quad \text{(assumption)}$$

$$\begin{aligned}
\Sigma &= \frac{1}{n}\mathbf{X}^T\mathbf{X} \\
&= \frac{1}{n}(\mathbf{Fw} + \epsilon)^T(\mathbf{Fw} + \epsilon) \\
&= \frac{1}{n}(\mathbf{w}^T\mathbf{F}^T + \epsilon^T)(\mathbf{Fw} + \epsilon) \\
&= \frac{1}{n}\{\mathbf{w}^T\mathbf{F}^T\mathbf{Fw} + \mathbf{w}^T\mathbf{F}^T\epsilon + \epsilon^T\mathbf{Fw} + \epsilon^T\epsilon\} \\
&= \frac{1}{n}\{\mathbf{w}^T\mathbf{w} + 0 + 0 + \epsilon^T\epsilon\} \\
&= \frac{1}{n}\mathbf{w}^T\mathbf{w} + \mathbf{\Psi}
\end{aligned}$$

$$\mathbf{U} = \mathbf{\Sigma} - \mathbf{\Psi} \quad \text{(positive symmetric)}$$

$$\begin{aligned}
\mathbf{U} &= \mathbf{CDC}^T \quad \text{(take a few eigenvectors)} \\
&\sim \mathbf{C}^*\mathbf{D}^*\mathbf{C}^{*T} \\
&= \mathbf{C}^*\mathbf{D}^{*1/2}\mathbf{D}^{*1/2}\mathbf{C}^{*T} \\
&= (\mathbf{C}^*\mathbf{D}^{*1/2})(\mathbf{C}^*\mathbf{D}^{*1/2})^T
\end{aligned}$$

$$\mathbf{w}^T\mathbf{w} \sim (\mathbf{C}^*\mathbf{D}^{*1/2})(\mathbf{C}^*\mathbf{D}^{*1/2}) \qquad (\mathbf{Qw})^T(\mathbf{Qw}) \ \text{if } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$$

solve $\mathbf{F}$ with $\mathbf{w}$

but how could we get ϵ (or Ψ) in the first place

assume $F_i \sim N(0,1) \rightarrow X_i \sim N(0, \Psi + \mathbf{w}^T\mathbf{w})$

$$L = \frac{-np}{2}\log 2\pi - \frac{\pi}{2}\log|\mathbf{\Psi} + \mathbf{w}^T\mathbf{w}| - \frac{n}{2}\mathrm{tr}((\Psi + \mathbf{w}^T\mathbf{w})^{-1}\mathbf{\Sigma})$$

starts with a guess about the unique variances

iterates to convergence

What is Interpretable Machine Learning (4)

- What should be interpretable or explainable in modeling
  - Relation between input and output

$$\mathbf{y} = w_1\mathbf{x_1} + w_2\mathbf{x_2} + \cdots + w_{512}\mathbf{x}_{512} \quad (p \gg n)$$

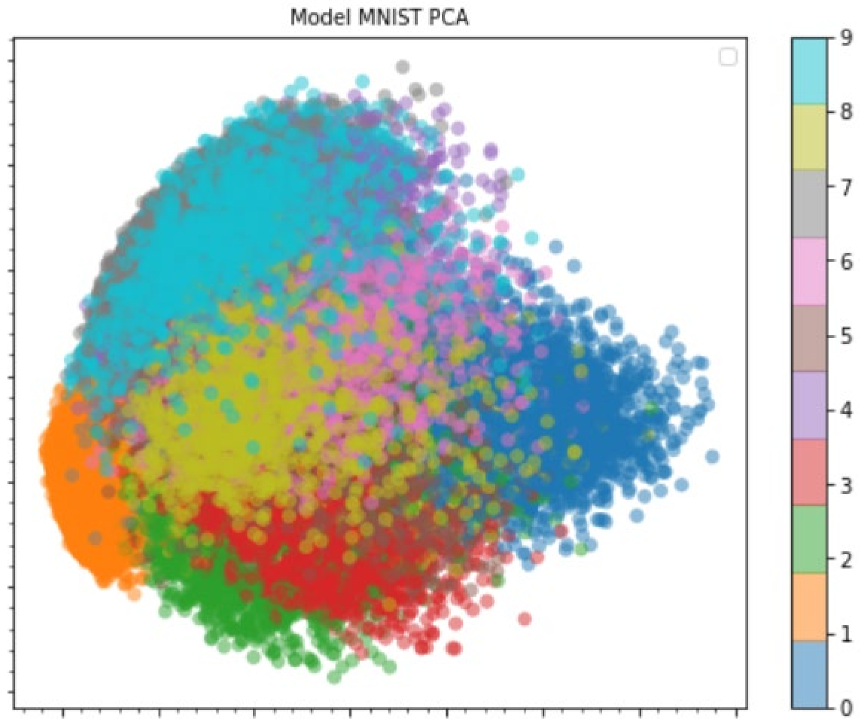
  - How the model is optimized

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (failed)$$

  - Principal component analysis
  - Factor analysis
- What if the features are related in a non-Eucleadian way

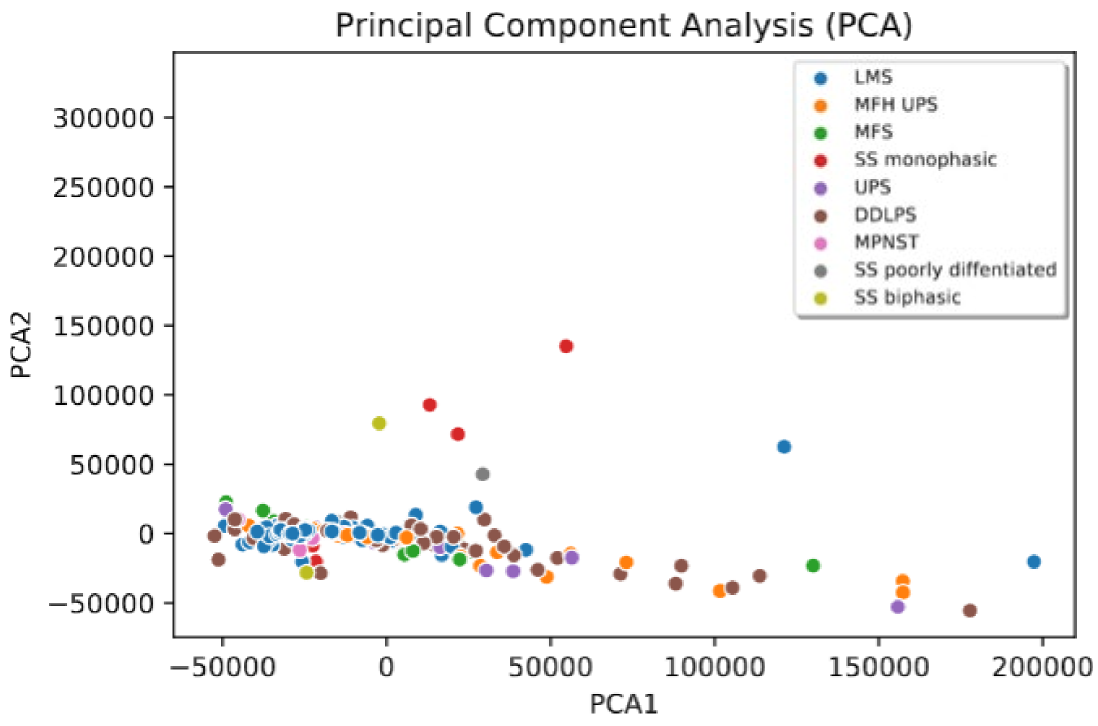
Do problems affect interpretability

How Could We Interpret the Result (1)



PCA on MNIST

(from Zeta Learning)

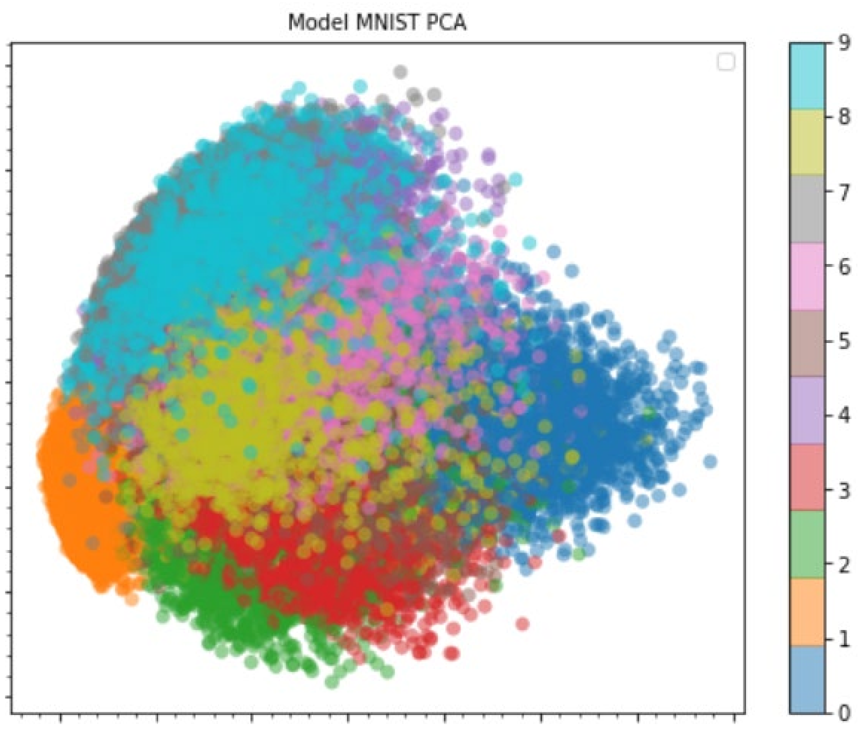


PCA on Network Edges for Sarcoma

(from in-house analysis by T. Belova)

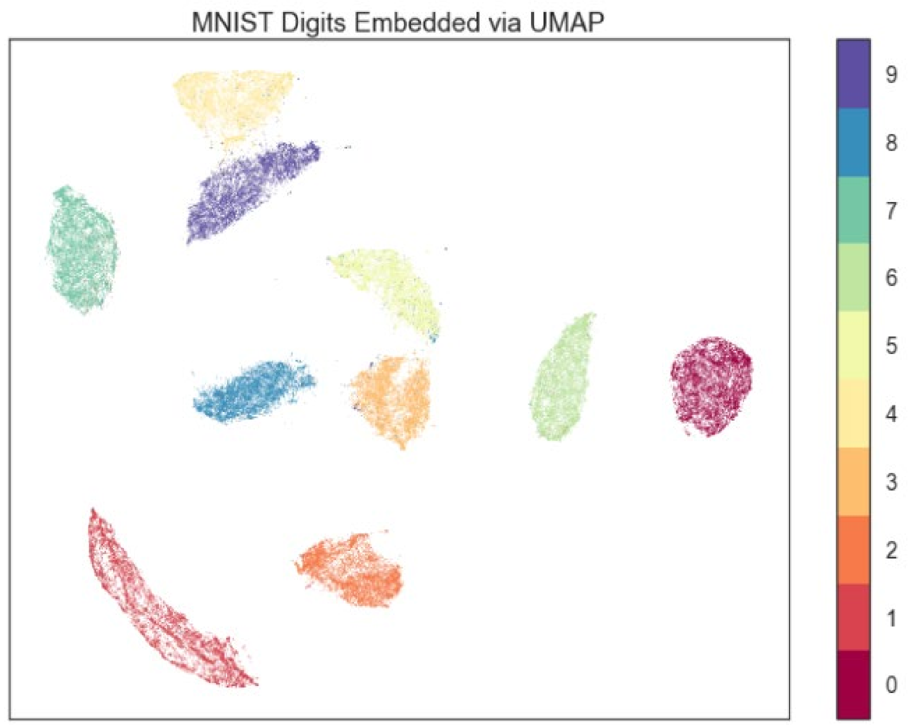
Do problems affect interpretability

How Could We Interpret the Result (2)



PCA on MNIST

(from Zeta Learning)



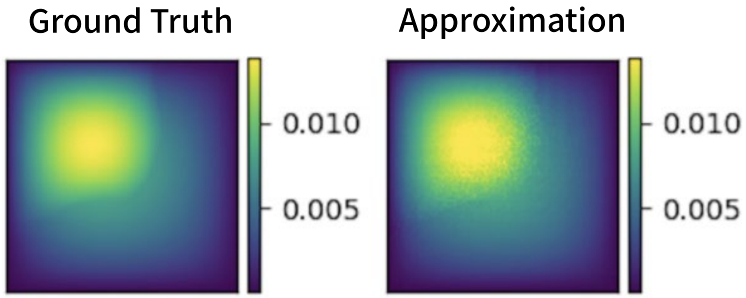
UMAP on MNIST

McInnes et al. (2018)

Do problems affect interpretability

Partial Differential Equation

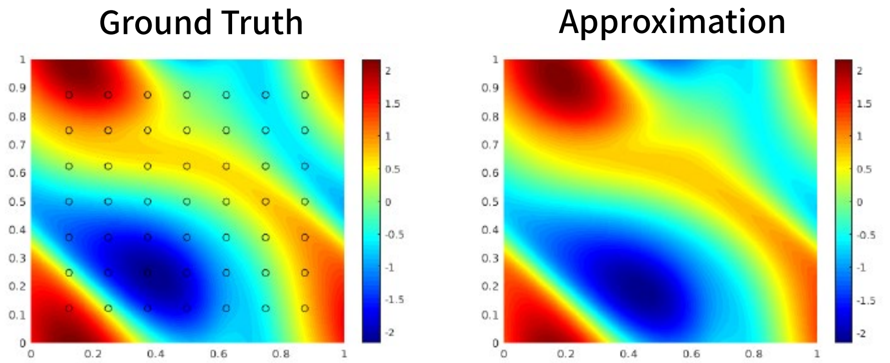
Neural Operator



$$
\begin{aligned}
(\mathcal{L}_a u)(x) &= f(x), \qquad && x \in D \\
u(x) &= 0, && x \in \partial D
\end{aligned}
$$

Li *et al.* (2020)

Fourier Neural Operator



$$
\begin{aligned}
\partial_t w(x,t) + u(x,t) \cdot \nabla w(x,t) &= \nu \Delta w(x,t) + f(x), \quad && x \in (0,1)^2, t \in (0,T] \\
\nabla \cdot u(x,t) &= 0, && x \in (0,1)^2, t \in [0,T] \\
w(x,0) &= w_0(x), && x \in (0,1)^2
\end{aligned}
$$

Li *et al.* (2020)

Properties of Interpretable Machine Learning

- Transparency (Model)
  - Simulatability
  - Decomposability
  - Algorithmic transparency

- Post-hoc Interpretability (Problem)
  - Local Explanation
  - Explain by example

Lipton (2017)

Transparency

- Simulatability
  - Whether the computation can be readily reproduced by human.
  - Depends not only on the model, but also on the dimensionality of data.
- Decomposability
  - Intuitive explanation for every part of the model.
  - We need to include feature engineering and anonymous features.
- Algorithmic transparency
  - Relation between input and output.
  - How the model is optimized.
  - Human lack of this as well.

Post-hoc Interpretability

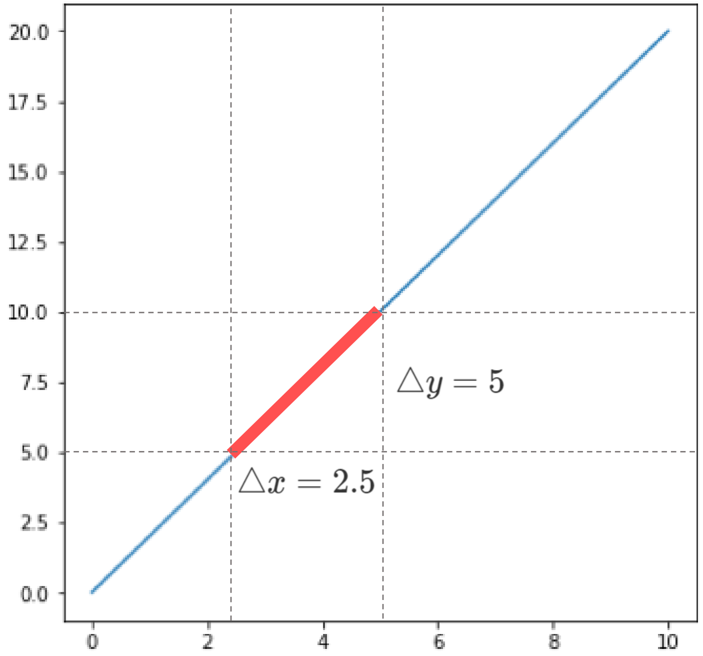
Gradient Based (1)

- What does the coefficient in linear regression mean to the model

$$f(x) = y = 2x$$

- How much the model output will change when the input is change.

$$\frac{\triangle y}{\triangle x} = 2 \rightarrow \nabla_x f(x) = 2$$



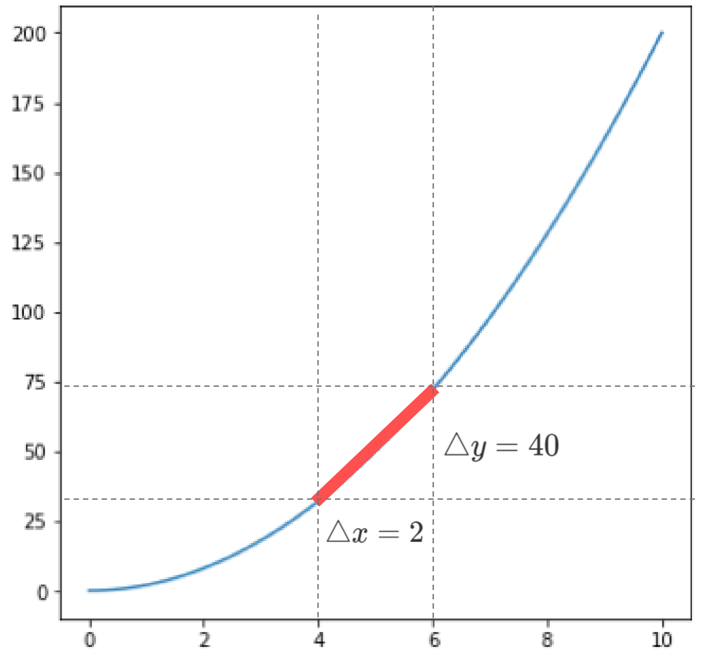
Post-hoc Interpretability

Gradient Based (2)

- How about non-linear function

$$f(x) = y = 2x^2$$

- How much the model output will change when the input is change.



Post-hoc Interpretability

Gradient Based (3)

- How about non-linear function

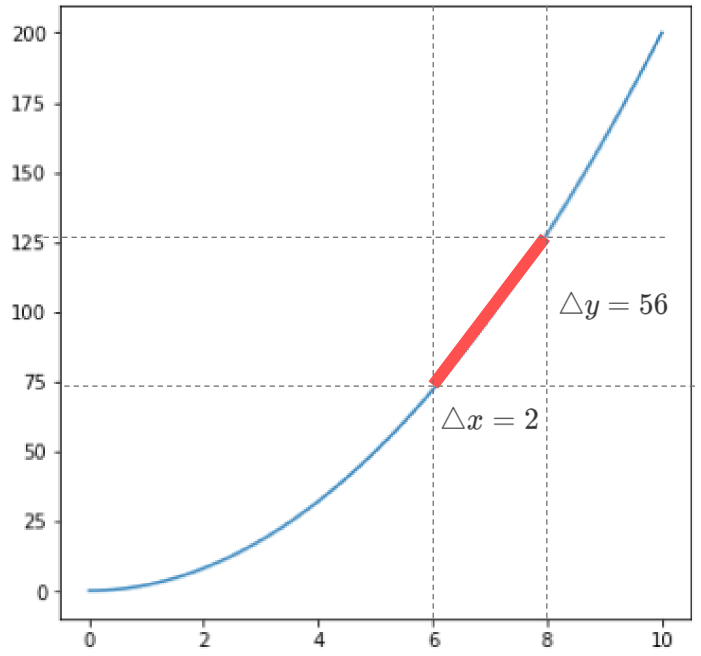
$$f(x) = y = 2x^2$$

- How much the model output will change when the input is change.

$$\nabla_x f(x) = 4x$$

- Generalize to higher dimension

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = f'(\mathbf{x})$$



Post-hoc Interpretability

Gradient Based (4)

- Gradient (Simonyan *et al.* 2014)

  $\nabla_{\mathbf{x}} f(\mathbf{x})$

- Implementation Invariance

- Sensitivity

  (1)

  if input and baseline differ only in one feature but have different predictions

  $\Downarrow$

  the contribution for that feature is non-zero

  (2)

  $$x'_j = 0 \Rightarrow \phi_i = 0$$

- Completeness

  $$\sum \psi = f(\mathbf{x}) - f(\mathbf{x}^{(b)})$$

Gradient Based (5)

- Gradient (Simonyan *et al.* 2014)

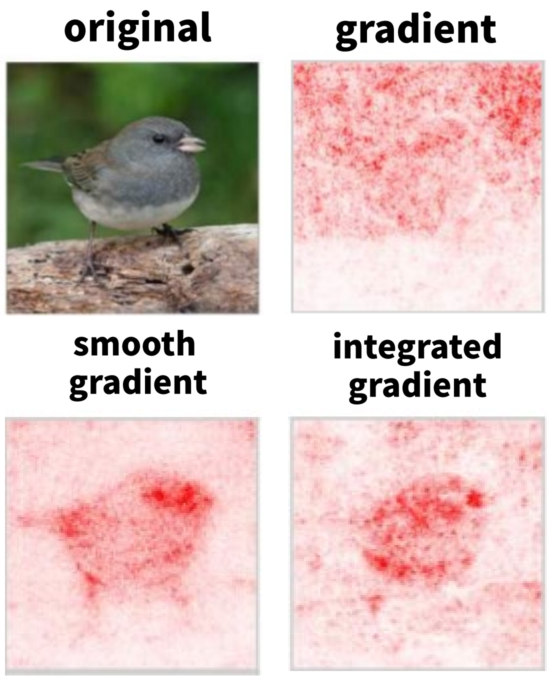
$$\nabla_{\mathbf{x}} f(\mathbf{x})$$

- Smooth gradient (Smilkov *et al.* 2017)

$$\frac{1}{N}\sum_{i=1}^{n} \nabla_{\mathbf{x}+\epsilon} f(\mathbf{x}+\epsilon)$$

- Integrated gradient (Sundararajan *et al.* 2017)

$$(\mathbf{x}-\mathbf{x}^{(b)}) \times \int_0^1 \nabla_{\mathbf{x}} f(\mathbf{x}^{(b)} + \alpha(\mathbf{x}-\mathbf{x}^{(b)}))d\alpha$$



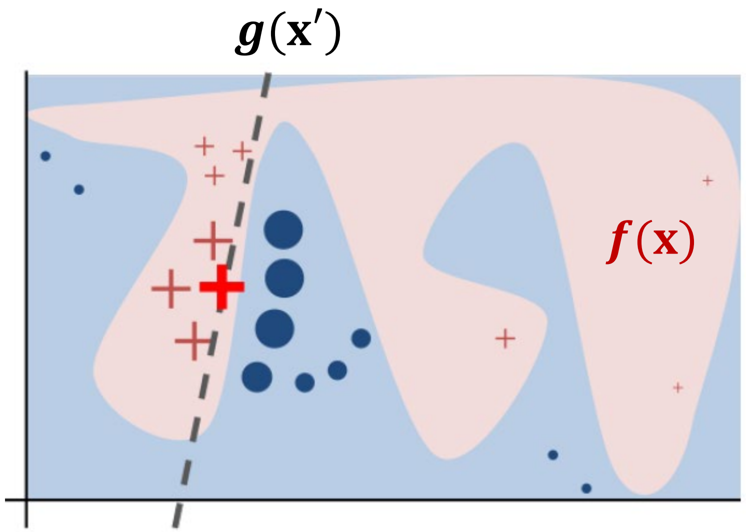
- ReLU activation function:
  - Guided Backpropagation (Springenberg et al., 2014)

- Convolutional Neural Network:
  - Deconvolutional Network (Zeiler & Fergus, 2014)
  - Grad-CAM (Selvaraju et al., 2016)
  - Grad-CAM++ (Chattopadhyay et al., 2018)

Surrogate Model - SHAP (1)

- Use a simple linear additive model to explain the complex model.

- The simplified model needs to follow
  1. Local Accuracy
  $$f(\mathbf{x}^{(\mathbf{i})}) = g_i(\mathbf{x}'^{(\mathbf{i})}) = \phi_0 + \sum_{j=1}^{n} \phi_j x'_j$$
  2. Missingness
  $$x'_j = 0 \Rightarrow \phi_i = 0$$
  3. Consistency
  $$f'(\mathbf{z}'^{(\mathbf{i})}) - f'(\mathbf{z}'^{(\mathbf{i})} \backslash j) \geq f(\mathbf{z}'^{(\mathbf{i})}) - f(\mathbf{z}'^{(\mathbf{i})} \backslash j)$$
  $$\Downarrow$$
  $$\phi_j(f', x) \geq \phi_j(f, x).$$



Ribeiro (2016)

- Exact solution

$$\phi_i = \sum_{S \subseteq F} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(\mathbf{x}_{S \cup \{j\}}) - f_S(\mathbf{x}_S)]$$
$$= \sum_{S \subseteq F} \frac{1}{|F|} \frac{1}{\binom{|F|-1}{|S|}} [f_{S \cup \{j\}}(\mathbf{x}_{S \cup \{j\}}) - f_S(\mathbf{x}_S)]$$

Shapley (1951)

Post-hoc Interpretability

Surrogate Model - SHAP (2)

- The exact solution is computationally intensive, the sampling approximation is usually used.

Exact Solution

$$\sum_{S\subseteq F} \frac{1}{|F|} \frac{1}{\binom{|F|-1}{|S|}} [f_{S\cup\{j\}}(\mathbf{x}_{S\cup\{j\}}) - f_S(\mathbf{x}_S)]$$

1. Compute the output

$f(A) = 0.85$
$f(B) = 1.00$
$f(C) = 0.25$
$f(A,B) = 1.25$
$f(A,C) = 0.50$
$f(B,C) = 0.75$
$f(A,B,C) = 0.95$

2. Compute the contribution

$$\phi_A = \frac{1}{3}\cdot\frac{1}{1}\cdot(0.85-0.00) + \frac{1}{3}\cdot\frac{1}{2}\cdot(1.25-1.00) + \frac{1}{3}\cdot\frac{1}{2}\cdot(0.50-0.25) + \frac{1}{3}\cdot\frac{1}{1}\cdot(0.95-0.75)$$

Approximate Solution

$$\phi_i = \frac{1}{|F|!} \sum_{S\in\pi(F)} [f_{p_j(S)\cup\{j\}}(\mathbf{x}_{p_j(S)\cup\{j\}}) - f_{p_j(S)}(\mathbf{x}_{p_j(S)})]$$

1. Random permutations

$$\pi(F) = \{ABC, ACB, BAC, BCA, CAB, CBA\}$$

2. Take elements precede the target

$$\{\varnothing, \varnothing, B, BC, C, CB\}$$

3. Compute contribution

$$\phi_A = \frac{1}{6}\cdot(0.85-0.00) + \frac{1}{6}\cdot(0.85-0.00) + \frac{1}{6}\cdot(1.25-1.00) + \frac{1}{6}\cdot(0.50-0.25) + \frac{1}{6}\cdot(0.95-0.75) + \frac{1}{6}\cdot(0.95-0.75)$$

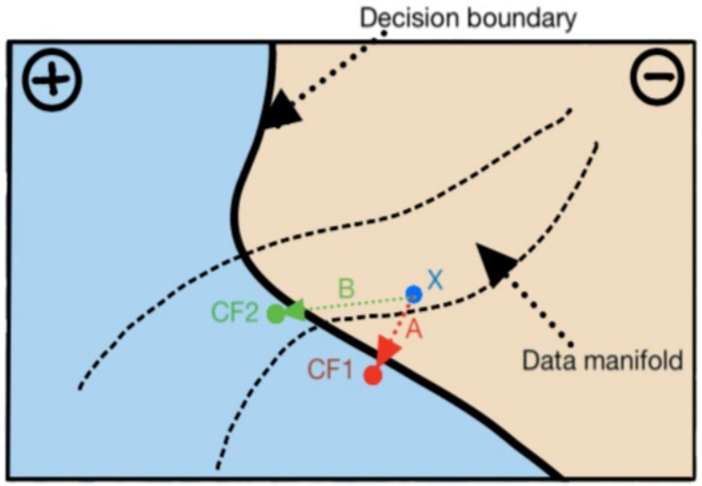
Sampling from the permutation

$$\phi_i = \frac{1}{K}\sum_{k=1}^{K} V_k$$

Štrumbelj et al. (2014)

Counterfactual Explanation

- How much the input needs to be changed in order to get the desired output?



Verma et al. 2020

$$\text{argmin}_{\mathbf{x}'} d(\mathbf{x}, \mathbf{x}') \quad s.t. \quad f(\mathbf{x}') = \mathbf{y}'$$

$$\Downarrow$$

$$\text{argmin}_{\mathbf{x}'} \lambda(f(\mathbf{x}') - \mathbf{y}') + d(\mathbf{x}, \mathbf{x}')$$

Wachter *et al.,* (2018)

$$\text{argmin}_{\mathbf{x} \in A} cost(\mathbf{x}, \mathbf{x}') \quad s.t. \quad f(\mathbf{x}') = \mathbf{y}'$$

Ustun *et al.,* (2019)

Discussion

- Interpretation and causality.
- Interpretation and sensitivity analysis.
- What kind of properties you think are important to your research.
- Tradeoff between accuracy and interpretability.
- Can we trust post-hoc explanation.
- Hyperparameter and interpretation.

The Great AI Debate, NeurIPS 2017

Thanks