Evaluating and Aggregating Feature-based Model Explanations

Umang Bhatt

Carnegie Mellon University and University of Cambridge

Joint Work with Adrian Weller (Cambridge) and José Moura (CMU)

usb20@cam.ac.uk

Overview

1. Why are feature level explanations important (in medicine)?
2. What are existing feature-based explanation techniques?
3. How do we evaluate feature-based explanations?
4. How do we aggregate feature-based explanations?
5. Future Work

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

- Doctors leverage **vital signs** and **indicators** to diagnose

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

- Doctors leverage **vital signs** and **indicators** to diagnose
  - This translates to a semantically meaningful feature vector $\boldsymbol{x}$

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

- Doctors leverage **vital signs** and **indicators** to diagnose
  - This translates to a semantically meaningful feature vector $\boldsymbol{x}$
- Doctors leverage insights from **past patients** to better diagnose current patients

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

- Doctors leverage **vital signs** and **indicators** to diagnose
  - This translates to a semantically meaningful feature vector $\boldsymbol{x}$
- Doctors leverage insights from **past patients** to better diagnose current patients
  - This translates to influential points in the training distribution

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

- Doctors leverage **vital signs** and **indicators** to diagnose
  - This translates to a semantically meaningful feature vector $\boldsymbol{x}$
- Doctors leverage insights from **past patients** to better diagnose current patients
  - This translates to influential points in the training distribution

Medical Diagnostic Tasks

If we understand how a doctor reasons to a diagnosis, then we can build models that mimic that decision making.

- Doctors leverage **vital signs** and **indicators** to diagnose
  - This translates to a semantically meaningful feature vector $\boldsymbol{x}$
- Doctors leverage insights from **past patients** to better diagnose current patients
  - This translates to influential points in the training distribution

We ask: can we validate the above intuition on a trained predictor, $\boldsymbol{f}$, using feature importance and sample importance?

Common feature-based explanations

SHAP (Lundberg and Lee. NeurIPS 2017)

$$\boldsymbol{g}_{ci} = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \phi_i = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{i\}} \binom{F-1}{S}^{-1} \left(\boldsymbol{f}(\boldsymbol{x}_{S \cup \{i\}}) - \boldsymbol{f}(\boldsymbol{x}_S)\right)$$

Provides a "fair" distribution of contribution over all features, since Shapley values satisfy **efficiency**, symmetry, additivity, and dummy (zero).

Common feature-based explanations

SHAP (Lundberg and Lee. NeurIPS 2017)

$$\boldsymbol{g}_{ci} = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \phi_i = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{i\}} \binom{F-1}{S}^{-1} \left(\boldsymbol{f}(\boldsymbol{x}_{S \cup \{i\}}) - \boldsymbol{f}(\boldsymbol{x}_S)\right)$$

Provides a "fair" distribution of contribution over all features, since Shapley values satisfy **efficiency**, symmetry, additivity, and dummy (zero).

Integrated Gradients (Sundarajan et al. ICML 2017)

Accumulates the gradients along a straight line path between $\boldsymbol{x}$ and $\bar{x}$, where $\boldsymbol{f}(\bar{x}) \approx 0$, and satisfies **completeness**, $\sum_{i=1}^{d} \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(x)$.

Common feature-based explanations

SHAP (Lundberg and Lee. NeurIPS 2017)

$$\boldsymbol{g}_{ci} = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \phi_i = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{i\}} \binom{F-1}{S}^{-1} \left(\boldsymbol{f}(\boldsymbol{x}_{S \cup \{i\}}) - \boldsymbol{f}(\boldsymbol{x}_S)\right)$$

Provides a "fair" distribution of contribution over all features, since Shapley values satisfy **efficiency**, symmetry, additivity, and dummy (zero).

Integrated Gradients (Sundarajan et al. ICML 2017)

Accumulates the gradients along a straight line path between $\boldsymbol{x}$ and $\bar{x}$, where $\boldsymbol{f}(\bar{x}) \approx 0$, and satisfies **completeness**, $\sum_{i=1}^{d} \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(x)$.

LIME (Ribeiro et al. KDD 2016)

$$\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \underset{\boldsymbol{g} \in \mathcal{G}}{\arg\min}\ \mathcal{L}(\boldsymbol{f}, \boldsymbol{g}, \pi_x) + \Omega(\boldsymbol{g})$$

Local surrogate model, $\boldsymbol{g}$, to approximate original model, f, in some kernelized region π_x, and encourages **sparsity** by keeping model complexity, $\Omega(\boldsymbol{g})$, low

Evaluating explanations

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z}) d\boldsymbol{z}$$

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z}) d\boldsymbol{z}$$

$$\mu_{\text{MAX}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \max_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))$$

Let D be the distance between explanations and ρ be the distance between inputs

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z})d\boldsymbol{z}$$

$$\mu_{\text{MAX}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \max_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))$$

Let D be the distance between explanations and ρ be the distance between inputs

Faithfulness

Does the explanation capture features important to the prediction?

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z})\leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z})d\boldsymbol{z}$$

$$\mu_{\text{MAX}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \max_{\rho(\boldsymbol{x},\boldsymbol{z})\leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))$$

Let D be the distance between explanations and ρ be the distance between inputs

Faithfulness

Does the explanation capture features important to the prediction?

$$\mu_{\text{F}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, S) = \text{corr}\left(\frac{1}{|S|}\sum_{i\in S}\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i, \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{[\boldsymbol{x}_s=\bar{\boldsymbol{x}}_s]})\right)$$

Fix a subset size and randomly sample subsets of that size from $\boldsymbol{x}$ to estimate the Pearson Correlation Coefficient

Evaluating explanations (cont.)

Evaluating explanations (cont.)

Complexity

Is the explanation digestible?

Evaluating explanations (cont.)

Complexity

Is the explanation digestible? We define an attribution contribution distribution:

$$\mathbb{P}_A = \left\{ \frac{|\boldsymbol{g}(\boldsymbol{x})_1|}{\sum\limits_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \frac{|\boldsymbol{g}(\boldsymbol{x})_2|}{\sum\limits_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \ldots, \frac{|\boldsymbol{g}(\boldsymbol{x})_d|}{\sum\limits_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|} \right\}$$

$$\mu_C(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}) = H(\boldsymbol{x}) = \mathbb{E}_i\big[-\ln(\mathbb{P}_A(i))\big] = -\sum_{i=1}^{d} \mathbb{P}_A(i) \ln(\mathbb{P}_A(i))$$

Evaluating explanations (cont.)

Complexity

Is the explanation digestible? We define an attribution contribution distribution:

$$\mathbb{P}_A = \left\{ \frac{|\boldsymbol{g}(\boldsymbol{x})_1|}{\sum\limits_{j \in [d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \frac{|\boldsymbol{g}(\boldsymbol{x})_2|}{\sum\limits_{j \in [d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \ldots, \frac{|\boldsymbol{g}(\boldsymbol{x})_d|}{\sum\limits_{j \in [d]} |\boldsymbol{g}(\boldsymbol{x})_j|} \right\}$$

$$\mu_{\mathrm{C}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}) = H(\boldsymbol{x}) = \mathbb{E}_i\big[-\ln(\mathbb{P}_A(i))\big] = -\sum_{i=1}^{d} \mathbb{P}_A(i) \ln(\mathbb{P}_A(i))$$

The least complex explanation is one where $\boldsymbol{g}(\boldsymbol{x})_i = 1$ and the most complex explanation is one where $\boldsymbol{g}(\boldsymbol{x})_i = \frac{1}{d}$.

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Three candidate methods for $h(\cdot)$.

- Convex Combination: $\boldsymbol{g}_{agg} = w\boldsymbol{g}_1 + (1 - w)\boldsymbol{g}_2$

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Three candidate methods for $h(\cdot)$.

- Convex Combination: $\boldsymbol{g}_{agg} = w\boldsymbol{g}_1 + (1 - w)\boldsymbol{g}_2$
- Centroid Aggregation: $\boldsymbol{g}_{\text{agg}} \in \arg\min_{\boldsymbol{g} \in \mathcal{G}} \sum_{i=1}^{m} d(\boldsymbol{g}, \boldsymbol{g}_i)$

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Three candidate methods for $h(\cdot)$.

- Convex Combination: $\boldsymbol{g}_{agg} = w\boldsymbol{g}_1 + (1-w)\boldsymbol{g}_2$
- Centroid Aggregation: $\boldsymbol{g}_{\text{agg}} \in \arg\min_{\boldsymbol{g} \in \mathcal{G}} \sum_{i=1}^{m} d(\boldsymbol{g}, \boldsymbol{g}_i)$
- Bayesian Optimization: $\max_{\boldsymbol{g}_{agg} \in \mathcal{G}} \mu(\boldsymbol{g}_{agg})$ where

$$k(\boldsymbol{g}_i, \boldsymbol{g}_j) = \mathbb{E}_{x \sim \mathcal{D}_x}\left[k(\boldsymbol{g}_i(\boldsymbol{x}), \boldsymbol{g}_j(\boldsymbol{x}))\right] = \mathbb{E}_{x \sim \mathcal{D}_x}\left[e^{-\frac{1}{2}||\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_j(\boldsymbol{x})||^2}\right]$$

Generalized Aggregation

Convex Combination

$$\boldsymbol{g}_{agg} = w^T G$$

$$G^T = \begin{pmatrix} | & | & & | \\ \text{SHAP}_1 & \text{LIME}_2 & \cdots & \text{IG}_m \\ | & | & & | \end{pmatrix}$$

Generalized Aggregation

Convex Combination

$$\boldsymbol{g}_{agg} = w^T G$$

$$G^T = \begin{pmatrix} | & | & & | \\ \text{SHAP}_1 & \text{LIME}_2 & \cdots & \text{IG}_m \\ | & | & & | \end{pmatrix}$$

$$w_{agg} \in \arg\max_{w} \sum_{i=1}^{m} \mu(w^T G_i)$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \dots IG_m\}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \dots IG_m\}$$

$$\boldsymbol{g}_c^S = \begin{bmatrix}1 & -2 & 7\end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix}.1 & .2 & .7\end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix}C & B & A\end{bmatrix}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \dots IG_m\}$$

$$\boldsymbol{g}_c^S = \begin{bmatrix} 1 & -2 & 7 \end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix} .1 & .2 & .7 \end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix} C & B & A \end{bmatrix}$$

$$\text{rank}^{S_1} = \begin{bmatrix} C & B & A \end{bmatrix} \quad \text{rank}^{S_2} = \begin{bmatrix} C & A & B \end{bmatrix} \quad \text{rank}^{S_3} = \begin{bmatrix} A & B & C \end{bmatrix}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

$$\boldsymbol{g}_c^S = \begin{bmatrix}1 & -2 & 7\end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix}.1 & .2 & .7\end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix}C & B & A\end{bmatrix}$$

$$\text{rank}^{S_1} = \begin{bmatrix}C & B & A\end{bmatrix} \quad \text{rank}^{S_2} = \begin{bmatrix}C & A & B\end{bmatrix} \quad \text{rank}^{S_3} = \begin{bmatrix}A & B & C\end{bmatrix}$$

$$\text{Borda Count:} \;\; \boldsymbol{g}_{agg} = \text{rank}^{agg} = \begin{bmatrix}C & A & B\end{bmatrix}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

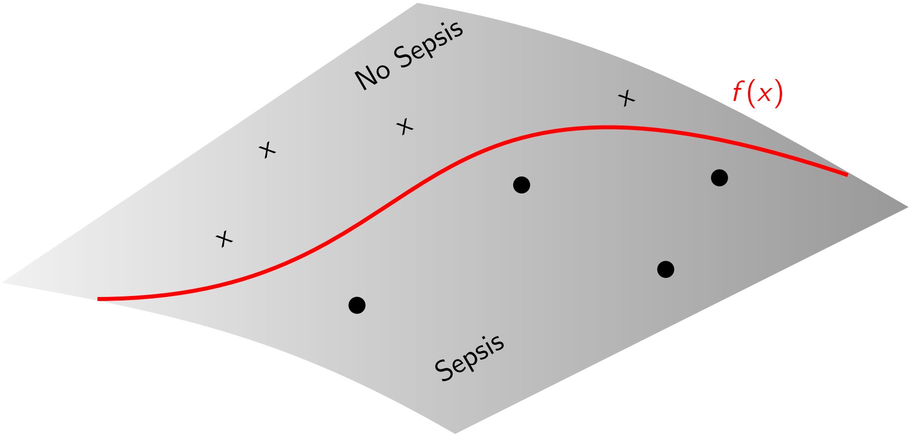
$$\boldsymbol{g}_c^S = \begin{bmatrix}1 & -2 & 7\end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix}.1 & .2 & .7\end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix}C & B & A\end{bmatrix}$$

$$\text{rank}^{S_1} = \begin{bmatrix}C & B & A\end{bmatrix} \;\; \text{rank}^{S_2} = \begin{bmatrix}C & A & B\end{bmatrix} \;\; \text{rank}^{S_3} = \begin{bmatrix}A & B & C\end{bmatrix}$$

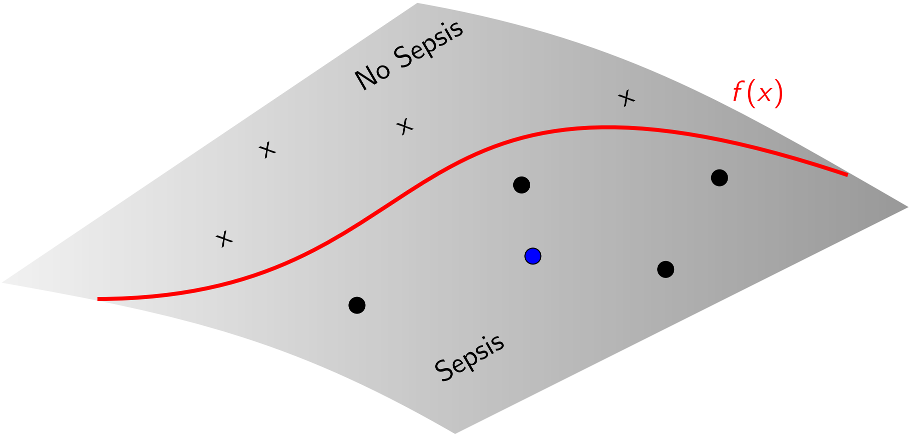
$$\text{Borda Count:} \;\; \boldsymbol{g}_{agg} = \text{rank}^{agg} = \begin{bmatrix}C & A & B\end{bmatrix}$$

$$\boldsymbol{g}_{agg} \in \arg\min_{\boldsymbol{g}} \sum_{\boldsymbol{g}_i \in \mathcal{G}_m} d(\boldsymbol{g}, \boldsymbol{g}_i)$$

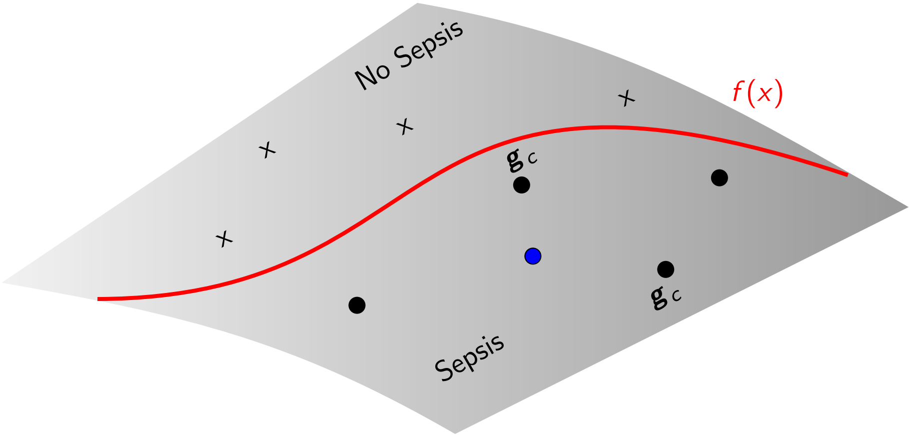
Aggregating Local Explanations



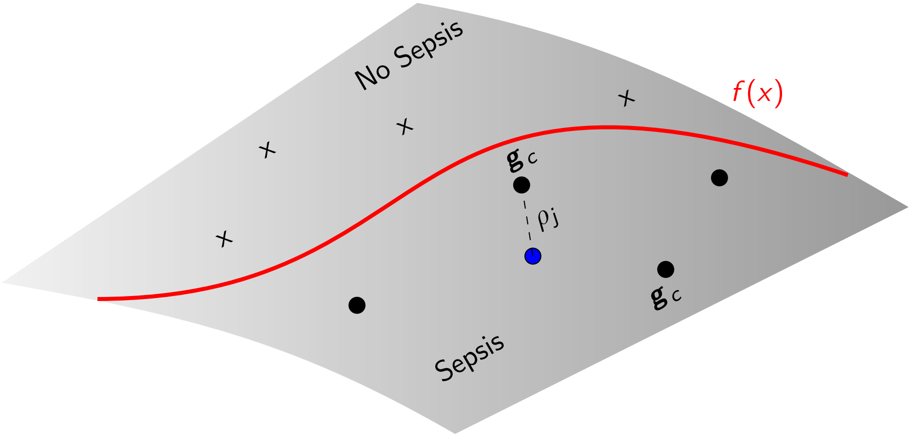
Aggregating Local Explanations



Aggregating Local Explanations



Aggregating Local Explanations



AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

1. Find k nearest neighbors, $\mathcal{N}_k$, of x_{test} and their weights, ρ_j

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon, x^{(j)}}, x_{\text{test}})\Big|_{\epsilon=0}$$

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D}, |\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

1. Find k nearest neighbors, $\mathcal{N}_k$, of x_{test} and their weights, ρ_j

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\Big|_{\epsilon=0}$$

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D},|\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

2. Calculate the attributions, $\boldsymbol{g}_c$, for all points in $\mathcal{N}_k$

$$\boldsymbol{g}_{ci} = \phi_i = \frac{1}{|F|}\sum_{S\subseteq F\setminus\{i\}} \binom{F-1}{S}^{-1} \left(f(x_{S\cup\{i\}}) - f(x_S)\right)$$

AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

1. Find k nearest neighbors, $\mathcal{N}_k$, of x_{test} and their weights, ρ_j

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\Big|_{\epsilon=0}$$

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D}, |\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

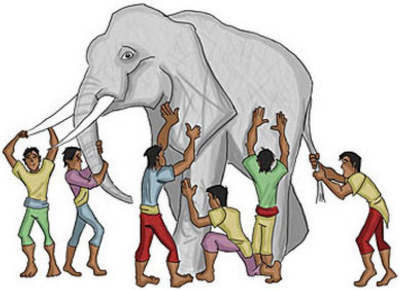
2. Calculate the attributions, $\boldsymbol{g}_c$, for all points in $\mathcal{N}_k$

$$\boldsymbol{g}_{ci} = \phi_i = \frac{1}{|F|}\sum_{S\subseteq F\setminus\{i\}} \binom{F-1}{S}^{-1} \left(f(x_{S\cup\{i\}}) - f(x_S)\right)$$

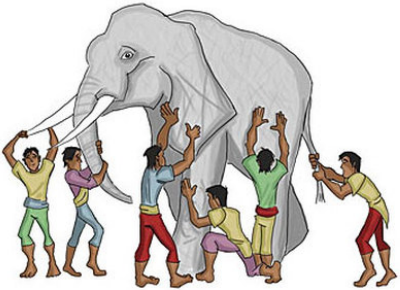
3. Aggregate the k explanations into a consensus, $\boldsymbol{g}_{agg}$

$$\boldsymbol{g}_{agg} = \sum_{x^{(j)}\in\mathcal{N}_k} \frac{\rho_j}{\rho}\boldsymbol{g}_c^j$$

Why aggregate?

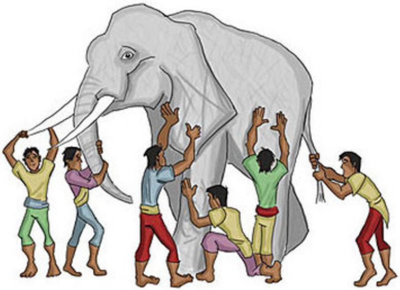


Why aggregate?



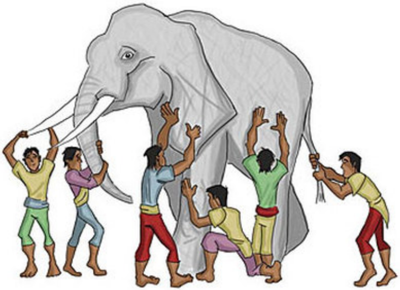
- Each training point has its own "learned" attribution

Why aggregate?



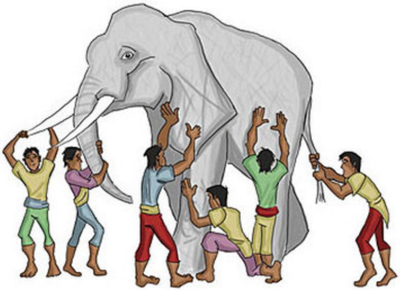
- Each training point has its own "learned" attribution
- Aggregate explanation now has lower sensitivity

Why aggregate?



- Each training point has its own "learned" attribution
- Aggregate explanation now has lower sensitivity
- Resulting attribution uses motivating reasoning of a doctor

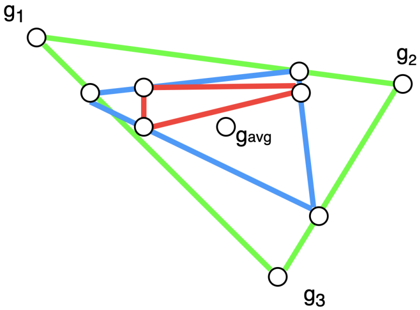
Why aggregate?



- Each training point has its own "learned" attribution
- Aggregate explanation now has lower sensitivity
- Resulting attribution uses motivating reasoning of a doctor
- **SUPER SUPER** cheap to compute

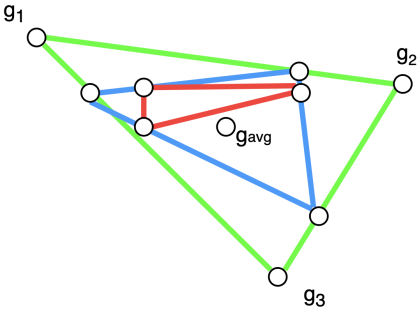
Minimizing Complexity

Region Shrinking Method

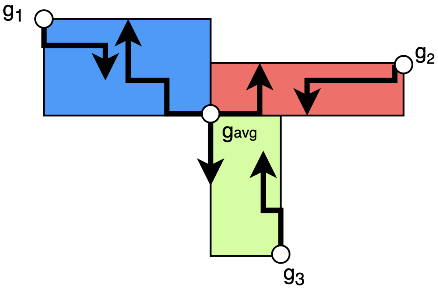


Minimizing Complexity

Region Shrinking Method



Gradient-Descent Style Method



Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
    - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.
3. Working with medical experts to find a $\boldsymbol{g}^*$

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
    - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
    - Consider counterfactuals, natural language explanations, etc.
3. Working with medical experts to find a $\boldsymbol{g}^*$
4. Multi-Objective optimization

Conclusion

Summary

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.
3. Working with medical experts to find a $\boldsymbol{g}^*$
4. Multi-Objective optimization
   - Resulting Setup

$$max \text{ faithfulness}(\boldsymbol{g}_{agg}) + \text{sensitivity}(\boldsymbol{g}_{agg})$$