再構成誤差のシャープレイ値による異常検知の説明

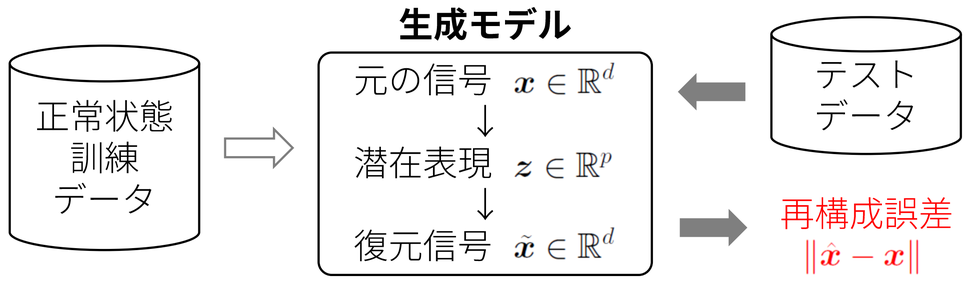
Explaining anomaly detection using Shapley value of reconstruction errors

武石 直也 (理研AIP)

arXiv:1909.03495

概要　異常検知におけるプラクティスとして，正常データに対してPCAなどの生成モデルを学習し，その再構成誤差を監視する方法がある．このとき，どの特徴が異常なのかを特定する手がかりとして，各特徴の再構成誤差を比較することがよくある．しかし，異常特徴の再構成誤差だけが常に大きくなるとは限らない．そこで本稿では，シャープレイ値を用いて異常特徴特定の手がかりを得ることを提案する．

背景　再構成誤差による異常検知 (detection)



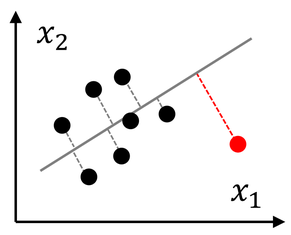
例　確率的主成分分析 (PCA)　[Tipping&Bishop 99]

生成モデル：

$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}_{\boldsymbol{x}}(\boldsymbol{W}\boldsymbol{z},\, \sigma^2 \boldsymbol{I})$$
$$p(\boldsymbol{z}) = \mathcal{N}_{\boldsymbol{z}}(\boldsymbol{0},\, \boldsymbol{I})$$

最適な再構成：

$$\hat{\boldsymbol{x}} = \boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{x}$$



単純だが，(ベースラインとして) よく用いられる．

目的　異常特徴の特定 (localization)

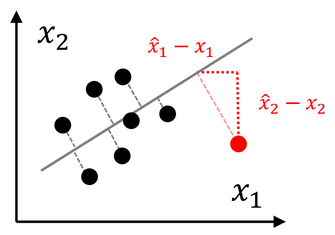
検知だけでなく，どの特徴が原因で異常となったのか特定したい．

※ 正しいモデルがないと「真の」原因 (因果) は特定できない．
　機械学習による方法では，*近似的に原因の見当をつける* という問題．

再構成誤差による異常特徴特定　各特徴の再構成誤差の大小を比べる．
再構成誤差の計算時に必ず出てくるので，大雑把な近似としては便利．

$$e(\boldsymbol{x}) = \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 = (\hat{x}_1 - x_1)^2 \;+\; \cdots \;+\; (\hat{x}_d - x_d)^2$$

どの特徴について再構成誤差が大きいか比べる



しかし，異常特徴の再構成誤差（だけ）が大きくなるとは限らない．

アイデア　シャープレイ値 (Shapley value) の利用

各特徴の再構成誤差そのものではなく，再構成誤差のシャープレイ値を見るとよいのではないか？

シャープレイ値　協力ゲームの利得を各プレイヤーに *公正に* 分配する方法のひとつ [Shapley 53]．いくつかの良い性質をもつ．

利得を表す特性関数 $v\colon \text{subsets of } \{1,\dots,d\} \to \mathbb{R}$ のもと，プレイヤー i のシャープレイ値は次のように定義される：

$$\phi_i(v) = \sum_{S \subseteq \{1,\dots,d\}\setminus\{i\}} \frac{|S|!(d-|S|-1)!}{d!}\big(v(S\cup\{i\}) - v(S)\big)$$

つまり，提携 S (プレイヤーの部分集合) に i が参加する影響の平均．

本研究のアイデア　「再構成誤差 = 利得，各特徴 = プレイヤー」と考えて，*再構成誤差の各特徴への分配* を異常特徴特定に利用する．

シャープレイ値の計算にあたって

ポイント①　特性関数の定義　v をどのように定義すればよいか？

「特徴集合 S のもとでの再構成誤差」では再訓練が必要になり不便．
→「S にない特徴について *再構成誤差を周辺化*」したものとして定義．
cf. [Štrumbelj&Kononenko 14; Lundberg&Lee 17]

$$v(S)\big|_{\boldsymbol{x}} = d^{-1}\,\mathbb{E}_{p(\boldsymbol{x}_{S^c}|\boldsymbol{x}_S)}\big[\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2\big]$$

ただし，S^c は S の補集合で，S, S^c による添字は部分ベクトル・部分行列．

ポイント②　条件付き分布に関する期待値計算

ひとまず，確率的PCAについては厳密に計算できる (他は要研究)．

$$d \cdot v(S)\big|_{\boldsymbol{x}} = \operatorname{tr}\big((\boldsymbol{I} - \boldsymbol{B}_{S^c})(\boldsymbol{V} + \boldsymbol{m}\boldsymbol{m}^\top)\big) - 2\operatorname{tr}\big(\boldsymbol{B}_{S^c,S}\boldsymbol{x}_S, \boldsymbol{m}^\top\big) + \operatorname{tr}\big((\boldsymbol{I} - \boldsymbol{B}_S)\boldsymbol{x}_S\boldsymbol{x}_S^\top\big)$$

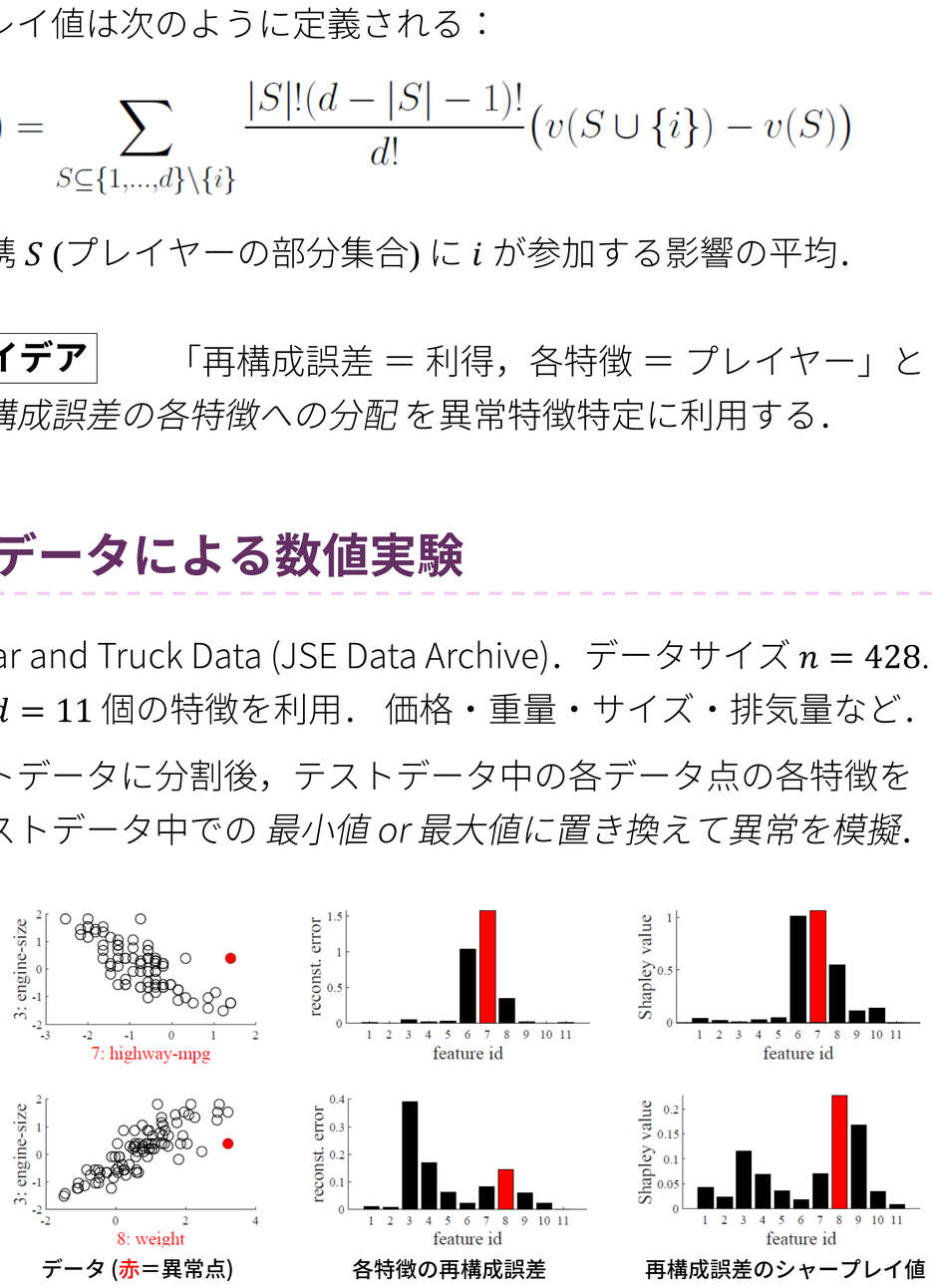
ただし，　$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$,　$\boldsymbol{C} = \sigma^2\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^\top$
$\boldsymbol{m} = \boldsymbol{C}_{S^c,S}\boldsymbol{C}_S^{-1}\boldsymbol{x}_S$,　$\boldsymbol{V} = \boldsymbol{C}_{S^c,S}\boldsymbol{C}_S^{-1}\boldsymbol{C}_{S^c,S}^\top$

模擬異常データによる数値実験

2004 New Car and Truck Data (JSE Data Archive)．データサイズ $n = 428$．
欠損のない $d = 11$ 個の特徴を利用． 価格・重量・サイズ・排気量など．
訓練・テストデータに分割後，テストデータ中の各データ点の各特徴を *それぞれテストデータ中での 最小値 or 最大値に置き換えて異常を模擬*．

例①
両方うまくいく

例②
再構成誤差では
特定失敗



データ (赤＝異常点)　各特徴の再構成誤差　再構成誤差のシャープレイ値

	MAX		MIN	
	Hits@1	Hits@3	Hits@1	Hits@3
reconstruction error	.316	.605	.271	.471
Shapley values	.484	.801	.484	.710

← 異常特徴特定の Hits@n (指標の上位n個によって異常特徴が正しく特定される割合)

シャープレイ値によってよりよく異常特徴を特定できる (かもしれない)．

議論①　再構成誤差と何が違うのか？

きわめて単純な例として，$\mathbb{R}^2$ 上の (正常) データ $\boldsymbol{x}$ の真の生成プロセスが右式で与えられる場合を考える（$z \in \mathbb{R}$ は潜在変数）．

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} z$$

これに対して，次のような異常発生パターンを考える (いずれも 特徴#2 が異常)．

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} z + \begin{bmatrix} 0 \\ \eta \end{bmatrix}$$
パターンA (加法的)

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \eta \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} z$$
パターンB (乗法的)

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 + \eta \end{bmatrix} z$$
パターンC (プロセス)

再構成誤差の比 e_2/e_1 および シャープレイ値の比 $|\phi_2|/|\phi_1|$ (> 1 となるべき) は次のようになる．

	A	B	C
$\left.\frac{e_2}{e_1}\right\|_\eta$	w_1^2/w_2^2	w_1^2/w_2^2	w_1^2/w_2^2
$\left.\frac{\phi_2(v)}{\phi_1(v)}\right\|_\eta$	$\frac{\eta^2\alpha + (\beta(\eta + w_2 z)^2 - \gamma - \delta)}{\eta^2\alpha - (\beta(\eta + w_2 z)^2 - \gamma - \delta)}$	$\frac{w_2^2 z^2\alpha(\eta-1)^2 + (w_2^2 z^2\beta\eta^2 - \gamma + \delta)}{w_2^2 z^2\alpha(\eta-1)^2 - (w_2^2 z^2\beta\eta^2 - \gamma + \delta)}$	$\frac{z^2\alpha\eta^2 + (z^2\beta(\eta + w_2)^2 - \gamma + \delta)}{z^2\alpha\eta^2 - (z^2\beta(\eta + w_2)^2 - \gamma + \delta)}$

※ $\alpha = (\sigma^2 + w_1^2)^2(\sigma^2 + w_2^2)^2$,　$\beta = \sigma^4(\sigma^2 + w_1^2)^2$,　$\gamma = w_2^2(\sigma^4 z^2 + \sigma^2 w_1^2 + w_1^4)(\sigma^2 + w_2^2)^2$,　$\delta = w_2^4(\sigma^2 + w_1^2)^2(\sigma^2 + w_2^2)$

議論②　VAEなどの場合どうするか？

どうすればいいですか？　例えばモデルが VAE　$p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}_{\boldsymbol{x}}\big(\boldsymbol{\mu}(\boldsymbol{z}),\, \boldsymbol{\Sigma}(\boldsymbol{z})\big)$ のとき，

$$p(\boldsymbol{x}_{S^c} \mid \boldsymbol{x}_S,\, \boldsymbol{z}) = \mathcal{N}_{\boldsymbol{x}_{S^c}}\big(\boldsymbol{\mu}_{S^c} + \boldsymbol{\Sigma}_{S^c,S}\boldsymbol{\Sigma}_S^{-1}(\boldsymbol{x}_S - \boldsymbol{\mu}_S),\, \boldsymbol{\Sigma}_{S^c} - \boldsymbol{\Sigma}_{S^c,S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S^c,S}^\top\big)$$

→ うまく $\boldsymbol{z}$ を追い出してください．　または，何か別のうまい方法？
　DNN出力のシャープレイ値を近似する方法も議論されている [Ancona+ 19 など]．

参考文献　[Tipping&Bishop 99] J. R. Stat. Soc. Ser. B. Stat. Methodol., 61(3):611-622.　[Shapley 53] Ann. of Math. Stud. 28, II:307-317.　[Štrumbelj&Kononenko 14] Knowl. Inform. Syst., 41(3):647-665.
[Lundberg&Lee 17] In NeurIPS 2017, pages 4765-4774.　[Ancona+ 19] In ICML 2019, pages 272-281.