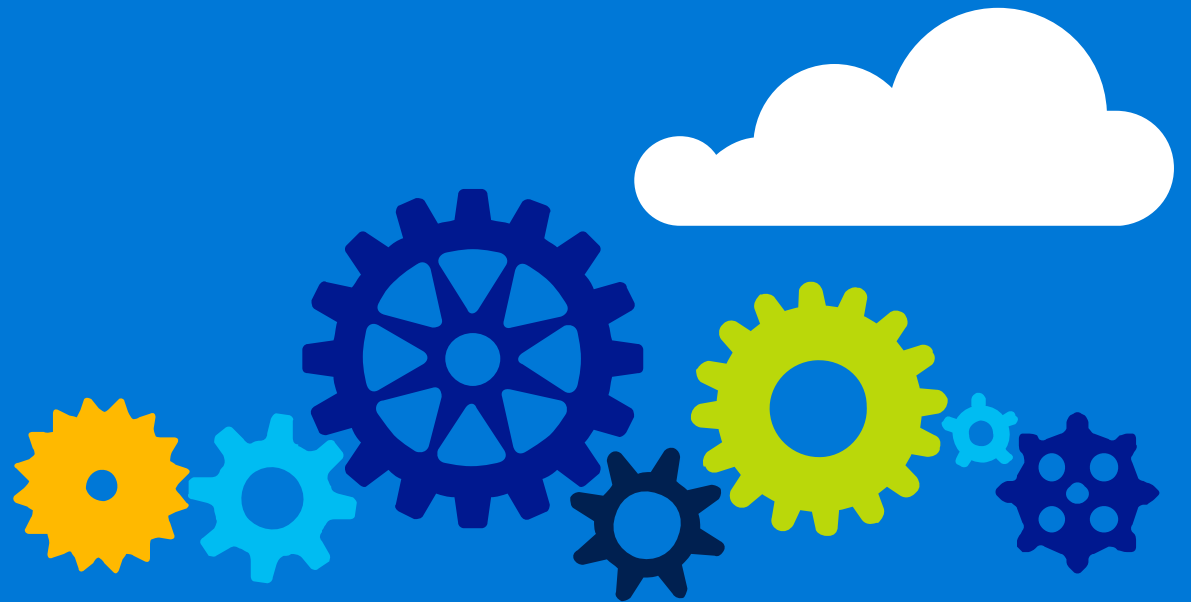


Azure Machine Learning による 異常検知手法

データミックス インテグレーションステップ

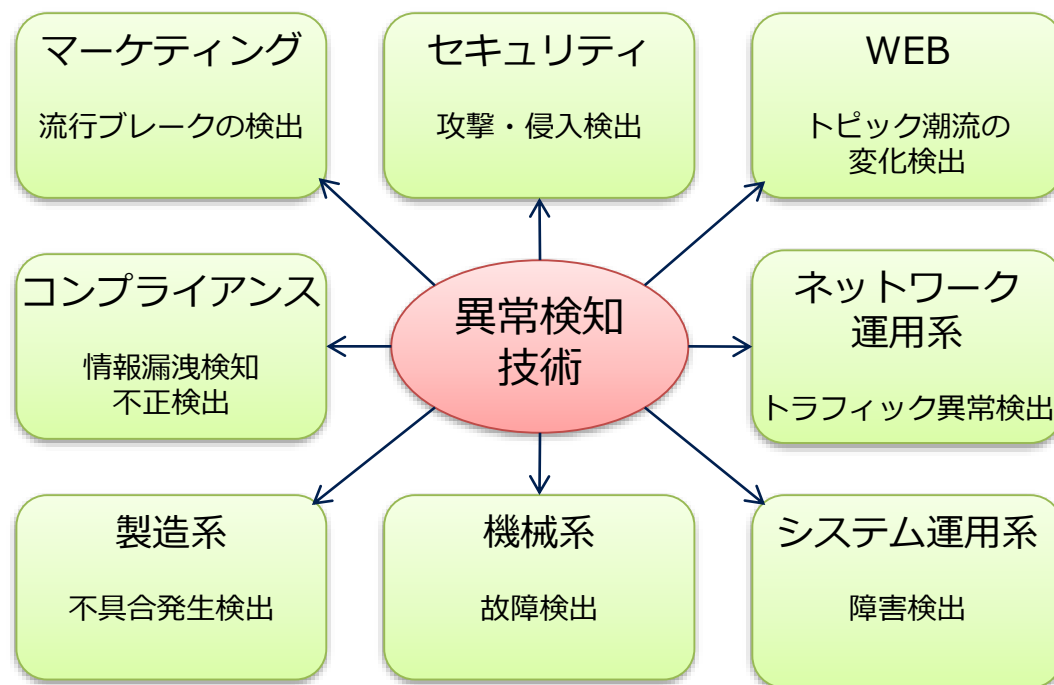
2017/12/23



課題と目的

- ・ 異常検知手法の習得

- ・ IoT と相性が良くビジネス的に応用範囲が大きいと思うため



←「データマイニングによる異常検知（山西健司）」より

- ・ Azure Machine Learning 利用方法の習得

- ・ GUI ベースの機械学習ツールの評価をしたかったため

異常検知の手法

- ・ A. 外れ値検出 \Rightarrow 仲間から値が外れている
- ・ B. 変化点検出 \Rightarrow 周波数の変化
- ・ C. 異常部位検出 \Rightarrow 異常なセッション

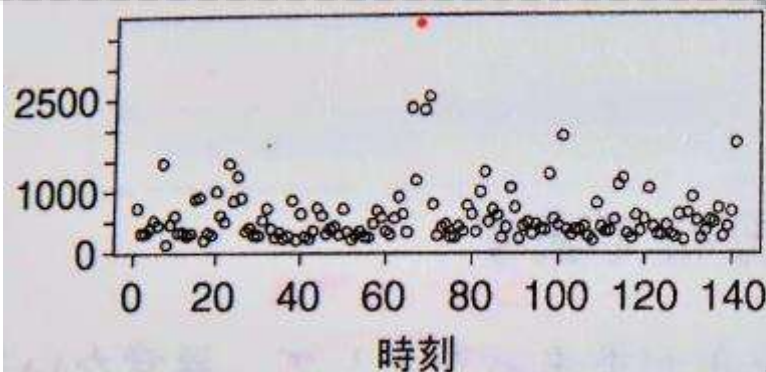
「異常検知と変化検知（井出剛／杉山将）」より



A. 外れ値検出

※時系列ではない

本日の範囲



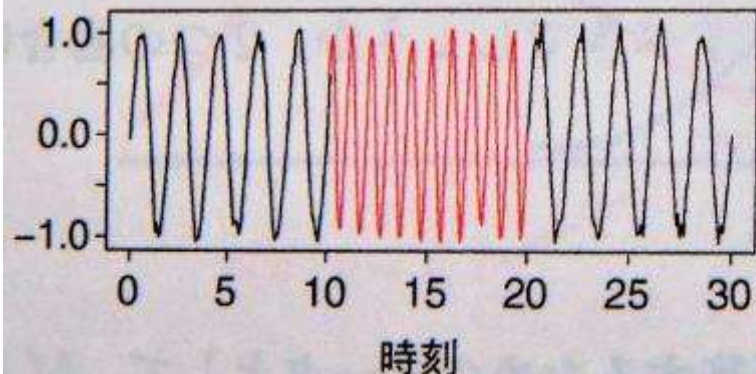
A. 外れ値検出

※時系列



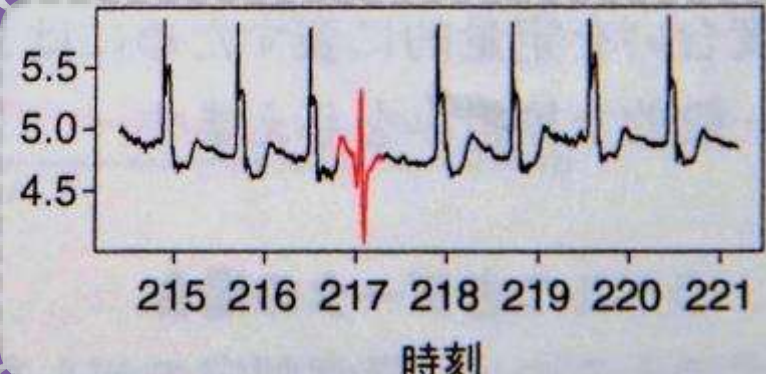
B. 変化点検知

※時系列



C. 異常部位検出

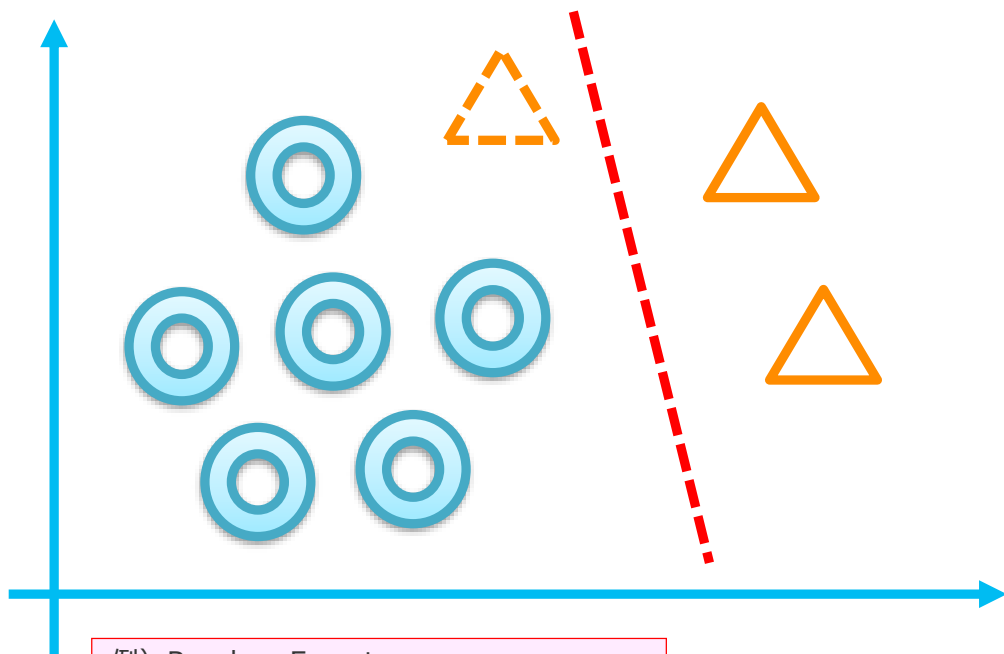
※時系列



異常検知の手法

教師あり異常検知

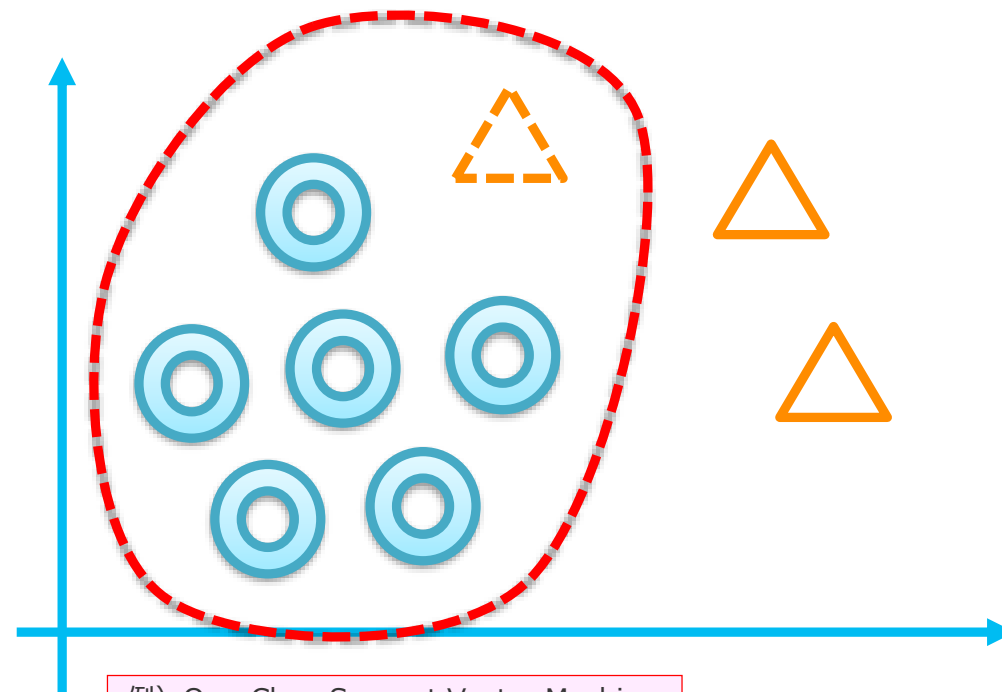
- 少数の既知異常データが存在
- 正常サンプルから判別する分類モデルを学習
- 既知異常に非常に有効



例) Random Forest

教師なし異常検知

- 異常データは用いない
- 与えられた正常データをモデリング
- 未知の異常にも有効



例) One-Class Support Vector Machine
例) PCA-based

異常検知のサンプル

- Cortana Intelligence Gallery
 - Azure Machine Learning のサンプル
 - <https://gallery.cortanaintelligence.com/>
- Anomaly Detection: Credit Risk
 - Attempts to predict credit risk as anomalies within the data.
 - <https://gallery.cortanaintelligence.com/Experiment/1219e87f8fb84e88a2e1b54256808bb3>

利用するデータ

- The UCI Machine Learning Repository
 - <https://archive.ics.uci.edu/ml/about.html>
 - カリフォルニア大学アーバイン校が運営している、機械学習アルゴリズムの実証分析用データの配布サイト
- Statlog (German Credit Data) Data Set
 - [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))
 - 各種属性を持つ人々を信用リスク（good or bad）で分類したデータ
 - Strathclyde 大学の Hofmann 教授が提供

利用するデータ

- Credit risk が正常／危険を表す
 - 1 → 正常（normal）。予測モデルの学習に利用。
 - 2 → 危険（risky）。予測モデルの評価に利用。

rows
1000









columns
21

Duration in months	Credit history	Purpose	Credit amount	Savings account/bond	Present employment since	Installment rate in percentage of disposable income	Personal status and sex	Other debtors/guarantors	Present residence since	Property	Age in years	Other installment plans	Housing	Number of existing credits at this bank	Job	Number of people being liable to provide maintenance for	Telephone	Foreign worker	Credit risk
6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	1
48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	2
12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	1
42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1
24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	2
36	A32	A46	9055	A65	A73	2	A93	A101	4	A124	35	A143	A153	1	A172	2	A192	A201	1
24	A32	A42	2835	A63	A75	3	A93	A101	4	A122	53	A143	A152	1	A173	1	A191	A201	1
36	A32	A41	6948	A61	A73	2	A93	A101	2	A123	35	A143	A151	1	A174	1	A192	A201	1
12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A201	1

データの傾向









Anomaly Detection: Credit Risk v1.1 > Summarize Data > Results dataset

rows: 62
columns: 23

	Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation
view as								
	Duration in months	1000	33	0	4	72	20.903	9.496524
	Credit amount	1000	921	0	250	18424	3271.258	2048.541528
	Installment rate in percentage of disposable income	1000	4	0	1	4	2.973	0.986182
	Present residence since	1000	4	0	1	4	2.845	1.00022
	Age in years	1000	53	0	19	75	35.546	9.026096
	Number of existing credits at this bank	1000	4	0	1	4	1.407	0.515262
	Number of people being liable to provide maintenance for	1000	2	0	1	2	1.155	0.26195
	Credit risk	1000	2	0	1	2	1.3	0.42

Anomaly Detection: Credit Risk v1.1 > Compute Linear Correlation > Results dataset

rows: 62
columns: 62

	Duration in months	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for	Credit risk
view as								
	1	0.624984	0.074749	0.034067	-0.036136	-0.011284	-0.023834	0.214927
	0.624984	1	-0.271316	0.028926	0.032716	0.020795	0.017142	0.154739
	0.074749	-0.271316	1	0.049302	0.058266	0.021669	-0.071207	0.072404
	0.034067	0.028926	0.049302	1	0.266419	0.089625	0.042643	0.002967
	-0.036136	0.032716	0.058266	0.266419	1	0.149254	0.118201	-0.091127
	-0.011284	0.020795	0.021669	0.089625	0.149254	1	0.109667	-0.045732
	-0.023834	0.017142	-0.071207	0.042643	0.118201	0.109667	1	-0.003015
	0.214927	0.154739	0.072404	0.002967	-0.091127	-0.045732	-0.003015	1

予測モデル

- ・ risky かどうかを判断するモデルを作成
- ・ 判断を誤る場合のコスト削減が目的
 - ・ 以下、Statlog (German Credit Data) Data Set に付属している誤判別コストのサンプル表 (損益行列)
- ・ risky を見誤るコストは、normal を見誤るコストの 5 倍

		判定結果	
		risky	normal
正解	risky	TP (True Positive)	FN (False Negative)
		コスト=0	コスト=5
	normal	FP (False Positive)	TN (True Negative)
		コスト=1	コスト=0

予測モデル

- One-Class Support Vector Machine

- 教師なし学習

- パラメータ

- This parameter determines the trade-off between the fraction of outliers and the number of support vectors.

- Specifies the stopping tolerance.

- PCA-based

- 教師なし学習

- パラメータ

- The number of PCA components

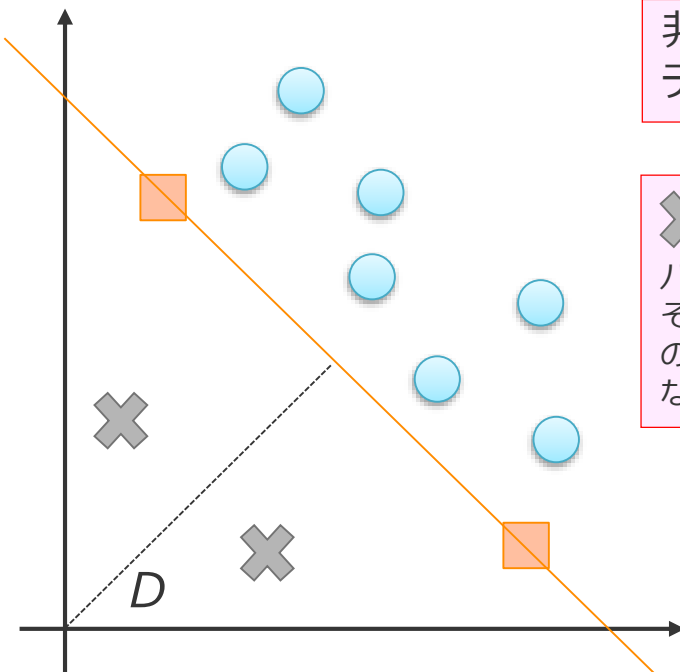
- The oversampling parameter used in randomized PCA

One-Class Support Vector Machine

- ・SVMを1クラスのみでの学習に用い、入力データがそのクラスに入るか入らないかのみを判断
- ・新規性判断、例外検出、外れ値検出などに利用

1クラス v-サポートベクトルマシン

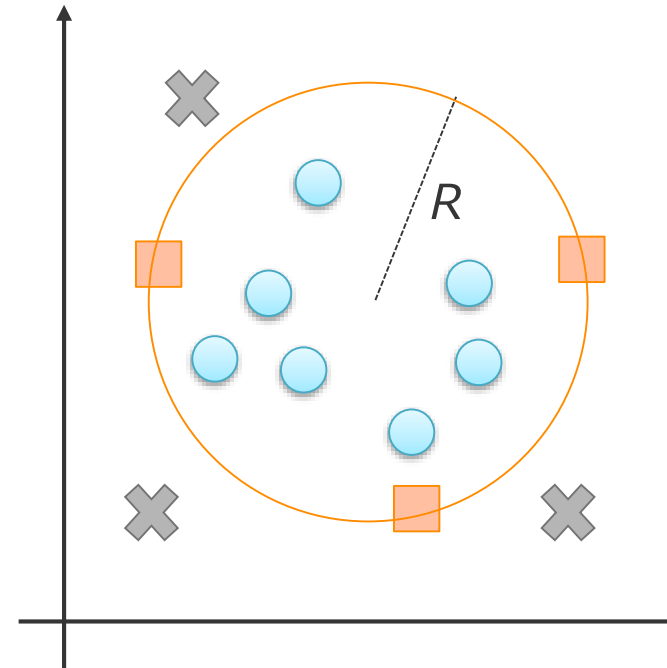
本日はこちら



非線形変換されたデータを超平面分割

✕ の割合の上限をハイパーパラメータで指定する。その制限下で、超平面と原点の距離 D が出来るだけ大きくなるように学習する。

サポートベクトル領域記述法



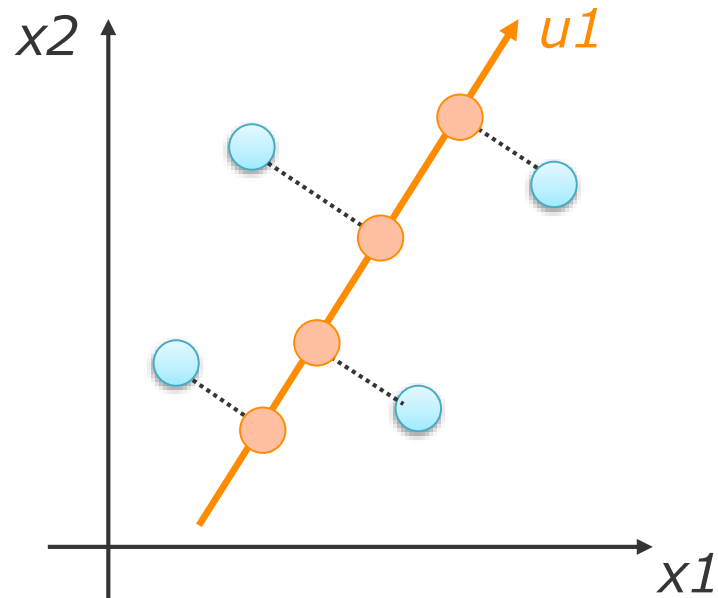
非線形変換されたデータを球面分割

✕ の割合の上限をハイパーパラメータで指定する。その制限下で、球面の半径 R が出来るだけ小さくなるように学習する。

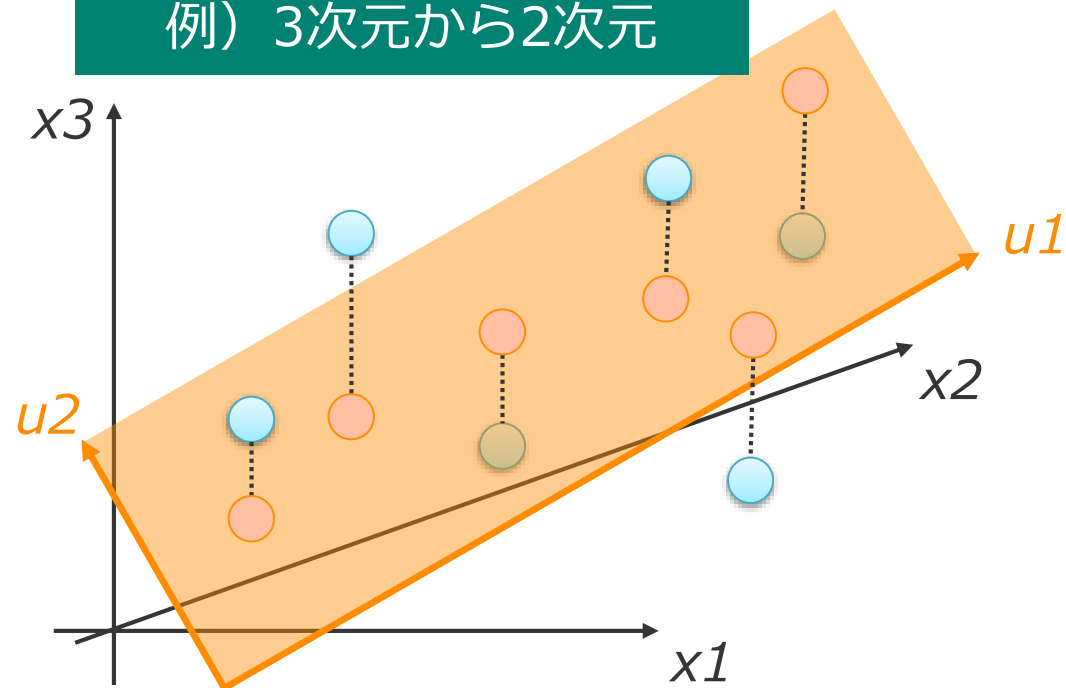
PCA based

- 主成分分析 (Principal Component Analysis)
 - 次元削減、非可逆データ圧縮、特徴抽出、データ可視化等に応用
 - 主成分空間と呼ばれる低次元への線形空間へ直交射影
 - 射影されたデータの分散が最大化されるように射影
 - もとのデータと射影されたデータの距離の総和を最小化することと同じ

例) 2次元から1次元

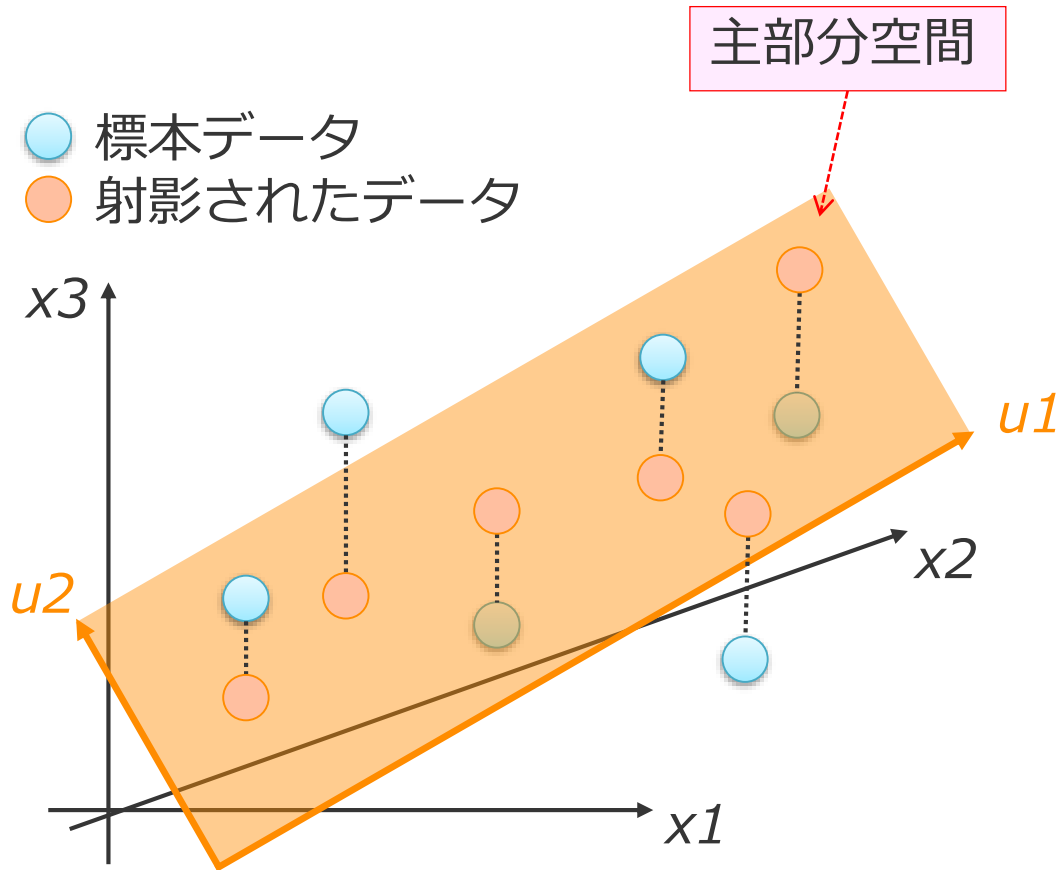


例) 3次元から2次元

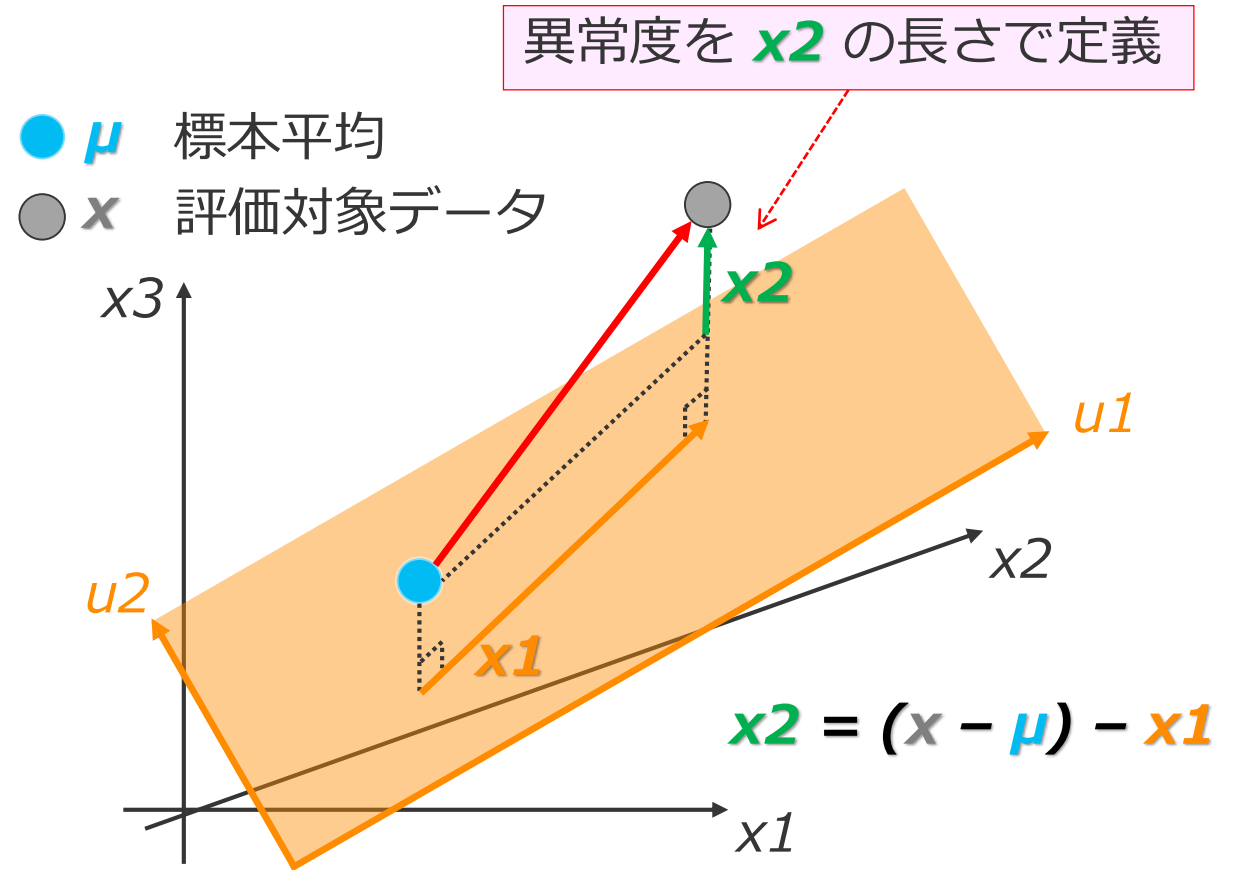


PCA based

・主成分分析（主部分空間）による異常検知

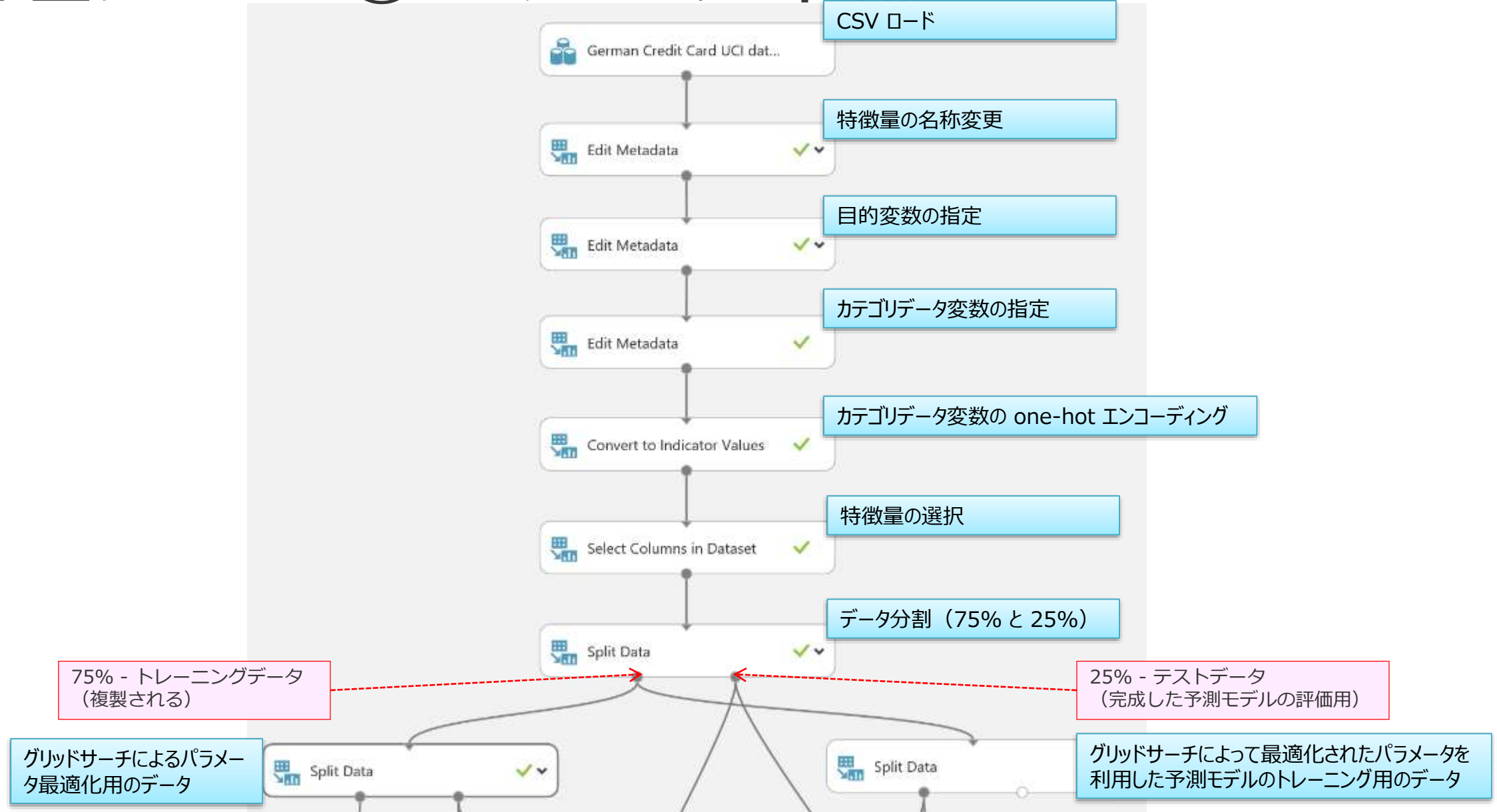


主部分空間の導出

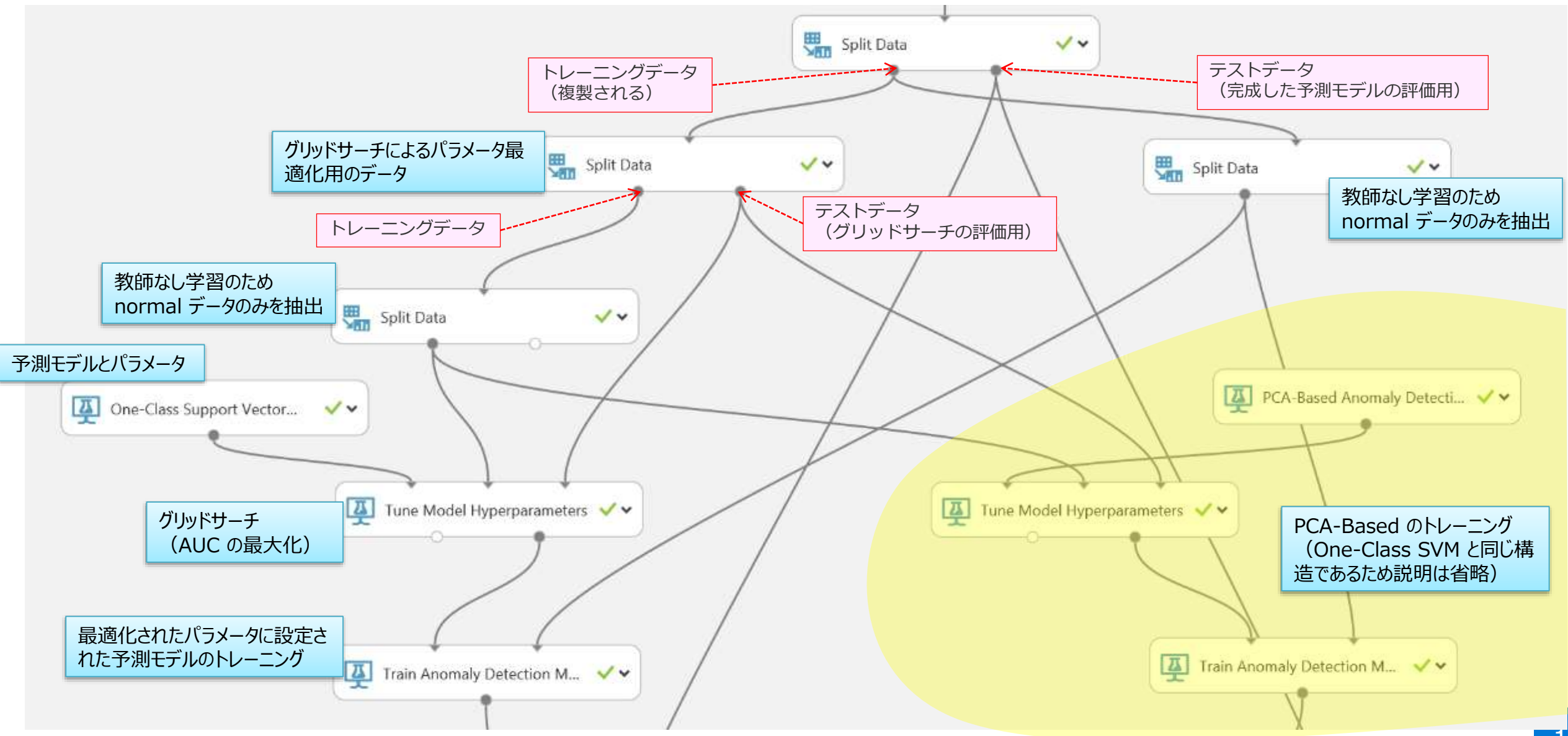


主部分空間による異常検知

処理フロー①：データ準備



処理フロー②：予測モデルのトレーニング



処理フロー③：予測モデルの評価

最適化されたパラメータに設定された予測モデルのトレーニング

完成した予測モデルの評価

PCA-Based のトレーニング
(One-Class SVM と同じ構造であるため説明は省略)

$$z = \frac{1}{1 + \exp(-x)}$$

Scored Probabilities の正規化

Foreign worker-A201	Foreign worker-A202	Credit risk	Scored Labels	Scored Probabilities
1	0	2	0	0.1073668
1	0	1	0	0.056063652
1	0	1	0	0.026780844
1	0	1	0	0.007402182
1	0	1	0	0.063828707
1	0	2	0	0.073522091
1	0	1	0	0.195270061
1	0	1	0	0.192638636
1	0	1	0	-0.05578303
1	0	1	0	0.022918463
1	0	2	0	-0.07668805

結果：データセット

One-Class SVM

Foreign worker-A201	Foreign worker-A202	Credit risk	Scored Labels	Scored Probabilities
1	0	2	0	0.526815945
1	0	1	0	0.514012243
1	0	1	0	0.506694811
1	0	1	0	0.501850537
1	0	1	0	0.515951761
1	0	2	0	0.518372248
1	0	1	0	0.518662085

PCA based

	Foreign worker-A201	Foreign worker-A202	Credit risk	Scored Labels	Scored Probabilities
0	1	0	2	1	0.665387273
0	1	0	1	1	0.61589545
0	1	0	1	1	0.642105877
0	1	0	1	1	0.672194362
1	1	0	1	1	0.667465866
1	1	0	2	1	0.735048056
1	1	0	1	1	0.707604527

risky かどうかを判定するための Scored Probabilities の閾値（Threshold／カットオフポイント）を設定する必要がある。

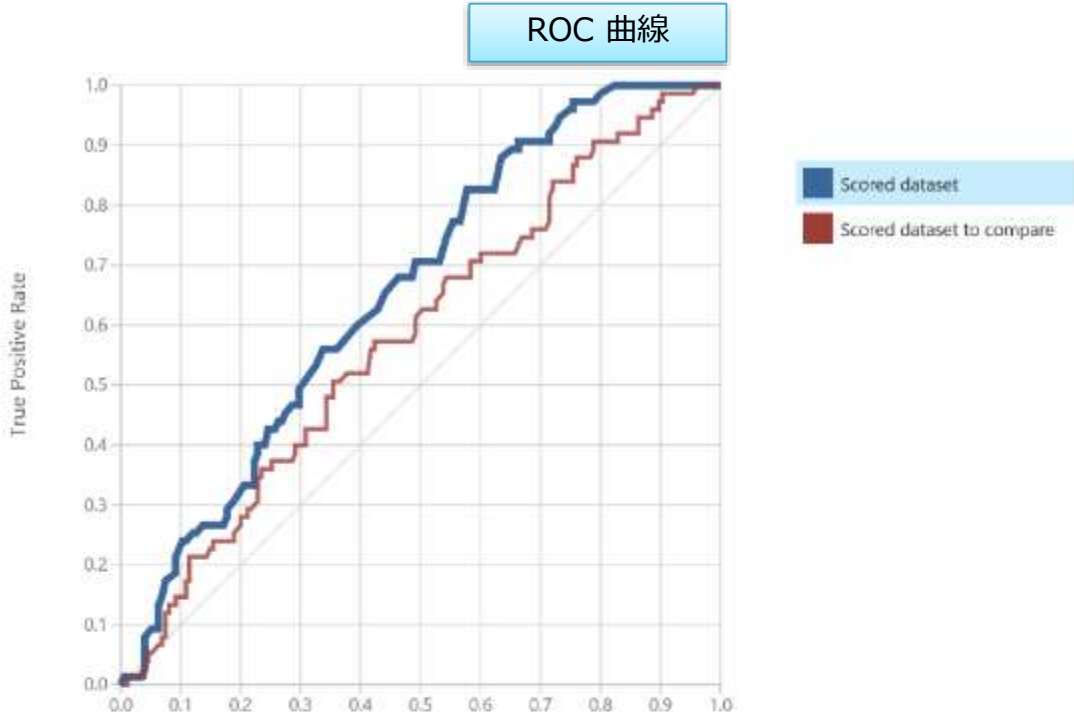
結果：コスト削減効果 [One-Class SVM]

One-Class Support Vector Machine

AUC=0.657 閾値=0.48		判定結果	
		risky	normal
正解	risky	True Positive	False Negative
		コスト=0	コスト=5
		75	0
	normal	False Positive	True Negative
		コスト=1	コスト=0
		144	31

コスト合計=144

- ・ 何もしない場合 (=全てを risky と判断する場合) のコストは 175
- ・ 175 から 144 にコストが減った



混同行列 (confusion matrix)

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
75	0	0.424	0.342	0.48	0.657
False Positive	True Negative	Recall	F1 Score		
144	31	1.000	0.510		
Positive Label	Negative Label				
2	1				

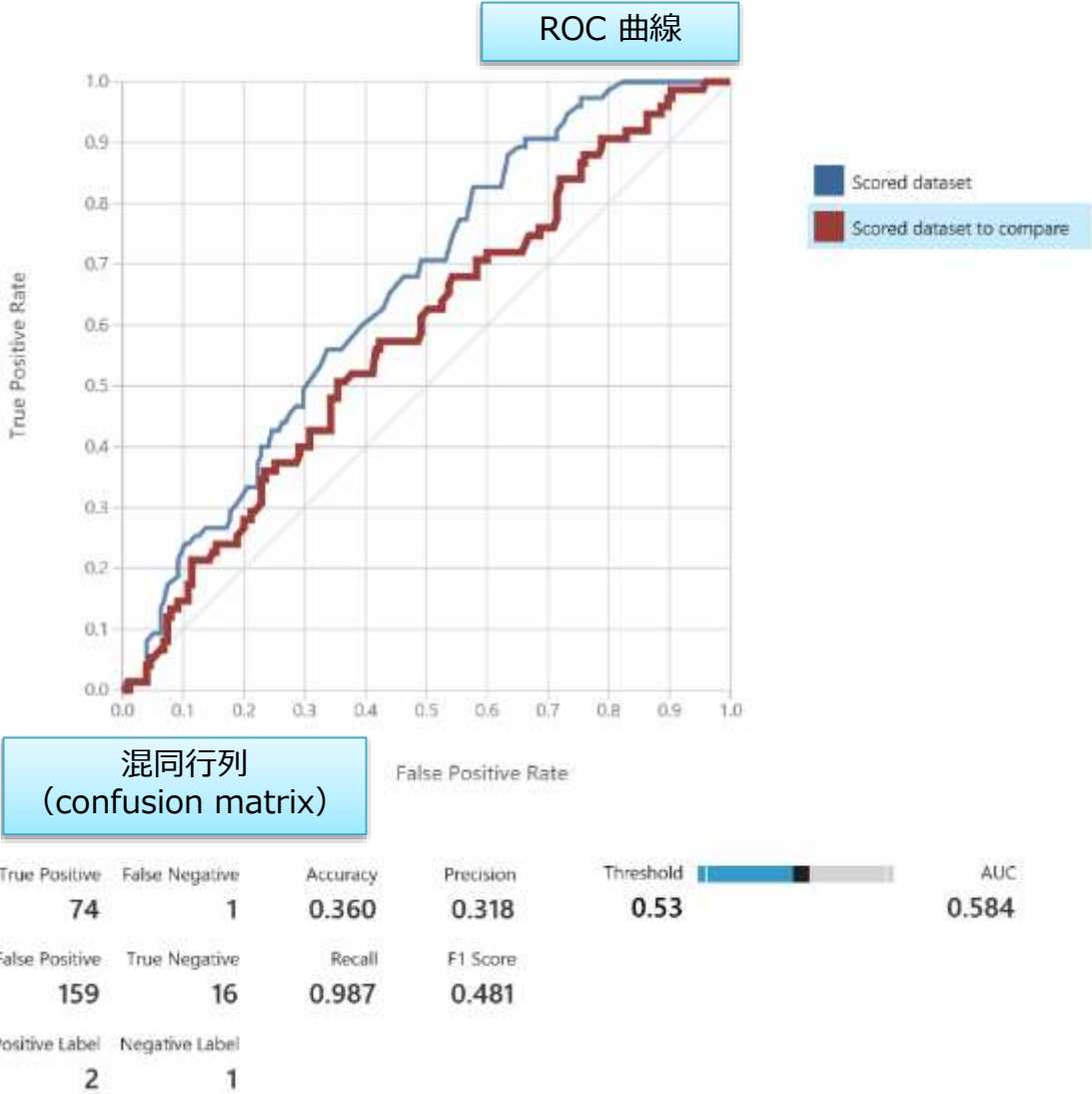
結果：コスト削減効果 [PCA-based]

PCA-based

AUC=0.584 閾値=0.53		判定結果	
		risky	normal
正解	risky	True Positive	False Negative
		コスト=0	コスト=5
		74	1
	normal	False Positive	True Negative
		コスト=1	コスト=0
		159	16

コスト合計=164

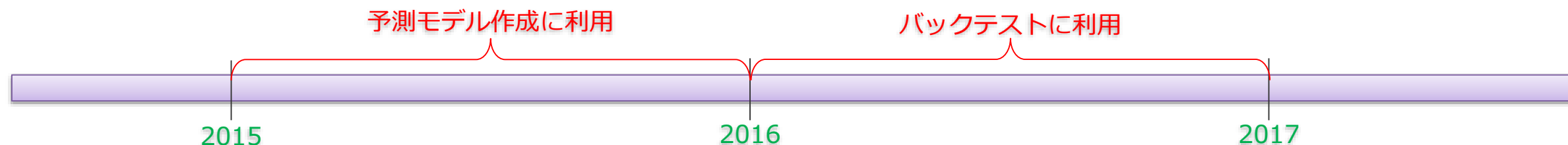
- ・ 何もしない場合 (=全てを risky と判断する場合) のコストは 175
- ・ 175 から 144 にコストが減った



検証計画

・バックテスト

- ・今回利用したデータが、例えば 2015 年のデータであったと仮定した場合、2016 年のデータを使って予測精度を評価する。

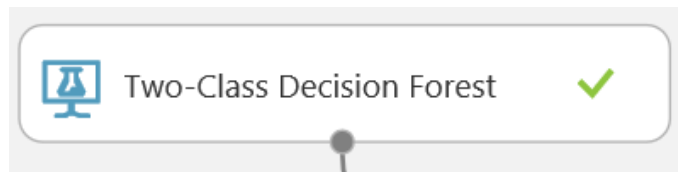


・A/B テスト

- ・Credit Risk 発生率が同程度の 2 店舗を選択して A/B テストを実施

		新規契約者数	Credit Risk 発生数	Credit Risk 発生率
予測モデルの利用	あり (渋谷支店)	1,000 人	100 人	10%
	なし (新宿支店)	5,000 人	350 人	7%
	合計	6,000 人	450 人	7.50%

教師あり学習との比較 [Random Forest]

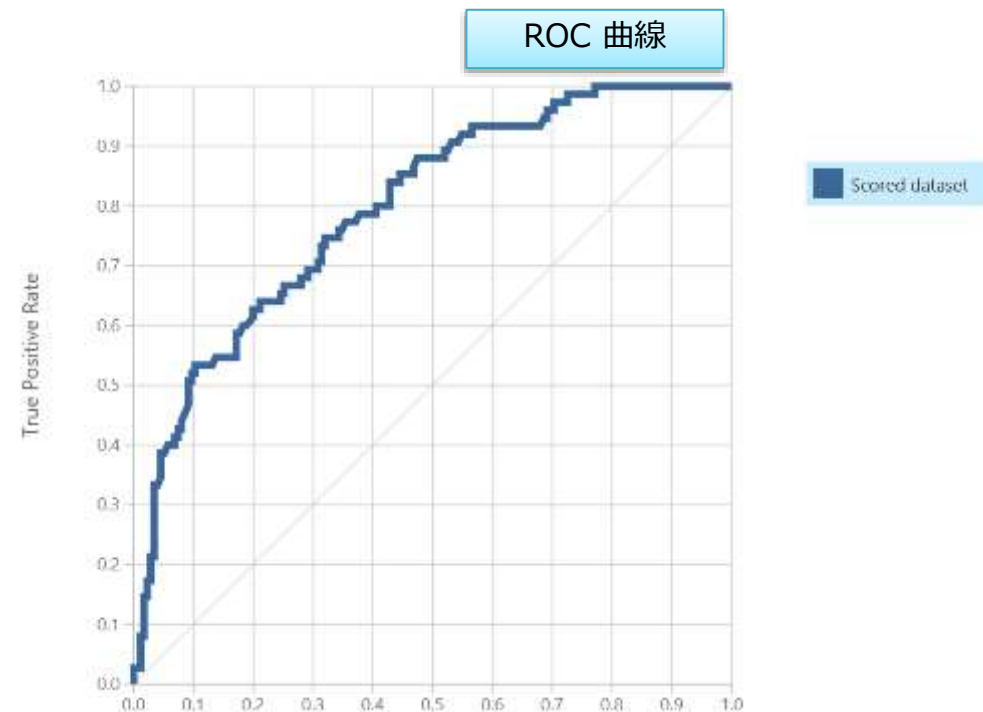


Random Forest

AUC=0.796		判定結果	
閾値=0.18		risky	normal
正解	risky	True Positive	False Negative
		コスト=0	コスト=5
		70	5
	normal	False Positive	True Negative
		コスト=1	コスト=0
		102	73

コスト合計=127

- 何もしない場合 (=全てを risky と判断する場合) のコストは 175
- 175 から 127 にコストが減った



混同行列 (confusion matrix)

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
70	5	0.572	0.407	0.18	0.796
False Positive	True Negative	Recall	F1 Score		
102	73	0.933	0.567		
Positive Label	Negative Label				
2	1				










Random Forest のハイパーパラメーター

リーフに入る標本の最低数

枝分かれ可能な最大数

決定木の深さ

決定木の数

Minimum number of samples per leaf node	Number of random splits per node	Maximum depth of the decision trees	Number of decision trees	Accuracy	Precision	Recall	F-Score	AUC
								
3	8	10	110	0.727273	0.727273	0.142857	0.238806	0.798937
3	8	10	120	0.737968	0.818182	0.160714	0.268657	0.797165
3	8	10	130	0.748663	0.846154	0.196429	0.318841	0.794575
3	9	10	110	0.754011	0.8125	0.232143	0.361111	0.793757
3	8	10	100	0.754011	0.8125	0.232143	0.361111	0.793621
3	9	10	100	0.754011	0.8125	0.232143	0.361111	0.792666
4	9	11	110	0.737968	0.684211	0.232143	0.346667	0.792666
3	7	12	120	0.748663	0.8	0.214286	0.338028	0.791576
4	9	11	130	0.743316	0.722222	0.232143	0.351351	0.791576
5	8	13	110	0.743316	0.785714	0.196429	0.314286	0.791576

教師あり学習の難しさ

- ・ 正常クラスと異常クラスの標本数に著しい偏りがある
 - ・ 正常標本が異常標本の1000倍であれば、何も工夫しなければ、すべてを正常と判定するモデルができる。それでも、予測精度は 99.9%
- ・ 不均衡データ (imbalanced data) への対応手段
 1. 重みづけ
 - ・ 正常標本が異常標本の10倍であれば、正常標本に0.1の重みを付けて学習
 2. 間引き (ダウンサンプリング)
 - ・ 正常標本が1000個、異常標本が100個であれば、正常標本から100個をランダムで取得して学習
 3. 水増し (アップサンプリング)
 - ・ 上の例で言えば、逆に、異常標本を復元抽出 (ブートストラップサンプリング) により1000個に水増し

今後さらに実施してみたいこと

1. 現実データでの異常検知

- ・ 不均衡データでの異常検知
- ・ KDD Cup 1999 Data での異常検知
 - ・ <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

2. 時系列データの異常検知

Appendix

Appendix 1. 参考資料

- ・ Azure Machine Learning
 - ・ One-Class Support Vector Machine
 - ・ <https://msdn.microsoft.com/en-us/library/azure/dn913103.aspx>
 - ・ PCA-Based Anomaly Detection
 - ・ <https://msdn.microsoft.com/library/en-us/Dn913102.aspx>
- ・ 書籍
 - ・ 異常検知
 - ・ 「入門 機械学習による異常検知—Rによる実践ガイド」 (井出剛)
 - ・ 「異常検知と変化検知」 (井出剛／杉山将)
 - ・ 「データマイニングによる異常検知」 (山西健司)
 - ・ SVM／PCA
 - ・ 「はじめてのパターン認識」 (平井有三)
 - ・ 「カーネル多変量解析」 (赤穂昭太郎)
 - ・ 「パターン認識と機械学習 下」 (C.M.ビショップ)
- ・ その他
 - ・ 異常検知の世界へようこそ
 - ・ <https://research.preferred.jp/2013/01/outlier/>
 - ・ KDD Cup 99 Dataおぼえがき
 - ・ <https://ntddk.github.io/2016/11/23/kdd-cup-99-data/>