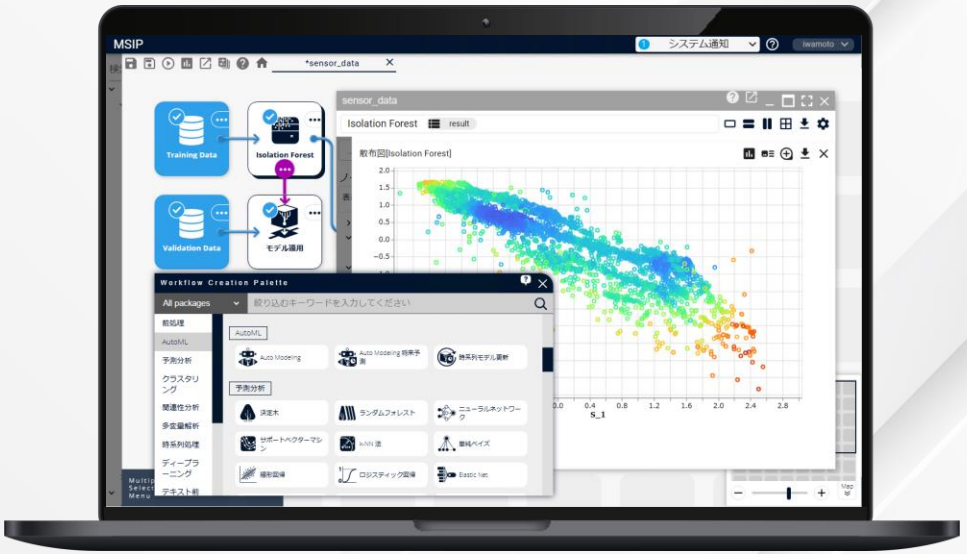
【アルカノ】
Alkano

テクニカルサンプルプロジェクト

異常検知モデルに対する
ドリフト検知



株式会社 NTTデータ数理システム

NTT DATA Mathematical Systems Inc.

このプロジェクト について

こんな方におすすめします

運用中の機械学習モデル・異常検知モデルに対し、精度低下を監視し再学習を行うタイミングを検知したい方

日々流入するデータについて、傾向の変化を数値的に確認したい方

何をするプロジェクト？

作成した機械学習モデルを使用していくと、最初のうちは精度よい予測が行えていたのに、徐々に予測の精度が落ちて来たということはありませんか。

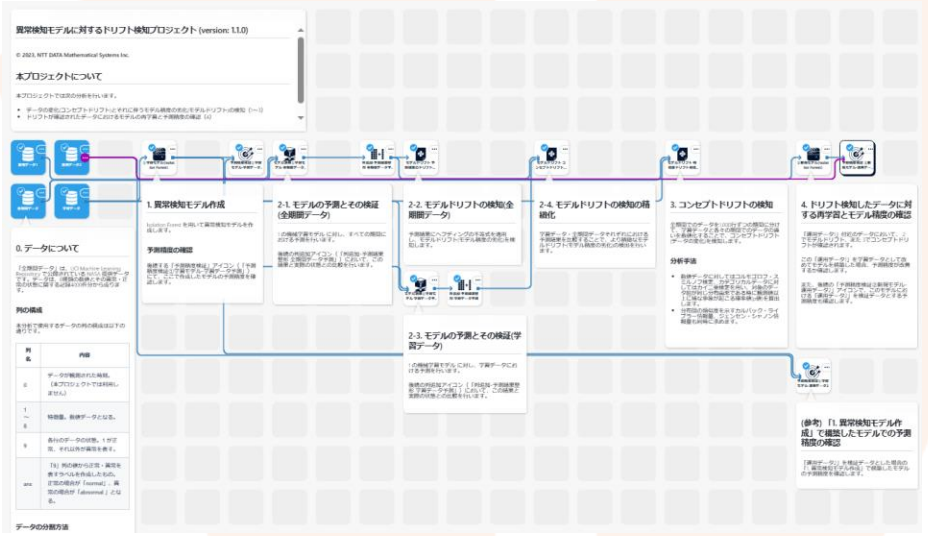
このような場合、データの変化（コンセプトドリフト）やそれに伴うモデル精度の劣化（モデルドリフト）を適切に検知し、新しいデータで再学習を行うことが一つの有効な方法とされています。

このプロジェクトでは、再学習の起点を知るためのドリフト検知手法を紹介いたします。

具体的には、入力となる特徴量、予測対象、予測結果の分布に対して

- **現在データが学習当時と同じ分布から生成されたものかを調べる**

といったことを行い、精度変化を検知し予測モデルを監視する方法を紹介いたします。



© 2023 NTT DATA Mathematical Systems Inc.

※p.28の「本文書・プロジェクトファイルのご利用にあたって」をお読みのうえ、ご利用ください。

目次

プロジェクト解説 p.4 - p.7

プロジェクトの流れについて解説します。

- 0. 対象データについて
- 1. 異常検知モデル作成
- 2. モデルドリフトの検知
- 3. コンセプトドリフトの検知
- 4. 再学習とモデル精度の確認

アウトプット p.8 - p.13

プロジェクトの出力について解説します。

- アウトプット（1.）
- アウトプット（2-2.）
- アウトプット（2-4.）
- アウトプット（3.）
- アウトプット（4.）

アイコン情報 p.14 - p.17

プロジェクトのコアとなっているアイコンについて解説します。

- 『モデルドリフト 予測結果のドリフト検知』アイコン
- 『モデルドリフト 予測結果のコンセプトドリフト検知』アイコン
- 『モデルドリフト 特徴量ドリフト検知 学習データ比較』アイコン

補足情報 p.18 - 28

プロジェクト内で使用している技術の紹介や
参照情報、利用規約について記載します。

- 1.コンセプトドリフトとモデルドリフト
- 2.異常検知とモデルドリフト検知の関係
- 3.データの比較方法
- 4.経験分布や度数分布を用いた比較
- 5.データ系列の変化を捉える
- 本文書・プロジェクトファイルのご利用にあたって

© 2023 NTT DATA Mathematical Systems Inc.

プロジェクト 解説

© 2023 NTT DATA Mathematical Systems Inc.

0. 対象データについて

UCI Machine Learning Repository で公開されている NASA から提供された数値データ 8 列(特徴量、説明変数)、ラベルデータ 1 列(目的変数)のデータを用います。ラベル 1 を正常、その他を異常として扱います。

0 列はデータが観測された時刻を示しますが、このプロジェクトでは説明変数としては利用しません。

・データ概要

列名	0	1	2	3	4	5	6	7	8	9	ans
説明	時間	数値	数値	数値	数値	数値	数値	数値	数値	ラベル	状態 (normal/abnormal)

特徴量
数値データ

1～9 に分類
1 を正常、それ以外を異常

正常/異常を normal/abnormal で入力

・データを可視化画面で表示

データは先頭から順次得られるとし

- 1- 1,000 行の期間を **学習データ**
- 2001 行- 3000 行の期間を **運用データ1**
- 3001-4000 行の期間を **運用データ 2**

とします。

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle)

© 2023 NTT DATA Mathematical Systems Inc.

プロジェクト 解説

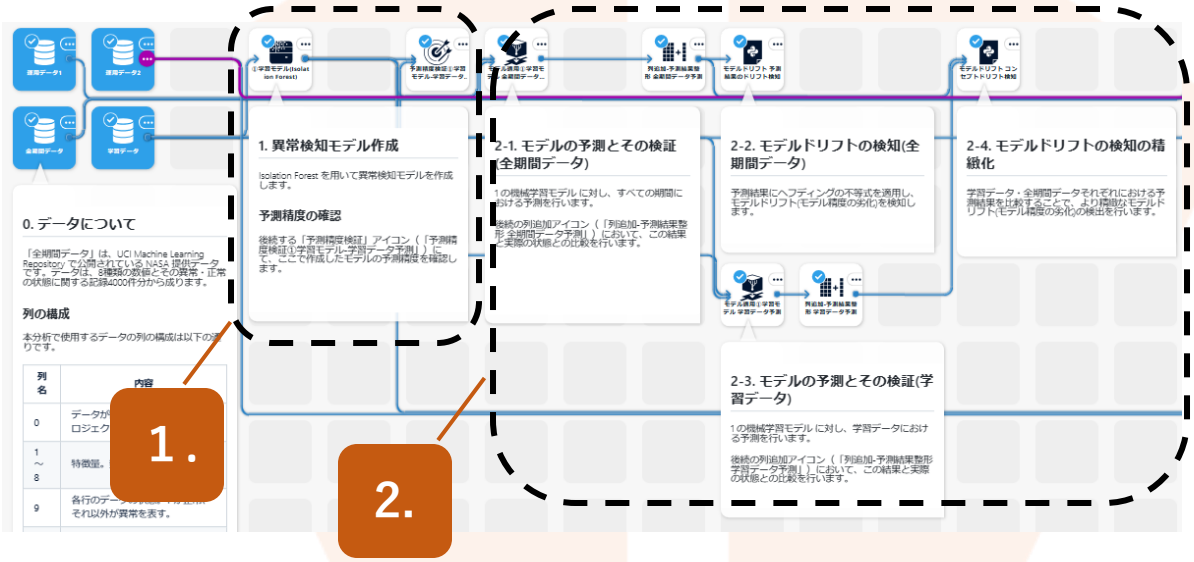
1. 異常検知モデル作成

Isolation Forest を用いて機械学習モデルを作成します。学習データを用いてモデルを作成し、予測精度検証アイコンを用いて精度を確認します【1.】。このモデルに対し、以降の操作でモデルドリフト検知を行います。

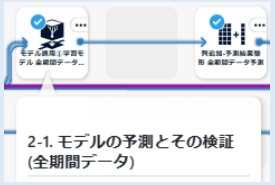
2. モデルドリフトの検知

1. で作成したモデルに対しすべての期間を通して予測を行い【2-1.】、その予測が合っているか間違っているかを評価することで、どの時点で精度が変化したか（モデルドリフトが起こっているか）を検知します【2-2.】。検知にはヘフディングの不等式を応用します。

また、予測結果だけではなく、予測に用いたデータを入力としてよりモデルドリフト検知を精緻化しています【2-3.】。ここでは、2 次元コルモゴロフ・スミルノフ検定を用います【2-4.】。



【2-1.】
この表記は、プロジェクト内コメントの対応箇所を示しています。



© 2023 NTT DATA Mathematical Systems Inc.

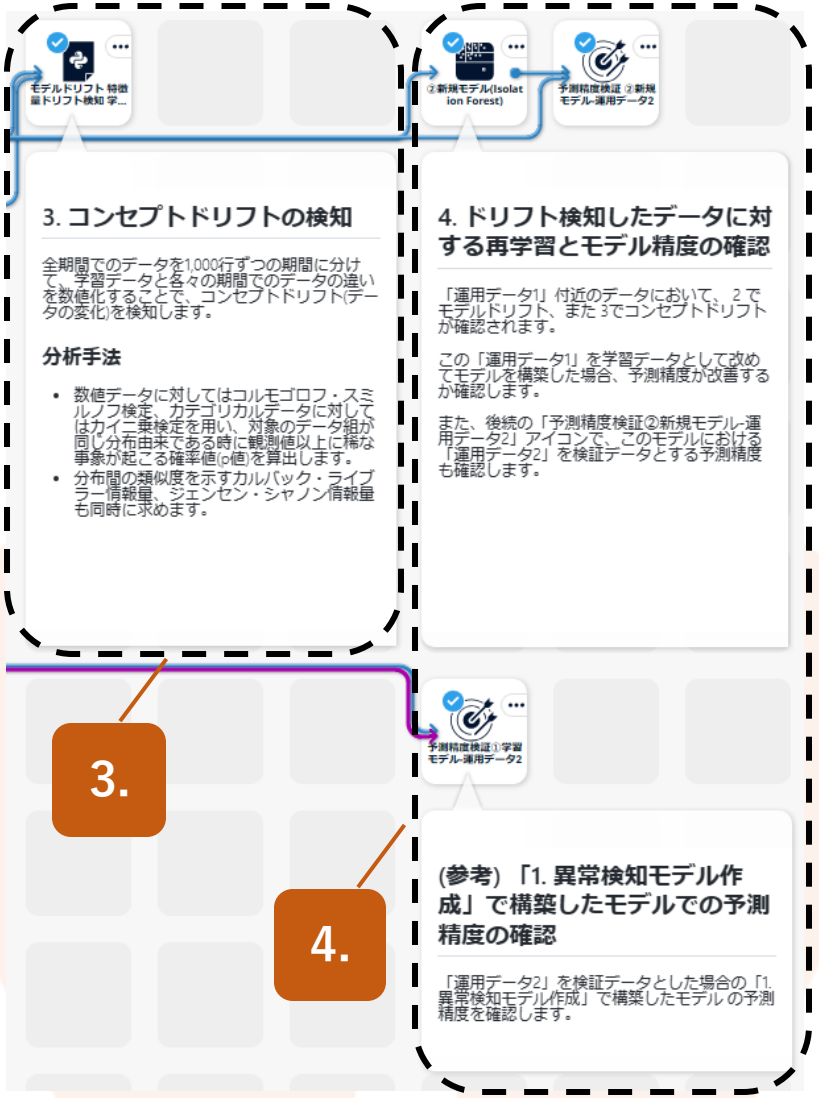
プロジェクト 解説

3. コンセプトドリフトの検知

全期間でのデータを1,000行ずつの期間に分けて、学習データと各々の期間でのデータの違いを数値化します。ここでは、数値データに対してはコルモゴロフ・スミルノフ検定、カテゴリカルデータに対してはカイ二乗検定を用い、対象のデータ組が同じ分布由来である時に観測値以上に稀な事象が起こる確率値（p値）を算出します。分布間の類似度を示すカルバック・ライブラー情報量、ジェンセン・シャノン情報量も同時に求めます【3.】。

4. 再学習とモデル精度の確認

ドリフト検知した地点でのデータを用いて再学習を行い新規に②の機械学習モデルを作成します【4.】。また、①の機械学習モデルの予測結果と比較します【参考】。



© 2023 NTT DATA Mathematical Systems Inc.

アウトプットの説明

© 2023 NTT DATA Mathematical Systems Inc.

アウトプット（1.）

学習データに対して「1. 異常検知モデル作成」で作成したモデルを用いて予測を行い、予測精度検証アイコンを適用し予測精度を算出します。

モデル作成に用いたデータそのものに対しての予測ですので、この結果をベースラインの参考値として扱うことにします。

この結果は、『予測精度検証①学習モデル-学習データ予測』に存在しています。この結果と、再学習後の結果との比較を行います。

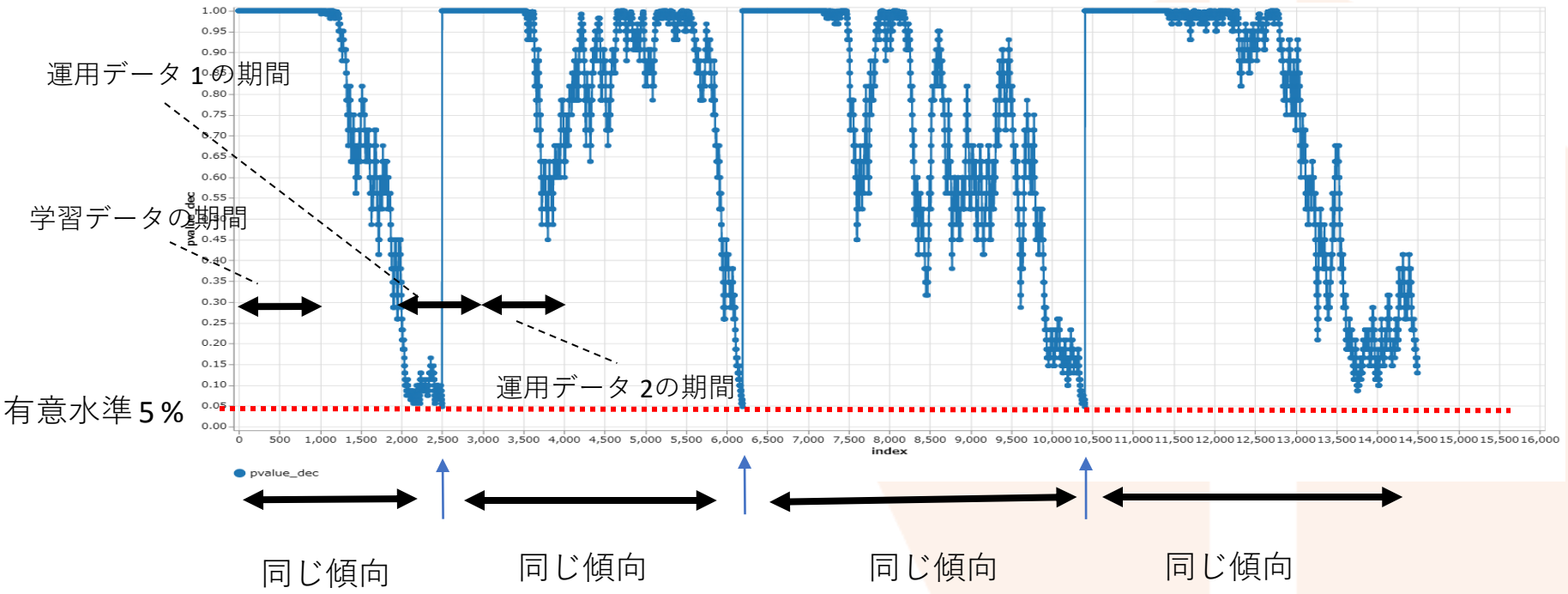
「1. 異常検知モデル作成」で作成したモデルの、学習データに対しての予測精度（scores テーブル）

No.	行名	precision FLOAT	recall FLOAT	f1-score FLOAT	support FLOAT	accuracy FLOAT
1	abnormal	0.535000	0.543147	0.539043	197.000000	*NA*
2	normal	0.887500	0.884184	0.885839	803.000000	*NA*
3	accuracy	*NA*	*NA*	*NA*	*NA*	0.817000

© 2023 NTT DATA Mathematical Systems Inc.

アウトプット（2-2.）

「1. 異常検知モデル作成」で作成したモデルに対し、すべての期間を通して予測を行い、その予測が確からしいかどうかの確率（p値）の最大値を計算します。この値は『モデルドリフト予測結果のドリフト検知』の pvalue_dec列に格納されています。
その結果をプロットしたものが以下です。可視化画面からは、『グラフ一覧』→『折れ線[モデルドリフト予測結果のドリフト検知-result:y(pvalue_dec)]』で確認できます。
pvalue_dec列の値小さいと予測精度が悪くなったと判断します。今回は有意水準を5%とし、これより確率が小さい時に予測精度が悪くなった、すなわちモデルドリフトが起こったと検知します。



有意水準 5 % として、運用データ 1 の期間(2,500 付近)で予測精度の低下が確認されました。

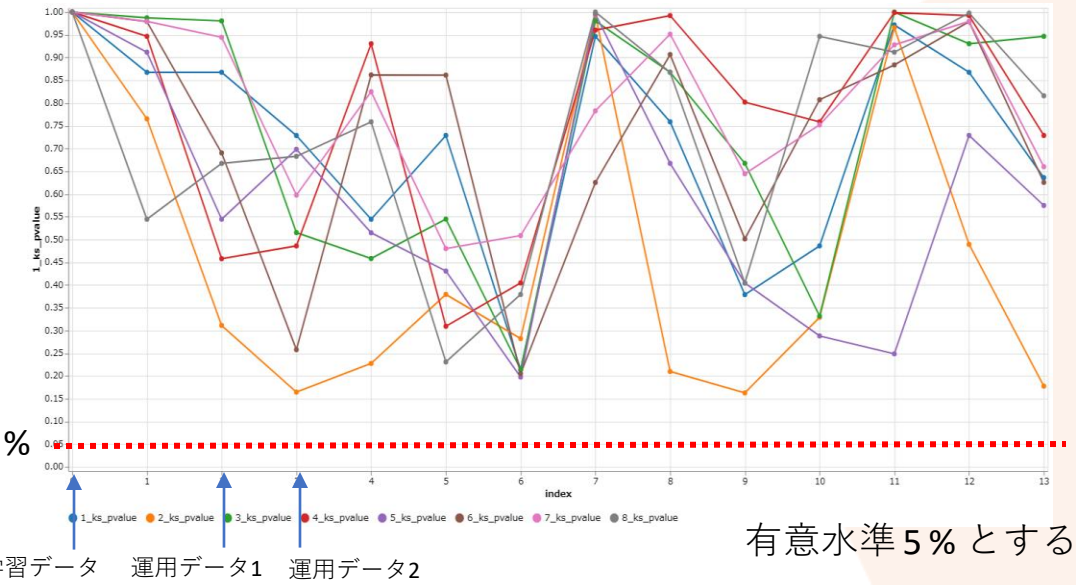
© 2023 NTT DATA Mathematical Systems Inc.

アウトプット（2-4.）

学習に用いたデータと予測結果の組が同じ分布に従うとしたとき、観測値以上に稀な事象が起こる確率値（p値）を求めています。「1」～「8」列に対応するp値が『モデルドリフトコンセプトドリフト検知』結果の1_ks_pvalue～8_ks_value列に格納されています。この値が小さいと、モデル変化が生じていないとした場合に、観測値以上に比較対象のデータから外れた値が得られる確率が低くなります。

その結果をプロットしたものが以下です。可視化画面からは、『グラフ一覧』→『折れ線[モデルドリフトコンセプトドリフト検知result:y(1_ks_pvalue,2_ks_pvalue,3_ks_pvalue,4_ks_pvalue,5_ks_pvalue,6_ks_pvalue,7_ks_pvalue,8_ks_pvalue)]』で確認きます。運用データ2付近で特徴量と予測結果の関係が変化しているように見えますが、有意水準を5%とするとドリフトは検知されていないことになります。

・学習データと全データとの比較



有意水準5％

学習データ　運用データ1　運用データ2

有意水準5％とすると、ドリフトは検知されていません。

© 2023 NTT DATA Mathematical Systems Inc.

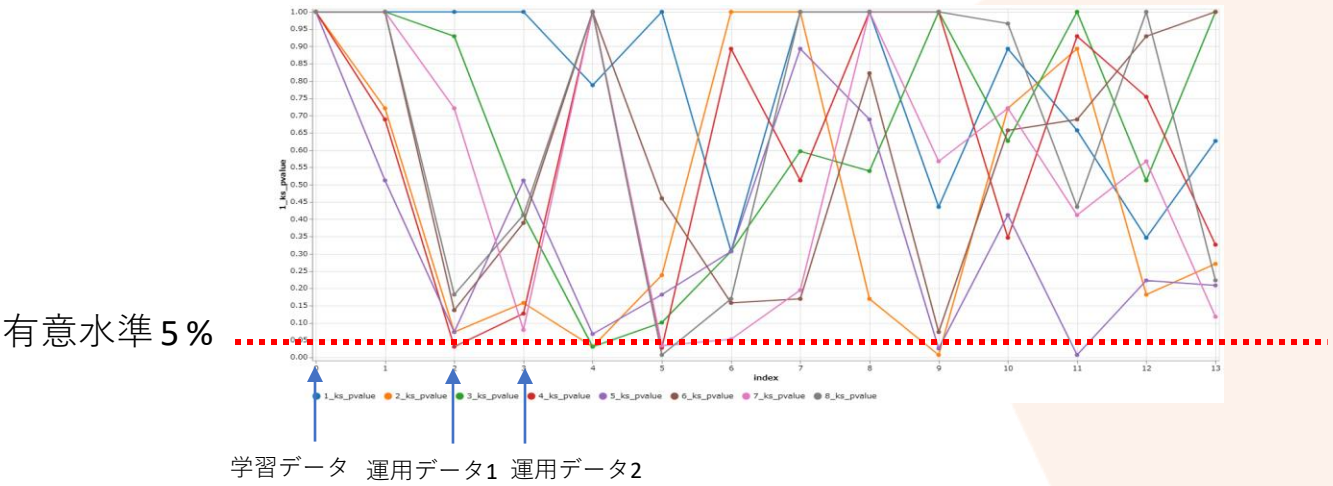
アウトプット（3.）

学習データと全期間のデータ、運用データ1と全期間のデータとの比較を行います。2つのデータが同じ分布に由来しているとしたとき、観測値以上に稀な事象が起こる確率値（p値）を求めています。「1」～「8」列に対応するp値が『モデルドリフト特徴量ドリフト検知学習データ比較』結果の1_ks_pvalue～8_ks_value列に格納されています。値が低いと、2つのデータが同じ分布であるとしたときに観測値以上に外れた値になる確率が低いことを表します。

その結果をプロットしたものが以下です。可視化画面からは、『グラフ一覧』→『折れ線[モデルドリフト特徴量ドリフト検知学習データ比較result:y(1_ks_pvalue,2_ks_pvalue,3_ks_pvalue,4_ks_pvalue,5_ks_pvalue,6_ks_pvalue,7_ks_pvalue,8_ks_pvalue)]』で確認きます。

運用データ1はアウトプット（2-4.）でモデルドリフトが検知された期間のデータです。プロットは列毎に行い、全期間データは1,000期間毎にデータを区切っています。

・学習データと全データとの比較



有意水準5％とすると、学習データと運用データ1では、傾向の違う列がみられます。

© 2023 NTT DATA Mathematical Systems Inc.

アウトプット（4.）

（2-2.）及び（3.）で、運用データ1付近でドリフトが検知されました。そこで、運用データ1を用いて再学習を行って新規に②の機械学習モデルを作成し、学習に用いなかった運用データ2において、ベースラインである「1.異常検知モデル作成」でのモデルの予測結果と比較します。

以下は各々『予測精度検証①学習モデル-運用データ2』と『予測精度検証②新規モデル-運用データ2』の結果です。
①の機械学習モデルより、再学習を行った②の機械学習モデルの方が予測精度が上がっています。

①学習モデル・運用データ 2 予測(scores テーブル)

No.	行名	precision FLOAT	recall FLOAT	f1-score FLOAT	support FLOAT	accuracy FLOAT
1	abnormal	0.500000	0.572115	0.533632	208.000000	NA
2	normal	0.883202	0.849747	0.866152	792.000000	NA
3	accuracy	NA	NA	NA	NA	0.792000

②新規モデル・運用データ 2 予測(scores テーブル)

No.	行名	precision FLOAT	recall FLOAT	f1-score FLOAT	support FLOAT	accuracy FLOAT
1	abnormal	0.521368	0.586538	0.552036	208.000000	NA
2	normal	0.887728	0.858586	0.872914	792.000000	NA
3	accuracy	NA	NA	NA	NA	0.802000

f1-score が上昇しているが、precision 及び recall ともに上昇している

今回は、（2-2.）での予測の確からしさの観点でモデルドリフトを検知しました。一方で、「学習データと予測結果の組が同じ分布にしたがうか」という観点（2-4.）ではドリフトは検知されませんでした。
（3.）の結果を考慮すると、特徴量ドリフトが起こったことによるモデルドリフトの可能性が高いと考えられます。

© 2023 NTT DATA Mathematical Systems Inc.

アイコン情報

© 2023 NTT DATA Mathematical Systems Inc.

アイコン情報

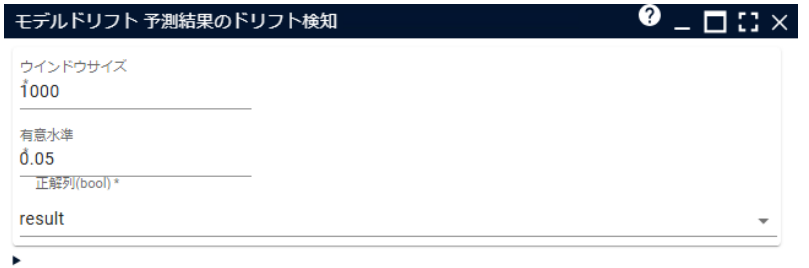
・『モデルドリフト 予測結果のドリフト検知』アイコン
概要
1つの入力データをウインドウサイズで指定した行ずつ抽出し、
先頭から1行ずつスライドしていきドリフトを検知します。
検知にはヘフディング不等式を用いた
Fast Hoeffding Drift Detection Method(FHDDM) を用います。
パラメータ設定画面は、アイコンをダブルクリックして開きます。

入力:
table : 比較データ

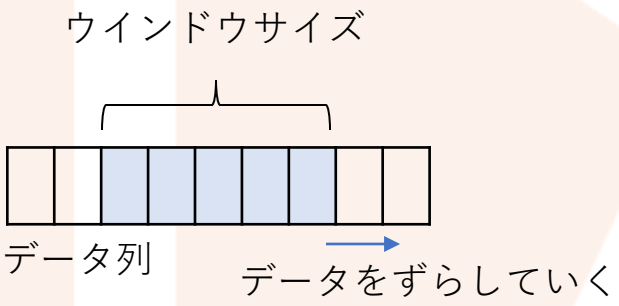
パラメータ:
ウインドウサイズ: 比較データからデータを抽出する行数、デフォルト値は 1,000
有意水準 : 検定を用いた時の有意水準、デフォルト値は0.05
正解列(bool) : 正解列名, bool 列を指定、 デフォルト値は result

出力:
ドリフト検知結果(有意水準として設定した値を用いる): detection
予測精度が向上していない確率値の最大値(p値 の最大値) : pvalue_inc
予測精度が低下していない確率値の最大値(p値 の最大値) : pvalue_dec

・パラメータ設定画面



・動作イメージ



ウインドウサイズ分だけ対象とするデータを取りだし、対象部分の正解率が以前のデータでの正解率と変化がないか判定していく

© 2023 NTT DATA Mathematical Systems Inc.

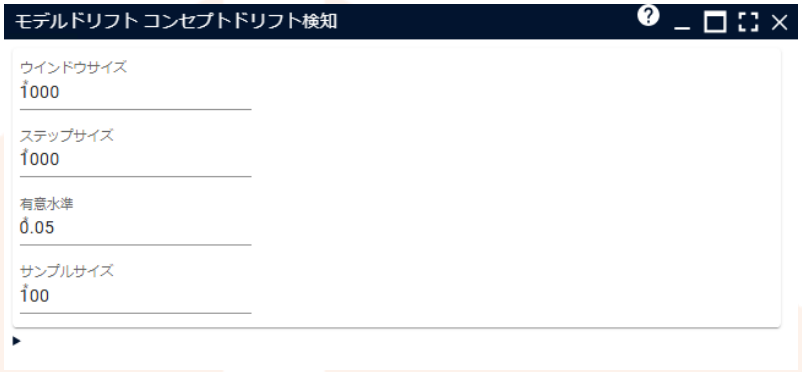
アイコン情報

・『モデルドリフト コンセプトドリフト検知』アイコン

概要

2 つの入力データの差異を計算します。
table1 を参照データとして、 table をウインドウサイズで設定した行ずつ抽出し、抽出したデータと参照データの各列とを組毎に比較します。table からの抽出は、先頭からステップサイズずつ行をずらし行います。
主に数値データを対象とします。
検知には、 2 次元コルモゴロフ・スミルノフ検定を用います。
パラメータ設定画面は、アイコンをダブルクリックして開きます。

・パラメータ設定画面

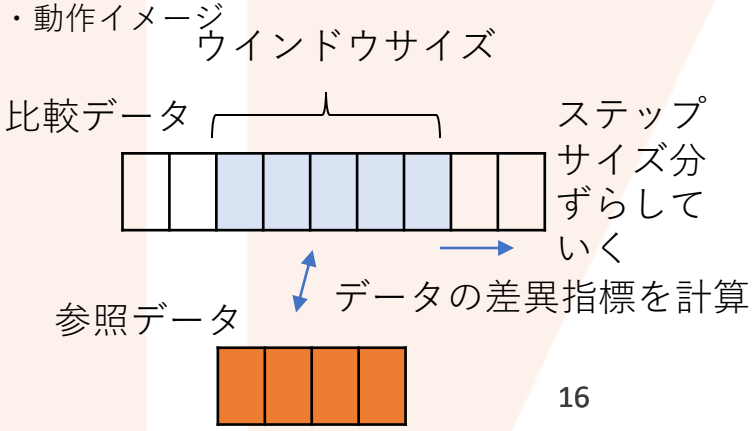


入力 :
table : 比較データ
table1 : 参照データ

パラメータ :
ウインドウサイズ :
比較データからデータを抽出する行数、デフォルト値は 1,000
ステップサイズ :
抽出する行をずらす幅、ステップサイズ、デフォルト値は 1,000
有意水準 : 検定を用いた時の有意水準、デフォルト値は 0.05
サンプルサイズ : 本サンプルでは用いない

出力 :
ドリフト検定結果(有意水準として設定した値を用いる) : (列名)_ks_detect
検定統計量 : (列名)_ks_stat
ドリフトが起きていない確率(p 値) : (列名)_ks_pvalue

・動作イメージ



© 2023 NTT DATA Mathematical Systems Inc.

アイコン情報

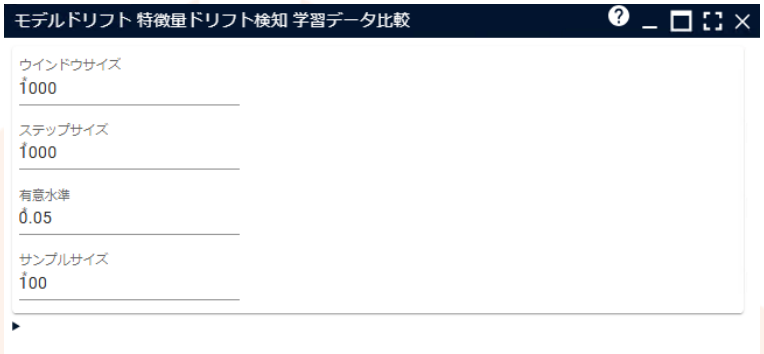
・『モデルドリフト 特徴量ドリフト検知 学習データ比較』アイコン
概要
2 つの入力データの差異を計算します。table1 を参照データとして、 table をウインドウサイズで指定した行ずつ抽出し参照データと各列毎に比較します。table からの抽出は、先頭から ステップサイズずつ行をずらし行います。
検知には、数値データの場合、コルモゴロフ・スミルノフ検定、カテゴリカルデータの場合、カイ二乗検定を用い、
同時にカルバック・ライブラー情報量 及びジェンセン・シャノン情報量 を算出します。
パラメータ設定画面は、アイコンをダブルクリックして開きます。

入力 :
table : 比較データ : table
table1 : 参照データ : table1

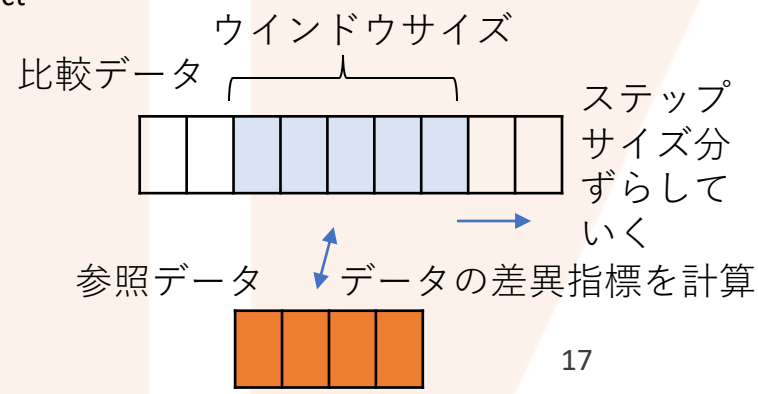
パラメータ :
ウインドウサイズ : 比較データからデータを抽出する行数、デフォルト値は 1,000
ステップサイズ : 抽出する行数をずらす幅、ステップサイズ、デフォルト値は 1,000
有意水準 : 検定を用いた時の有意水準 、デフォルト値は 0.05
サンプルサイズ : 本サンプルでは用いない

出力 :
- 数値データの場合
コルモゴロフ・スミルノフ検定結果(有意水準として設定した値を用いる) : (列名)_ks_detect
コルモゴロフ・スミルノフ検定・検定統計量 : (列名)_ks_stat
コルモゴロフ・スミルノフ検定・p 値(ドリフトが起きていない確率) : (列名)_ks_pvalue
- カテゴリカルデータの場合
カイ二乗検定結果(有意水準として設定した値を用いる) : (列名)_chi2_detect
カイ二乗検定結果・検定統計量 : (列名)_chi2_stat
カイ二乗検定検定・p 値(ドリフトが起きていない確率) : (列名)_chi2_pvalue
カルバック・ライブラー情報量 : (列名)_kl_div
- 共通
ジェンセン・シャノン情報量 : (列名)_js_div

・パラメータ設定画面



・動作イメージ



© 2023 NTT DATA Mathematical Systems Inc.

補足情報

技術的な情報や利用規約について

© 2023 NTT DATA Mathematical Systems Inc.

1. コンセプトドリフトとモデルドリフト

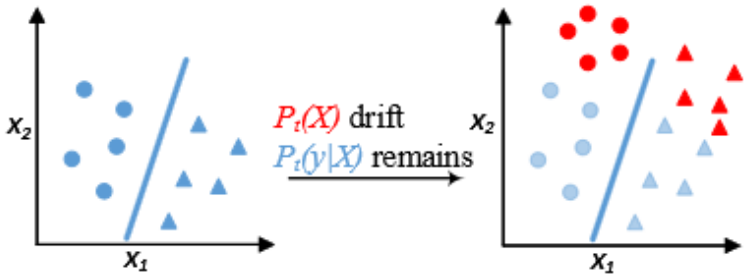
コンセプトドリフトとは、時間経過とともにデータの傾向が変化すること、また、それに伴いモデルの予測精度が劣化することをモデルドリフトと呼びます。

コンセプトドリフトやモデルドリフトは、以下の 2 つの変化によって起こるとし、検知手法が開発されています。

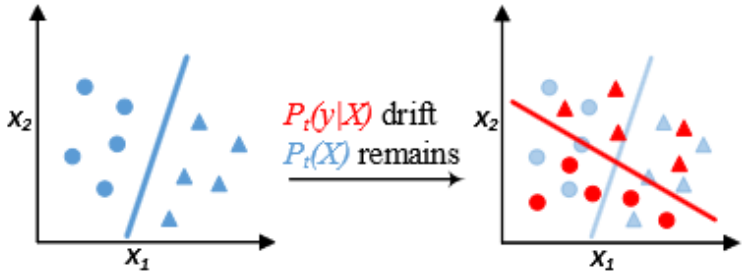
・特徴量ドリフト、共変量シフト ：特徴量の分布が変化すること

・(真の)コンセプトドリフト　　：特徴量と予測対象との関係が変化すること

●ドリフトのイメージ

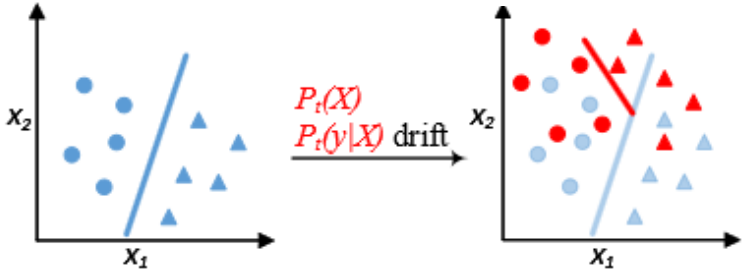


時刻 t　　特徴量ドリフト　　時刻 t+1



時刻 t　　時刻 t+1

コンセプトドリフト



時刻 t　　時刻 t+1

特徴量ドリフトとコンセプトドリフト(同時)

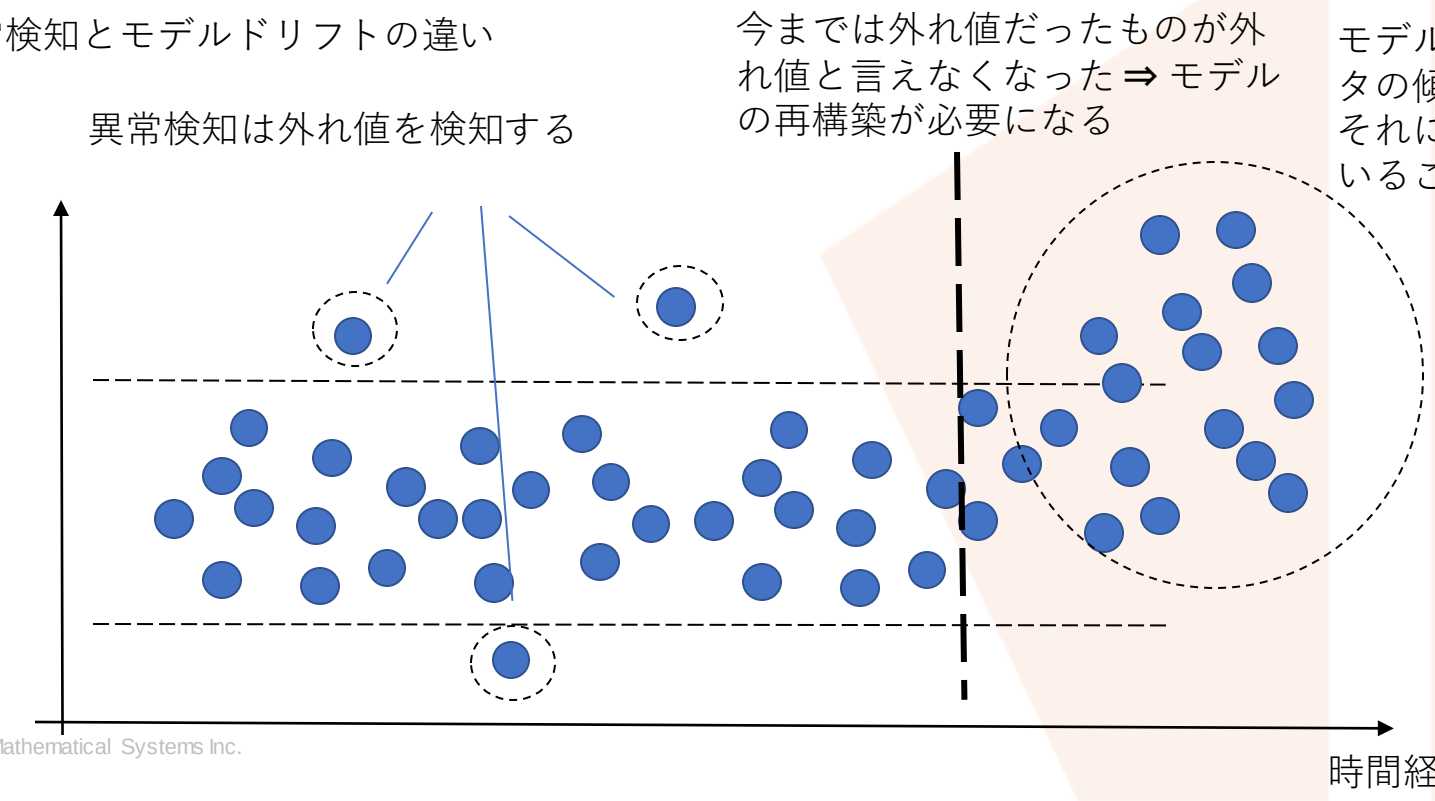
Lu, Jie, et al. "Learning under concept drift: A review." *IEEE Transactions on Knowledge and Data Engineering* 31.12 (2018): 2346-2363.

© 2

2. 異常検知とモデルドリフト検知の関係

異常検知は、データの中から外れ値を検知し、モデルドリフト検知は今までのデータの傾向からのずれを検知します。
どちらも入って来るデータに対して、今までのデータと違っているかどうかを判定する手法ですが、異常検知は元々の分布は同じですが、モデルドリフトの場合、データの分布自体が変化する状況を検知します。
実際にやっていることは似ているのですが、求められる精度や計算速度の違いから、採用する手法にも違いがあります。

●異常検知とモデルドリフトの違い



© 2023 NTT DATA Mathematical Systems Inc.

3. データの比較方法

データの集まりを評価する時、データを数値やラベルの列として扱うのではなく、経験分布や度数分布のような分布関数の形にして評価することがあります。

数値列の場合は、経験分布と呼ばれるデータ点が増える毎に階段上に分布の値を増やしていく分布を作成し、これを累積分布関数として扱ったり、一定の区間ごとに分けその区間の中にあるデータ点の数で分布関数を構成するヒストグラムが使われます。

カテゴリカルデータの場合、度数分布や割合を用います。

・数値データ

元データ

8	2	6	0	4

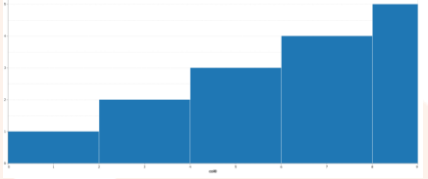
→

昇順に並べ替え

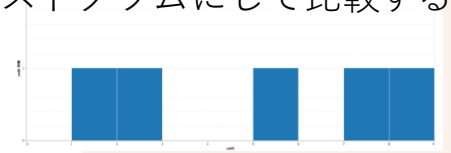
0	2	4	6	8

→

経験分布関数にして比較する



ヒストグラムにして比較する



・カテゴリカルデータ

元データ

A	B	B	C	B

→

度数分布や割合にして比較する

A	B	C
1	3	1

A	B	C
0.2	0.6	0.2

© 2023 NTT DATA Mathematical Systems Inc.

4. 経験分布や度数分布を用いた比較

コルモゴロフ・スミルノフ検定

2 つのデータがあった時に、経験分布ががそれぞれ$F_n(x), F_m(x)$と表されるとすると、

2 つのデータが同じ分布から生成されたとしたとき、差の絶対値の上界

$$D_{nm} = \sup_x |F_n(x) - F_m(x)|$$

の確率が、n, mが十分大きい時経験分布がブラウン橋過程に従うことを利用し、漸近的に

$$Prob(\sqrt{\frac{mn}{m+n}} D_{nm} \leq x) \to \frac{\sqrt{2\pi}}{x} \sum_{i=1}^{\infty} \exp(-\frac{(2i-1)^2\pi^2}{8x^2})$$

となることを用います。

2 つのデータが同じ分布から生成されていることを前提として(帰無仮説)、

得られたデータが、あらかじめ決めたある水準(有意水準)より実現する確率が低い場合に、

同じ分布から生成されていない(対立仮説)を採用します。

Marsaglia, George, Wai Wan Tsang, and Jingbo Wang. "Evaluating Kolmogorov's distribution." *Journal of statistical software* 8 (2003): 1-4.
Dmitry Panchenko, MIT OpenCourseWare “Mathematics Theory of Probability, Spring 2005 Lecture 29”, https://dspace.mit.edu/bitstream/handle/1721.1/37302/18-175Spring-2005/OcwWeb/Mathematics/18-175Spring-2005/CourseHome/index.htm

© 2023 NTT DATA Mathematical Systems Inc.

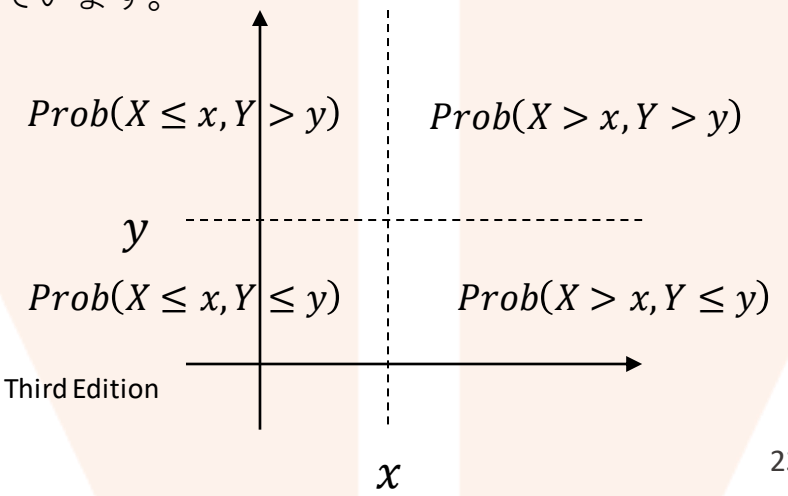
4. 経験分布や度数分布を用いた比較

2 次元・2 標本コルモゴロフ・スミルノフ検定

1 次元の場合は、軸が 1 つでありデータの傾向を表す経験分布を容易に決まりますが、 2 次元の場合、以下の 4 つの経験分布毎に傾向が違ってきます。

$$F(x,y) = Prob(X \leq x, Y \leq y)$$
$$F(x,y) = Prob(X > x, Y \leq y)$$
$$F(x,y) = Prob(X > x, Y > y)$$
$$F(x,y) = Prob(X \leq x, Y > y)$$

なので、 4 つの場合を各々に評価し、一番確率の低い分布関数を採用することで、 2 次元の場合の検定を実現しています。また、計算量自体の削減の工夫もされています。



Press,Teukolsky, Vetterlling, and Flannery, “Numerical Recipes: The Art of Scientific Computing, Third Edition in C++ (2007)”, Cambridge University Press, pp.762-766

© 2023 NTT DATA Mathematical Systems Inc.

4. 経験分布や度数分布を用いた比較

2 標本カイ二乗検定

カテゴリカルデータの比較に2標本カイ二乗検定を用います。

検定統計量として

	1	2	…	k	総数
サンプル1	R_1	R_1	…	R_k	$\sum_{i=1}^{k} R_i$
サンプル2	S_1	S_2	…	S_k	$\sum_{i=1}^{k} S_i$

$$\chi^2 = \sum_{i=1}^{k} \frac{(K_1 R_i - K_2 R_i)^2}{R_i + S_i}, K_1 \equiv \frac{\sum_{i=1}^{k} S_i}{\sum_{i=1}^{k} R_i}, K_2 \equiv \frac{\sum_{i=1}^{k} R_i}{\sum_{i=1}^{k} S_i}$$

が、サンプル1とサンプル2が同じ分布から生成されているとすると、

サンプル1とサンプル2が同じ総数の時、自由度k-1のカイ二乗分布に従い、

総数が異なる場合、自由度kのカイ二乗分布に漸近的に従うことを用います。

サンプル1とサンプル2が同じ分布から生成されていることを前提として(帰無仮説)、

得られたサンプルがあらかじめ決めたある水準より(有意水準)実現する確率が低い場合に、

同じ分布から生成されていない(対立仮説)を採用します。

Press,Teukolsky, Vetterlling, and Flannery, “Numerical Recipes: The Art of Scientific Computing, Third Edition in C++ (2007)”, Cambridge University Press, pp.730-734

© 2023 NTT DATA Mathematical Systems Inc.

4. 経験分布や度数を用いた比較

カルバック・ライブラー情報量

2 つの確率分布 がどの程度似ているかを表す尺度です。分布$p(x)$を基準に分布q(x)がどの程度違っているかを以下の式で測ります。

$$KL(p||q) = \int p(x) log \frac{p(x)}{q(x)} dx$$

離散分布の場合、基準となる分布をP、比較する分布をQとなります。

$$KL(P||Q) = \sum_i P_i log \frac{P_i}{Q_i}$$

以下の特徴を持ちます。

$$KL(p||q) \geq 0$$

$$KL(p||q) = 0 \Leftrightarrow p = q$$

$$KL(p||q) \neq KL(q||p)$$

© 2023 NTT DATA Mathematical Systems Inc.

4. 経験分布や度数を用いた比較

ジェンセン・シャノン情報量

カルバック・ライブラー情報量は対称性を持たないため使いづらい面があるのですが、対称性を持たせたものになります。

$$JS(p||q;\lambda) = \lambda KL(q||\lambda q + (1-\lambda)p) + (1-\lambda)KL(p||\lambda q + (1-\lambda)p)$$

$\lambda = \frac{1}{2}$ の時対称になります。

$$JS(p||q;\frac{1}{2}) = JS(q||p;\frac{1}{2})$$

© 2023 NTT DATA Mathematical Systems Inc.

5. データ系列の変化を捉える

ヘフディングの不等式

独立な確率変数$X_1, \cdots, X_n$とし、 $x_i \in [a, b]$の範囲である時、任意のtの実数に対して漸近的に以下が成り立ちます。

$$Prob\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \geq t\right) \leq \exp(-\frac{2nt^2}{(b-a)^2})$$

E はアンサンブルを表し、真の値を取得することを意味します。

上式をヘフディングの不等式と呼び、サンプルプロジェクトでは、この式を応用し、分類機の正解不正解の情報を持つバイナリデータ列を扱い、データ列に変化が現れたかどうかを Fast Hoeffding Drift Detection Method(FHDDM) を用いて検知する例を記載しています。 FHDDM では、データ列に変化が起きない場合、$\frac{1}{n}\sum_{i=1}^{n} X_i$ の値は同じか上昇すると仮定し、$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$をデータ列の最大値として、その値と比較しドリフトが起きたかどうかを検知しています。

検知基準としては、許容できる確率値を$\alpha \in (0,1)$とし 、確率値が$\alpha = \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$以下になる場合、つまり、$\frac{1}{n}\sum_{i=1}^{n} X_i$と$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$との差が$\mathrm{t} = \sqrt{\frac{-(\mathrm{b}-\mathrm{a})^2}{2\mathrm{n}}} log\alpha$以上になる場合にドリフトを検知したと考えます。アイコンでは、不等式であることから$\exp(-\frac{2nt^2}{(b-a)^2})$を p 値の最大値と捉え、 αを最大値での有意水準として記載しています。

Pesaranghader, Ali, and Herna L. Viktor. "Fast hoeffding drift detection method for evolving data streams." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II 16. Springer International Publishing, 2016.

© 2023 NTT DATA Mathematical Systems Inc.

本文書・プロジェクトファイルのご利用にあたって

本文書ならびにプロジェクトファイルは、（株）NTT データ数理システム （以下「弊社」）が開発・販売する分析プラットフォーム Alkano についての情報提供として弊社が作成を行ったものです。弊社による事前の許可なしに、本文書の再配布や引用の範囲を超える複製といった行為、およびリバースエンジニアリングを禁じます。

本文書ならびにプロジェクトファイルのご利用に際して、ご利用者様および第三者に損害が発生したとしても、弊社は責任を負わないものとします。

プロジェクトファイルは、その中に同梱されているデータを利用し、本文書内で解説している設定可能なパラメータで動作させた場合についてのみ、弊社にて動作の検証を行っております。これを超えるような状況における動作は保証いたしません。

本プロジェクトファイル は、MSIP1.10.0 および Alkano1.4.0 にて動作確認を行っております。

© 2023 NTT DATA Mathematical Systems Inc.

【アルカノ】

Alkano

データ活用の確かなパートナー

お問い合わせ・詳細:
株式会社NTTデータ数理システム
https://www.msi.co.jp/solution/alkano/