テクニカルサンプルプロジェクト

テキストの話題分析



株式会社 NTTデータ数理システム

NTT Data © 2025 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

- ・ テキストデータや属性データを利用して、予測モデルを作成したい方
- ・ テキストデータを利用した機械学習を行いたい方

何をするプロジェクト?

このプロジェクトでは、テキストデータを利用して、機械学習 の有名な手法であるサポートベクターマシンで分類・予測モ デルを作成する一連の流れを紹介します。 この流れは、テキストデータを利用した機械学習の一般的 なフローとなりますので、これを応用することで様々な機械 学習手法を用いてテキストデータを扱うことができます。 また、付随する属性データも機械学習で利用し、分析す ることができます。



プロジェクトの解説

© 2025 NTT DATA Mathematical Systems Inc.

プロジェクトを構成する要素

本プロジェクトは大きく分けて以下の3つの要素に分けられます。

本サンプルプロジェクトでは、テキストの分割の粒度を変えて予測を行っています。「自動連結あり」の結果は、「形態素・構文解析+」アイ コンにて分割の粒度を大きく設定したフロー、「自動連結なし」の結果は、分割の粒度を小さく設定したフローとなっています。

次ページからは「自動連結あり」の結果のフローについて、各要素を構成するアイコンについて説明します。

対象データ			テキスト前処理			分類と予測	IJ
E	Стр. — — — — — — — — — — — — — —	RED 20902					2
1.データについて ▲ ホ分析では、ホームペーカリーに対するレビューデータを使用します。 ボータの列構成 ボータの列構成 ホローカリーに対するレビューデータを使用します。 アータの列構成 10 レビューごとに対与される四	3. テキストの分割 ▲ 「税務素・様文解析+」アイコン 「ビビュー」列のテキストを単語単位に 分割します。さらに、単語の公開や部り ▼	4. 分析に用いる単語の抽出 ▲ 「諸句のフィルタリング」アイ コン マモス上金分裂して守られた単原から. ▼	5. テキストのベクトル化と腐性データの 付与 ロエのにのアイコンを利用して、テキストを登場のペジトル にを変換した後、分類・予測に除いる開作テータを接合しま ・ は、「マージー予想対象現性」アイコン ・ は、「マージー予想対象現性」アイコン ・ は、「マージー予想対象現性」アイコン ・ いがを用いて、文章タベクトル(含価摂取)に変通します。 ガボートベクターマシンと介持体行うとなく、アドリックス 和学 (現件アータ) におします。	 6.学習データと検証データの作成 「データ分割」アイコン ↑ ↑ ↑ ↑ ↑ (* 金沢竹焼き(売,5,7) = 0 	 7.学習データに重み列を追加 コエロロのアイコンを利用して、分別体の学校データに、 が確心データが認知になる場合が必然します。 これにより、データの不可能を解消する物いがあります。 (1)「除計」アイコン (2)「ゆー3」アイコン (a)「集計」アイコン (j)「集計」アイコン (j)「集計」アイコン 	8. 分類モデルの作成 A 「サポートへクターマシン」アイ コン v	9. 評価の予 「モデル總用 ^{検ロデークに対し} ^{検ロデークに対し}
PE レビューの### 中国 レビューの### 中国 レビュワー編#: 「自動連結 年代 レビュワー編#: 「自動連結 第代 レビューが兼め集団の講解	あり」の結果 なし」の結果				(b) 「96歳加」アイコン (a) 空気振力会: 急行端における天正公に中心的中心が用い、 (b) 「96歳加」アイコン	VIII	

プロジェクト解説 — 対象データ

1. HBレビューデータ.dft

ECサイトで様々なホームベーカリーに対してのレ ビューをまとめた、仮想の口コミデータです。MSIP の上では、csv形式のデータをdft形式に変換し、 シナリオ編集エリア上に配置して使用します。1行 が1レビューに対応します。

今回は口コミテキストの入ったレビュー列を利用し ます。データに含まれる列の詳細については、右 の表をご覧ください。

2. HBレビューデータ_ユーザー辞書.dft

既存の辞書にはないような、ユーザー独自の単 語を追加するためのデータです。テキストの分割 処理を行った結果、つながって出てきてほしい複 合語が、いくつかの単語として分かれて出てきてし まう場合などに利用します。

列名	内容
ID	レビューごとに付与され るID
評価	レビューの評価列(1 ~5の5段階)
製品名	レビュー対象の製品名 や型番
性別	レビュワー情報 : 「男 性」もしくは「女性」
年代	レビュワー情報 : 10歳 刻みの年代
購入価格	レビュー対象の製品の 価格
書き込み日	レビューが投稿された日 付
レビュー	レビュー内容 分析対象のテキスト列
価格分類	価格を1万円ごとにまと めた価格帯
メーカー名	製品名から取得された メーカー名
評価まとめ	「評価」列を「5」「4」 「低評価」にまとめた列

TextExtension



データの列構成

本データの列構成は以下の通りです。

列名	内容
ID	レビューごとに付与されるIC
評価	レビューの評価列(1~5の5
製品名	レビュー対象の製品名や型番
性別	レビュワー情報:「男性」も
年代	レビュワー情報:10歳刻みの
購入価格	レビュー対象の製品の価格
書き込み日	レビューが投稿された日付
レビュー	レビュー内容(※分析対象の
価格分類	価格を1万円ごとにまとめた
メーカー名	製品名から取得されたメーカ
評価まとめ	「評価」列を「5」「4」「個

2.	HBレビューデータ ユーザー辞書。dt	

HBレビューラ	データ_ユーザー辞書.dft	t-data 列数: 2 行数: 2
No.	表記 Category	品詞 Category
1	パン焼き機	名詞 一般
2	a lot of	形容詞 一般

プロジェクト解説 — テキスト前処理

3. テキストの分割

テキストデータを分析する際、記載されている文章の長さや内容が統一され ていないため、テキストデータそのままでは分析を行うことができません。そこで、 「形態素・構文解析+」アイコンを利用して、テキストデータを文節単位に分 割し、単語や品詞、係り受け情報を抽出します。詳細は補足情報の『テキ ストのベクトル化:「自動連結あり」と「自動連結なし」の違い』をご参照くだ さい。

今回は、「HBレビューデータ.dft」データの「レビュー」列に入っているテキスト データが分割の対象です。

4. 分析に用いる単語の抽出

対象とする単語を品詞と頻度の観点から絞り込みます。ここでは意味のある 単語でベクトルを作成するために、品詞が「名詞」「動詞系」「形容詞・形容 動詞系」「副詞」の単語を取り出しています。更に、その中でベクトルの次 元数を調整するため、頻度2~100、文字数2以上の単語のみを取り出し ています。



プロジェクト解説 ― テキスト前処理

5. テキストのベクトル化と属性データの付与

テキストを数値のベクトルに変換した後、分類・予測に用いる属性 データを結合します。

(a) ここでは、BoW (Bag of Words) を用いて、「どの単語が 何件出現しているかという数値のベクトル表現を獲得します。 また、分析アイコンであるサポートベクターマシンの入力として利用 するために、横持データであるマトリックス形式のベクトル表現を作 成します。

(b) (a)の結果に、分類・予測に用いたい属性を 性データには、サポートベクターマシンの予測対象 である「評価まとめ」が含まれます。



	分かち書きの	Dフィルタリング-result	列数: 13 行数: 7,08	1) 🛥 🛓 🗙
	TokenID Integer	form String	lemma Category	replaced Category	pos Category	pos_detail Category
に用いたい属性を紐づけます。 属	1	もともと	ちともと	ಕಿಂಗಿ ಕೆಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೆಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೆಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೊಂಗಿ ಕೆಂಗಿ ಕೊಂಗಿ ಕೊ	副詞	
	2	焼き立ての	焼き立て	焼き立て	名詞	一般
フここの又測み合 (口的亦粉)	5	オーブンで	オーブン	オーブン	名詞	一般
マンノのア測刈豕(日的変数)	7	オーブンが	オーブン	オーブン	名詞	一般
	8	故障したのと、	故障	故障	名詞	サ変可能
ਰ੍ਹ	12	パン焼きに	パン焼き	パン焼き	名詞	一般
	14	ホームペーカリを	ボームペーカリ ***	ホームペーカリ	26月	-#5
	2	思んに ポイントけ	送えていた	選ぶ	之词	
	4	11/12/18	10121		-0.04	
	文章ベクトル	L/化-result 列数: 1,069:	行数: 363			• ± >
	No.	RowID Integer	もともと Integer	焼き立 Integer	て オーブン r Integer	
	1	1		1	1 2	
「立聿ベクトリル」マイコンで畄語列に	じけ	2		0	0 0	
又盲へノリバレノノコノし半品ノリレ	旧た	3		0	0 0	
た「replaced 列の単語が列名となり	、その	4		0	0 0	
	+ +	6		0	0 0	
申 語の 山 現 凹 叙 か て の 列 の 恒 に な り さ	ま9。	7		0	0 0	
	8	8		0	0 0	
	9	9		0	1 0	

プロジェクト解説 — テキスト前処理 学習データと検証データ

6. 学習データと検証データの作成

学習データと検証データを8:2に分割します。データ分割の比率は、 用いるデータ件数や問題設定により、おおよそ5:5~9:1の範囲で違いはありますが、7:3や8:2が多く用いられる傾向にあります。



プロジェクト解説 ― テキスト前処理 学習データと検証データ

7. 学習データに重み列を追加

分割後の学習データに、各評価のデータ件数に応じた重み列を追加 します。これにより、データの不均衡を解消する狙いがあります。 重み列は、各評価の件数を「集計」アイコンで集計し、その逆数を重 み値として採用することで、サポートベクターマシンの重み列指定に利 用することができます。

- (a) 「評価まとめ」の値ごとにデータの件数を集計します。
- (b) (a)の結果から、各評価に対するデータ件数の逆数を計算し、 重み列を作成します。作成した重み列は「集計」アイコンの結果 に追加されます。
- (c) (a)と(b)の結果を結合し、学習データに重み列を追加します。



プロジェクト解説 — 分類モデルの作成

8. 分類モデルの作成

評価を予測する予測モデルを構築します。予測モデルには決定木や ランダムフォレスト、ディープラーニングなど様々な手法がありますが、今 回はデータ数が多くないため、ある程度の件数でも精度が見込めるサ ポートベクターマシンを採用しています。



プロジェクト解説 — 予測

9. 評価の予測

サポートベクターマシンで構築したモデルを検証データに適用し、予測 値を出力します。今回はデータ分割後の検証データを利用しています が、(同じ形式の)新規の口コミデータがある場合、それを入力とし て評価値の予測を行うことができます。

9.	
9. 評価の予測 「モデル適用」 アイコン	

アウトプットの説明

© 2025 NTT DATA Mathematical Systems Inc.

アウトプット(サポートベクターマシン)

「サポートベクターマシン」アイコンの結果を確認すると、predicted_value、評価まとめ、元データの順で列があることがわかります。 「predicted_value」列が予測モデルにより予測された値(予測値)、「評価まとめ」列が元データにあった値(正解値)です。

学習データに対するモデルの適用なので、全体として正しく予測できていることがわかりますが、28,29 行目では、誤った予測がされている データがあることも確認できます。

サポートべく	クターマシン-result	列数: 1,083 行数: 290			5 •	e III 🖺 👁 🛨 🗙
No.	predicted_value Category	評価まとめ _{Category}	RowID Integer	もともと Integer	焼き立て Integer	オーブン Integer
15	5	5	349	0	1	0
16	5	5	348	1	0	0
17	低評価	低評価	347	0	0	0
18	4	4	346	0	0	0
19	4	4	344	0	0	0
20	5	5	343	0	0	0
21	5	5	342	0	0	0
22	5	5	341	0	0	0
23	5	5	340	0	0	0
24	5	5	339	0	0	0
25	5	5	338	0	0	0
26	5	5	337	0	0	0
27	5	5	336	0	0	0
28	4	5	335	0	0	0
29	4	5	334	0	0	0
30	5	5	333	0	0	0
31	低評価	低評価	332	0	0	0
32	予測値	正解值	329	0	0	0

TextExtension

アウトプット(モデル適用)

「モデル適用」アイコンの結果を確認すると、predicted_value、元データの順で列があることがわかります。「predicted_value」列は予測 モデルにより予測された値(予測値)です。「サポートベクターマシン」アイコンと異なり、未知の新規データに対して予測を行っているので、 正解値の列はありません。

今回のデータの場合、データ分割により交差検証の形で分析フローが構成されているため、「モデル適用」アイコンの代わりに、「予測精度 検証」アイコンを利用して予測モデルの精度を確認することができます。

モデル適用-1	result 列数: 1,081 行数:	73				III 🖺 👁 🛨
No.	predicted_value Category	RowID	もともと Integer	焼き立て ^{Integer}	オーフン Integer	故障 Integer
1	5	4	0	0	0	0
2	5	6	0	0	0	0
3	低評価	7	0	0	0	0
4	5	9	0	1	0	0
5	5	10	0	0	0	0
6	5	13	0	0	0	0
7	5	21	0	0	0	0
8	5	27	0	0	0	0
9	4	30	0	0	0	0
10	5	43	0	0	0	0
11	5	46	0	0	0	0
12	5	52	0	0	0	0
13	4	53	0	0	0	0
14	4	54	0	0	0	0
15	5	56	0	0	0	0
16	4	57	0	0	0	0
17	5	58	0	0	0	0
18	4	62	0	0	0	0
	予測値		-	-	-	-

アイコンの設定 アイコンの入力設定や処理実行時の設定項目について

アイコン – 形態素・構文解析+_自動連結あり

インプット設定

テキストデータと辞書ファイルの設定を行います。

ここでは、分割対象の対象のテキスト列を含むデータを「table」、HB レビューデータ_ユーザー辞書を「usrdic」に指定します。

辞書はそれぞれ、類義語辞書を「syndic」、ユーザー辞書を 「usrdic」、分割辞書を「sepdic」に設定します。いずれの辞書も必 須ではありません。詳細は補足情報の『辞書ファイル』をご参照ください。

対象テキスト列

● テキスト列

分割処理の対象としたい列を指定します。1列のみの指定 が可能です。ここでは「レビュー」列を対象とします。

Input Matching (Controller				×
		table	syndic	usrdic	sepdic
HBレビューデータ.dft	data	•• @ ••			
HBレビューデータ	data			-0-	
* 複数可					

形態素・構文解析・	▶_自動連結あり			? _ ×
対象テキスト列				
テキスト列 レ	<u>-</u>	•	ļ	
※String型・Category	/型の列を選択			
言語の選択				
● 日本語 〓				
○ 英語				
■ 構文解析と自動連	結を行う 🗮			
文章の区切りとみれ	なす文字			
☑ 句点(。)	☴ ☑ 疑問符(?)	🔲 🖌 感嘆	時(!)	
□ 空白	☰□ 改行	=		
その他				=
並列処理数				
1	=			
▼ 原文参照のための	オブジェクトを出力する	=		
				展方
				1#15

アイコン - 語句のフィルタリング①

インプット設定

「形態素・構文解析+」アイコンの結果のうち分割結果のテーブルである 「result」をフィルタリング対象として「table」に指定します。

品詞フィルタ

よく利用される品詞セットは「デフォルト品詞セット」として設定されています。 名詞/動詞系/形容詞・形容動詞系/副詞の選択が可能です。詳細に設 定する場合には「オリジナル設定」を選択し、利用する品詞を個別に指定し ます。

頻度フィルタ

● 対象列

頻度を指定して抽出したい単語列を指定します。

● 最低頻度を設定

指定した値以上の出現頻度の単語を抽出します。頻度の小さい単 語を除外することでノイズを減らします。

● 最高頻度を設定

指定した値以下の出現頻度の単語を抽出します。

Input Matching C	ontroller		×
		table	
	result	-θ-	
形態素・構文解析+	originaldata		
	morphtable		
- 複数可			

語句のフィルタリング ? _ X
☑ 品詞フィルタを設定する
● デフォルト品詞セット ○ オリジナル設定
抽出する品詞
✓ 名詞 ✓ 動詞系
☑ 形容詞・形容動詞系 ☑ 副詞
☑ 頻度フィルタを設定する
対象列 replaced v ···· ※String型・Category型の列を選択
✓ 最低頻度を設定 2
☑ 最高頻度を設定 100
□ 上位N単語を除外する 5
□ 上位N単語を抽出する 20
□ 文字列フィルタを設定する
☑ 文字数フィルタを設定する
対象列 replaced 👻 🚥
※String型・Category型の列を選択
 ✓ 最小文字数を設定 2
□ 最大文字数を設定 10
□ 文字数を結果出力
実行して閉じる ▼ 保存

TextExtension

アイコン -語句のフィルタリング②

文字数フィルタ

● 対象列

文字数を指定したい単語列を指定します。ここでは「replaced」列を 選択します。

● 最小文字数を設定

対象列のうち、指定した文字数以上の単語を抽出します。ここでは、2文字以上の単語を抽出します。

input Matching (Jontroller		X	
		table		
	result	-0-		
形態素・構文解析+	originaldata			
	-			
	morphtable			
* 複数可				
語句のフィルタリング				9_×
☑ 品詞フィルタを設	定する			
◉ デフォルト品詞セット	○ オリジナル設定			
抽出する品詞				
✓ 名詞	✓ 動詞系			
✓ 形容詞・形容動詞系	✓ 副詞			
■ 頻度フィルタを設	定する			
対象列 replaced	-	•••		
※String型・Category型の	列を選択			
✓ 最低頻度を設定 2 ──				
✓ 最高頻度を設定 100				
□ 上位N単語を除外する	5			
□ 上位N単語を抽出する	20			
	処守する			
	設定する			
対象列 replaced	-			
※String型・Category型の	列を選択			
■ 最小文子数を設定 2 				
 最大文字数を設定 10)			

TextExtension

アイコン - 文章ベクトル化①

変数選択

● 単語列

ベクトル化の対象とする単語列を指定します。ここでは、類義語辞 書を適用した後の単語の列である「replaced」列を選択します。

● キー列

ベクトルを生成するためのキー列を指定します。「形態素・構文解 析+」アイコンの結果を利用する場合、以下の列を選択します。

- 1行(1セル)単位のベクトル化: RowID
- 1文単位のベクトル化: RowID, SntID
 ここでは、1行単位でベクトル化を行うため、「RowID」列を選択します。

文章ベクトル	Ł		? _ ×
変数選択			
列名		列型	単語列 キー列
RowID		整数	
SntID		整数	
TokenID		整数	
form		文字列	
lemma		カテゴリ	
replaced		カテゴリ	✓ □ -
モデルの設定			
モデル	● BoW		\equiv
	◯ tf-idf		
	⊖ SWEM		
SWEMの設定	1		
計算方法	● 平均 ○ 最大		=
乱数シード	0		生成 🗮
出力形式			
◉ マトリックス	ス形式 ○ リスト形式		≡
			実行 ▼ 保存

アイコン - 文章ベクトル化②

モデルの選択

「形態ベクトル表現のモデルを選択します。モデルの種類は、単語の出現 状況から文章データをベクトル化する手法として、

- BoW (Bag of Words)
- tf-idf (Term Frequency-Inverse Document Frequency)
 単語の埋め込み表現を利用してベクトル化する手法として、
- SWEM (Simple Word-Embedding-based Methods)
 があります。詳細はマニュアルをご参照ください。

ここでは「BoW」を選択します。

出力形式

ベクトル化したデータの出力形式を指定します。マトリックス形式はキー列 で指定した単位1行ごとにベクトル表現を出力します。リスト形式は、キー 列・単語・値の組を出力します。

サポートベクターマシンの入力はマトリックス形式のため、ここでは「マトリック ス形式」を選択します。

文章ベクトル	化		? _ ×
変数選択			
列名		列型	単語列 キー列
RowID		整数	
SntID		整数	
TokenID		整数	
form		文字列	
lemma		カテゴリ	
replaced		カテゴリ	✓ □
モデルの設定			
モデル	● BoW		=
	⊖ tf-idf		
	⊖ SWEM		
SWEMの設定	Ē		
計算方法	◉ 平均 🔿 最大		=
乱数シード	0		生成 📃
出力形式			
◉ マトリック	ス形式 🔿 リスト形式		\equiv
			実行 ▼ 保存

アイコン - マージ_予測対象属性①

接続リンクを追加

既存のノードへ接続リンクを追加するには、ノードメニュー内の「接続リ ンク追加」ボタンをクリックします。クリックすると、ワークフロー画面が接 続先選択モードとなり、接続可能なノードがフォーカスされた状態にな ります。この状態で接続先のノードをクリックすることで、そのノードへの 接続リンクが新しく追加されます。

このようにして、「マージ」アイコンに2つの入力を設定します。

インプット設定

ベクトル化したテキストデータと、分類・予測対象である目的変数を紐 づけます。

「文章ベクトル化」アイコンの結果テーブルである「result」を左テーブ ルとして「left」に、「形態素・構文解析+」アイコンの結果のうち、オリ ジナルテーブルである「original_data」を右テーブルとして「right」に 指定します。





アイコン – マージ_予測対象属性②

入力設定

フィルタリング結果と列属性変更結果のデータを紐づけます。 マージ設定

• マージモード

紐づけを行う際の結合方法を指定します。ここでは、「文章ベクトル 化」結果(インプット設定「left」)全体に、元のテキストデータの属 性データテーブル(インプット設定「right」)の該当する情報のみ を紐づけるため、「左外部結合」を指定します。

● 左テーブルマージキー列

入力設定の「left」で指定したテーブルにおいて紐づけのキーとなる 列を指定します。ここでは「RowID」を指定します。

● 右テーブルマージキー列

入力設定の「right」で指定したテーブルにおいて紐づけのキーとなる列を指定します。ここでは「RowID」を指定します。

	hing Controller			×	
			left	right	
語句のフィルタ	リング	result	•••@••		
列属	性変更	result		•• D ••	
* 複数可					
					_
マージ					? _ ×
マージ設定					
マージモード 〇	内部結合	=			
۲	左外部結合				
0	右外部結合				
0	完全外部結合				
左テーブルマージキ	⊨—列	右テーブルマ・	-ジキー列		
RowID	•	RowID		<u> </u>	×
欠損補填					
□ 整数	0	=			
□ 実数	0	=			
□ 日付	2022/12/26				
□ 日時	2022/12/26 19	:51 =			
	0				
🗌 タイムデルタ	U	\equiv			

アイコン - データ分割

分割

データ全体を学習用/検証用 に分割します。 今回は、学習用:検証用= 8:2 に設定します。

● 学習用

学習データとして保持する割合を指定します。

● 検証用

検証データとして保持する割合を指定します。

詳細設定

● 層対象列

指定した列の層(種別)ごとに分割を行います。ここでは設定しませんが、値が不均衡な列を層対象列にすることで、元データの分布と同じ分布の学習データを作成します。

データ分割		? _ ×
分割設定		
学習用	8	=
検証用	2	=
詳細設定		
グループキー列		• •••
屆対象列		-
乱数シード値	0	生成 📃
	_	
		実行 ▼ 保存

アイコン - 集計

インプット設定

重み付け列を作成したいデータを指定します。ここでは学習データに重み 付けを行うため「training」テーブルを指定します。

分析設定

指定した列の統計量を計算します。ここではキーに指定した列の件数を求めたいので、集計項目は選択しません。

キー列

カテゴリ値の集計を行うため、キー列の集計機能を利用します。

● キー列の件数

ここでは件数を集計します。

集計キー

重み付け対象となる列を指定します。複数列の指定が可能です。 ここでは「評価まとめ」列を指定します。

		table	
	training	••@••	
テージカ制	validation		
★ 複数可			

集計					? _ ×
集計項目					
列名	列型	項目数	合計	平均	最小値
RowID	integer				
もともと	integer				
焼き立て	integer				
パン	integer				
食べる	integer				
		_	_	_	•
□ 集計項目を	マトリックス形式	式で出力する	\equiv		
キー列					
🗹 キー列の4	+数 ☰				
□ キー列の4	╄数割合 📃				
集計≠−	評価まとめ《	3	•		_
オプション					
□ 重み付ける	೬する ≡				
	ſ				•
				実行	▼ 保存

アイコン - 列追加

列追加

● 追加する列名

新しく作成する列名を任意で指定します。

● 計算式

新しい列に入力する値を指定します。ここでは「集計」アイコンで得られた各属性の件数列をもとに、件数の逆数を新しく値とするために、計算式に 1/table["件数"] と記述します。

No. 評価まとめ Category	I	件数 Integer
1 5		181
2 4		58
3 低評価		51

列追加		?_×
追加する列名	計算式	
重み列	1/table["件粪	ģ"] ×
		Ŧ
入力補助		_
列名	列型	関数名 説明
評価まとめ	category	table[columnIr 列を取得します。列 [;] ^
件数	integer	table.get_valu 指定されたセルの値:
		table.nrow() 行数を取得します。
		table.ncol() 列数を取得します。
		table.colname 列名を取得します。
		table.rowname 行名を取得します。
		table.dtypes[c 列型を取得します。
		table.scale(tar 正規化を行います。
		table.scale_by 区間指定による正規
		select(conditic 値や DataFrame の症
		table.sort(colu ソートを行います。
		table.sort_ind: 行名でソートを行い:
		実行 ▼ 保存

アイコン – マージ

入力設定

学習データと重み列を紐づけます。

マージ設定

• マージモード

紐づけを行う際の結合方法を指定します。ここでは、学習データ (インプット設定「left」)全体に重み列テーブルを紐づけるため、 「左外部結合」を指定します。

● 左テーブルマージキー列

入力設定の「left」で指定したテーブルにおいて紐づけのキーとなる 列を指定します。ここでは「評価まとめ」を指定します。

● 右テーブルマージキー列

入力設定の「right」で指定したテーブルにおいて紐づけのキーとなる列を指定します。ここでは「評価まとめ」を指定します。



マ−ジ ? _X
マージ設定
マージモード 〇 内部結合 🛛 💳
● 左外部結合
○ 右外部結合
○ 完全外部結合
左テーブルマージキー列 右テーブルマージキー列
評価まとめ ▼ … 評価まとめ ▼ … ×
•

アイコン – サポートベクターマシン

変数設定

● 目的変数

分類・予測対象とする列を指定します。ここでは「評価まとめ」を指 定します。

● 説明変数

分類・予測の説明変数を指定します。ここでは、単語列をすべて選択します。性別や年代といった属性列は選択しません。

SVM

● 重み付け

学習データの行ごとに重み付けを行います。ここでは作成した重み 列を指定します。

• 列指定:「重み列」

サポートベクタ-	-マシン			?_×
変数選択				
列名		列型	目的変数 ;	説明変数
評価まとめ		カテゴリ	\checkmark	
RowID		整数		
もともと		整数		\checkmark
焼き立て		整数		\checkmark
パン		整数		\checkmark
食べる		整数		-
モデルタイプ 分類 図 入力データを出	出力に含める			=
SVM				
カーネル関数	ガウシアン			≡
分散 ♂2	⑧ 自動 ☰			
	○ 指定	0.1		
Slack変数の係数	1			≡
回帰分析の精度	0.1			=
重み付け				
○ なし	=			
○ クラス均等化				
 列指定 		重み列 		•
			実行・	保存

TextExtension

補足情報 技術的な情報や利用規約について

© 2025 NTT DATA Mathematical Systems Inc.

テキストのベクトル化:「自動連結あり」と「自動連結なし」の違い

サポートベクターマシンで分類するためには、テキストデータを数値データに変換して入力する必要があります。文章を単語に分割して、 単語毎の頻度を集計しますが、単語の切れ目が異なると、最終的に得られるベクトルも変わります。「自動連結なし」では、「自動連 結あり」に比べて、単語が細かく分割されるため、その分だけ同じ単語を持つ文章が増えて、類似性を表現しやすいベクトルを作成で きます。例えば、次の3文に対して、「自動連結あり」と「自動連結なし」でベクトル化すると下記のようになります。「自動連結なし」 で作成したベクトルは、「パン」という共通部分がありますが、「自動連結あり」では共通部分がなくベクトルからは類似していると判断が 難しくなります。

(A) コッペパンを食べる
(B) カレーパンを作る
(C) メロンパンが好き

自動連結あり		コッペパン	食べる	カレーパン	作る	メロンパン	好き	
	А	1	1	0	0	0	0	
	В	0	0	1	1	0	0	
	С	0	0	0	0	1	1	
自動連結なし		אעב	パン	食べる	カレー	作る	メロン	好き
	А	1	1	1	0	0	0	0
	В	0	1	0	1	1	0	0
	С	0	1	0	0	0	1	1

TextExtension

辞書ファイル

「形態素・構文解析+」アイコンを利用する際には、ユーザー辞書、類義語 辞書、分割辞書を利用することができます。

ユーザー辞書

単語の切れ目を変える辞書です。主に、つながって出てきてほしい複合語が、いくつかの単語として分かれて出てきてしまう場合などに利用します。

分割辞書

単語の切れ目を変えるために用いる辞書です。「構文解析と自動連結を 行う」にチェックを入れて単語の分割処理を行う際に、登録した内容に応じ て「連結しないように」します。

類義語辞書

類義語をまとめ上げるために用いる辞書です。表記ゆれのまとめ上げに有 用です。

これらの辞書はテキストの分割処理が行われる際、右図のような流れで用いられます。



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 MSIP および Alkano と TextExtension の機能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、TextExtension1.2.0 にて動作確認 を行っております。

お問い合わせ:株式会社NTTデータ数理システム 営業部 WEB: https://www.msi.co.jp/solution/analytics/index.html

株式会社 NTTデータ数理システム

NTT Data © 2025 NTT DATA Mathematical Systems Inc.