Wekaの基礎

櫻井彰人
慶應義塾大学理工学部

Weka

- ニュージーランドのワイカト大学が開発（University of Waikato, New Zealand）
- Waikato Environment of Knowledge Analysis の略
- Weka: 探求心旺盛な飛べない鳥

Weka の特徴

- Java言語で記述（使う人にとっては関係ないことですが）
  - しかし、そうはいっても、すぐどこでも動くかつ安全なことは安心材料
- フリーソフト
  - 営利目的以外には自由に使用可能。改変可
- 機能の追加が可能

Wekaの特徴（2）

- 日本語化が比較的容易（Javaがそうだから）
- 欠点：機能が少ない
  - 特に GUI（graphical user interface）が貧弱
  - 営利目的でない以上、ある程度は我慢すべし
  - 無保証（これは商用ソフトも似たようなもの）

最初に: 対象とするデータ

```
@relation 天気とテニス

@attribute 天気予報 {晴, 曇, 雨}
@attribute 気温 real
@attribute 湿度 real
@attribute 風 {強, 弱}
@attribute テニス {行う, 止め}
```

天気とテニス.arff の内容

```
@data
晴,29,85,弱,止め
晴,27,90,強,止め
曇,28,86,弱,行う
雨,21,96,弱,行う
雨,20,80,弱,行う
雨,18,70,強,止め
雨,18,70,強,止め
曇,18,65,強,行う
晴,22,95,弱,止め
晴,21,70,弱,行う
雨,24,80,弱,行う
晴,24,70,強,行う
曇,22,90,強,行う
曇,27,75,弱,行う
雨,22,91,強,止め
```

Excel の表形式で書いたもの

天気予報	温度	湿度	風	テニス
晴	29	85	弱	止め
晴	27	90	強	止め
曇	28	86	弱	行う
雨	21	96	弱	行う
雨	20	80	弱	行う
雨	18	70	強	止め
曇	18	65	強	行う
晴	22	95	弱	止め
晴	21	70	弱	行う
雨	24	80	弱	行う
晴	24	70	強	行う
曇	22	90	強	行う
曇	27	75	弱	行う
雨	22	91	強	止め

Wekaバージョンに関する注意

		メニュー	arffファイル中の2バイト文字	決定木の表示	
				日本語	英語
Weka 3.6.13	Windows	日本語化	文字化け	文字化け	yes
	others	日本語化	yes	yes	yes
Weka 3.7.13	Windows	英語	文字化け	文字化け	yes
	others	英語	yes	yes	yes

プラットフォームとして others を選んだ場合：
ファイルをダウンロード後、（全部を解凍してもよいが）weka.jar を解凍する。
そして、ある場所に、java –jar weka.jar だけを含む RunWeka.bat を作成する。
または次のスライドに示す RunWeka.bat を作成する。
起動はこれをクリックする。

なお、文字化けはプラットフォームの違いによるものではなく、起動の仕方による。
Windows版 でも RunWeka.ini 中の fileEncoding=Cp1252 を fileEncoding=SJIS とすれば、Shift JISコード文字の文字化けはしない。なお、UTF-8 を用いる場合には、勿論、fileEncoding=UTF-8 とすればよい

RunWeka.bat

次のように何もしない(メモリは1Gとっているが)コマンドで十分

```
@echo off
javaw -Xmx1024M -classpath . -jar .¥weka.jar
```

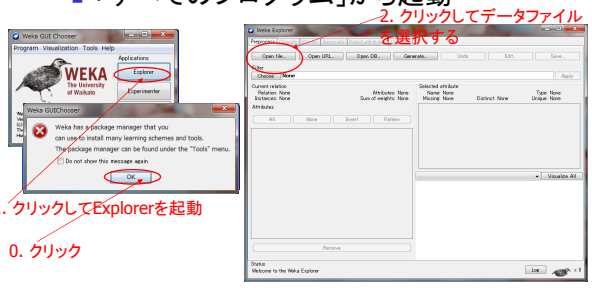
Windows版標準では RunWeka.java を用いて初期化している。

```
@echo off
set _cmd=%1
set _java=javaw
if "%_cmd%"=="" set _cmd=default
if "%_cmd%"=="-h" set _java=java
%_java% -classpath . RunWeka -i .¥RunWeka.ini -w .¥weka.jar -c %_cmd% "%2"
```

この場合、標準の RunWeka.ini 内で fileEncoding変数に Cp1252 を設定している。
すなわち fileEncoding= Cp1252 としている。
Windowsでは多くの場合、Shift-JIS コードを用いているため、これでは文字化けが起こる。
そのような場合には、SJIS を設定する。すなわち、fileEncoding=SJIS とする。なお、UTF-8 を用いる場合には、勿論、fileEncoding=UTF-8 とすればよい

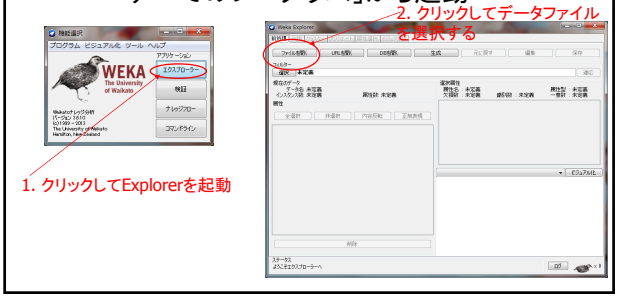
使ってみよう（Weka-3-7-13, 英語）

- 「すべてのプログラム」から起動

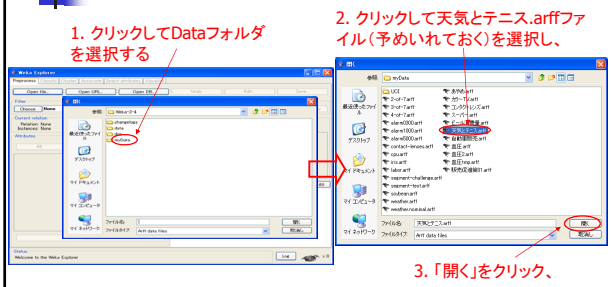


使ってみよう（Weka-3-6-13）

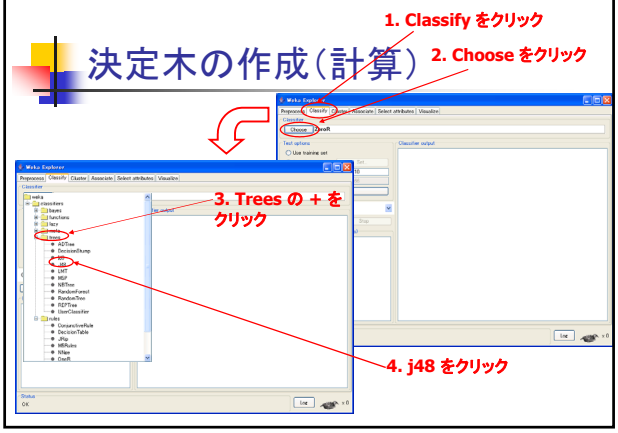
- 「すべてのプログラム」から起動



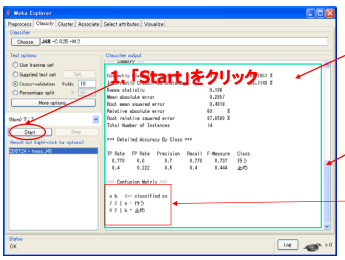
対象データファイルの指定



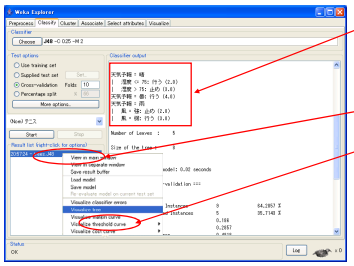
決定木の作成（計算）



結果の確認



結果の確認と図示

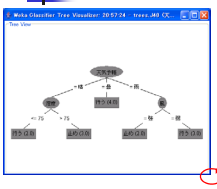


1. 決定木を文字列で表現したもの

2. この上で「右」クリック

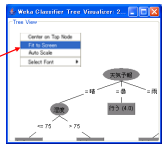
3. 「Visualize tree」「木構造をビジュアル化」の上でクリック

図示された木の変形



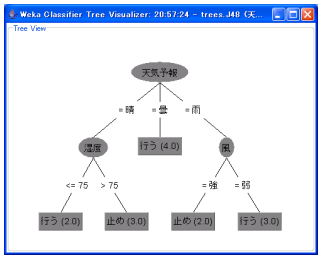
1. マウスカーソルをこの角にもってくると ↘ に変わる。その状態でドラッグすると、このウィンドウの形・大きさが変更できる

2. このスクリーン上で「右」クリック。Fit to Screen をクリックすると、スクリーンの大きさにあった大きさの木になり、Auto Scale でクリックすると木が適度にコンパクトになる。文字の大きさを変えるには Select Font でクリック
木をドラッグすることもできる



3.6.11 では働かない

決定木の例

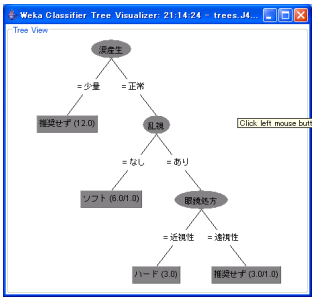


意味：
天気予報が雨であれば
　そして風が強ければ、止め
　風が弱ければ、行う
天気予報が曇りであれば、
　行う
天気予報が晴れだれば
　そして湿度が75%より高ければ、止め
　湿度が75%以下であれば
　行う

コンタクトレンズの例

contact-lenses.arff
コンタクトレンズ.arff

年齢	眼鏡処方	乱視	涙産生	コンタクトレンズ
若年期	近視性	なし	少量	推奨せず
若年期	近視性	なし	正常	ソフト
若年期	近視性	あり	少量	推奨せず
若年期	近視性	あり	正常	ハード
若年期	遠視性	なし	少量	推奨せず
若年期	遠視性	なし	正常	ソフト
若年期	遠視性	あり	少量	推奨せず
若年期	遠視性	あり	正常	ハード
前老眼期	近視性	なし	少量	推奨せず
前老眼期	近視性	なし	正常	ソフト
前老眼期	近視性	あり	少量	推奨せず
前老眼期	近視性	あり	正常	ハード
前老眼期	遠視性	なし	少量	推奨せず
前老眼期	遠視性	なし	正常	ソフト
前老眼期	遠視性	あり	少量	推奨せず
前老眼期	遠視性	あり	正常	推奨せず
老眼期	近視性	なし	少量	推奨せず
老眼期	近視性	なし	正常	推奨せず
老眼期	近視性	あり	少量	推奨せず
老眼期	近視性	あり	正常	ハード
老眼期	遠視性	なし	少量	推奨せず
老眼期	遠視性	なし	正常	ソフト
老眼期	遠視性	あり	少量	推奨せず
老眼期	遠視性	あり	正常	推奨せず



分類問題

- 分類問題は、統計的には「判別問題」として扱われるが結構難しい。数多くの手法がある（Excel にはツールがない）
- 人工知能では古典的な課題である
- Fisher (統計学者)が扱った「あやめの分類問題」を考えてみる

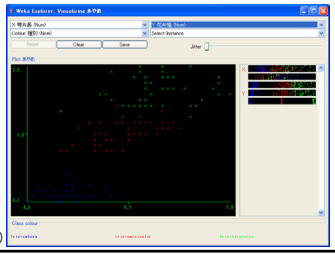
Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems.
Annals of Eugenics 7: 179-188.
(http://digital.library.adelaide.edu.au/coll/special/fisher/138.pdf)

あやめの分類問題

iris.arff
あやめ.arff

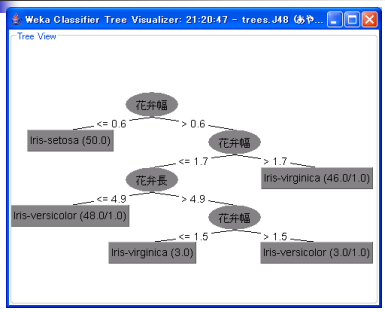
- 萼片長、萼片幅、花弁長、花弁幅とあやめ（setosa, versicolor, virginica の3種）の値が150組。

萼片長	萼片幅	花弁長	花弁幅	種別
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa



（横軸：萼片長、縦軸：花弁幅）

分類結果



労使間交渉の決着状況

labor.arff

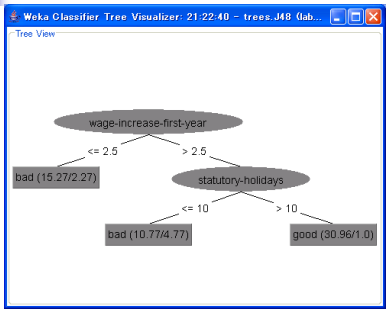
- カナダ労使間交渉の決着状況を、賃金・手当等との組みで表したもの
- 欠損値が多い（ごく普通の状況）：理論的・アルゴリズム的に困難な課題

労使間交渉データ

属性	型	1	2	3	...	40
継続期間	(年数)	1	2	3		2
賃上げ(第1年)	百分率	2	4	4.3		4.5
賃上げ(第2年)	百分率	?	5	4.4		4
賃上げ(第3年)	百分率	?	?	?		?
生活費保証	[none, tcf, tc]	none	tcf	?		none
労働時間/週	時間数	28	35	38		40
年金	[none, ret-allw, empl-cntr]	none	?	?		?
stand-by pay	百分率	?	13	?		?
変則勤務手当て	百分率	?	5	4		4
教育手当て	[あり、なし]	あり	?	?		?
土曜休業	休日数	11	15	12		12
休暇	[平均以下、平均、平均以上]	平均	平均以上	平均以上		平均
長期傷害助成	[あり、なし]	なし	?	?		あり
歯科診療保険助成	[なし、半分、完全]	なし	?	完全		完全
死別助成	[あり、なし]	なし	?	?		ari
健康保険助成	[なし、半分、完全]	なし	?	完全		半分
対応	[良い、悪い]	悪い	良い	良い		良い

（縦横がこれまでと逆なので注意）

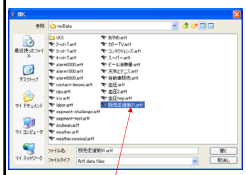
労使間交渉データの結果



判断値が数値のとき

- これまでは、if ... then ... の then のあとがカテゴリ変数（クラス、分類）であった
- 数値のときを、次に扱う
- 回帰と類似であるが、説明変数にカテゴリ変数があること、一次式（直線）で説明できない場合を扱うことが特徴

ファイルの選択

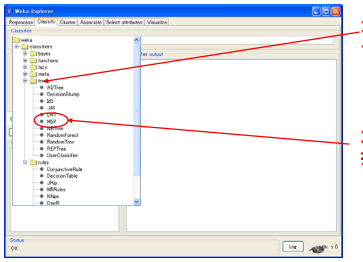


1. 販売促進01.arffファイル（どこかにある）クリック、

```
@relation '販売促進策01'
@attribute 月 real
@attribute 日 real
@attribute 曜日 {日, 月, 火, 水, 木, 金, 土}
@attribute 天候 {晴, 雨, 曇り}
@attribute 客数 real
@attribute 備考 {オートコール, 通常}
@data
7, 1, 金, 曇り, 491, 通常
7, 2, 土, 雨, 432, 通常
7, 3, 日, 晴, 514, 通常
7, 4, 月, 晴, 457, 通常
7, 5, 火, 曇り, 451, 通常
7, 6, 水, 雨, 441, 通常
7, 7, 木, 雨, 604, 通常
7, 8, 金, 曇り, 467, 通常
7, 9, 土, 晴, 408, 通常
7, 10, 日, 雨, 457, 通常
7, 11, 月, 雨, 484, 通常
7, 12, 火, 雨, 506, 通常
7, 13, 水, 曇り, 474, 通常
7, 14, 木, 晴, 666, 通常
7, 15, 金, 雨, 479, 通常
7, 16, 土, 曇り, 478, 通常
7, 17, 日, 晴, 640, 通常
7, 18, 月, 晴, 497, 通常
7, 19, 火, 晴, 473, 通常
7, 20, 水, 晴, 468, 通常
7, 21, 木, 晴, 875, オートコール
7, 22, 金, 晴, 829, オートコール
```

月	日	曜日	天候	客数	備考
7	1	金	曇り	491	通常
7	2	土	雨	432	通常
7	3	日	晴	514	通常
7	4	月	晴	457	通常
7	5	火	曇り	451	通常
7	6	水	雨	441	通常
7	7	木	雨	604	通常
7	8	金	曇り	467	通常
7	9	土	晴	408	通常
7	10	日	雨	457	通常
7	11	月	雨	484	通常
7	12	火	雨	506	通常
7	13	水	曇り	474	通常
7	14	木	晴	666	通常
7	15	金	雨	479	通常
7	16	土	曇り	478	通常
7	17	日	晴	640	通常
7	18	月	晴	497	通常
7	19	火	晴	473	通常
7	20	水	晴	468	通常
7	21	木	晴	875	オートコール
7	22	金	晴	829	オートコール
7	23	土	晴	597	通常
7	24	日	雨	633	通常
7	25	月	曇り	476	通常
7	26	火	晴	480	通常
7	27	水	晴	408	通常
7	28	木	晴	544	通常
7	29	金	雨	365	通常
7	30	土	晴	380	通常
7	31	日	晴	448	通常

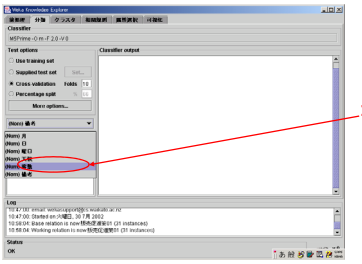
使うアルゴリズムの選択



1. Tree の右にある + をクリック

2. M5P というのを選択する

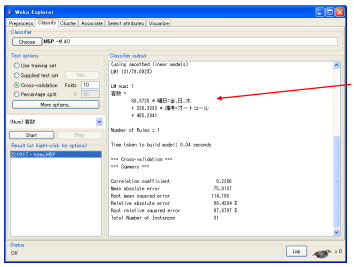
被説明変数の指定



1. 「客数」の上でクリック

黙っているとデータ（表）のなかの最も右の属性が用いられる。
今回は、「最も右」ではないのでここで指定する

結果の解析

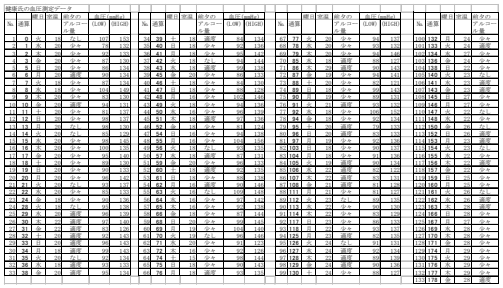


客数 =
60.3725 * 曜日=金,日,木
+ 326.3333 * 備考=オートコール
+ 465.2941

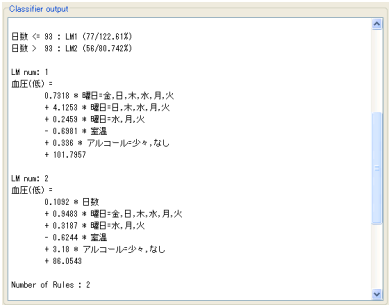
オートコールを行った方が客数が増加することがわかる

血圧の測定データ

血圧.arff

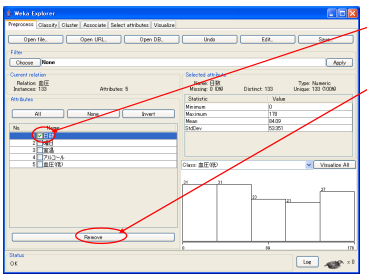


Weka による分析結果



```
Classifier output

日脈 <= 93 : LM1 (77/122.61%)
日脈 > 93 : LM2 (56/80.742%)

LM num: 1
血圧(低) =
        0.7818 * 曜日=金,日,木,水,月,火
        + 4.1253 * 曜日=日,木,水,月,火
        + 0.2459 * 曜日=水,月,火
        - 0.6981 * 室温
        + 0.338 * アルコール=少々,なし
        + 101.7957

LM num: 2
血圧(低) =
        0.1092 * 日脈
        + 0.8483 * 曜日=金,日,木,水,月,火
        + 0.3187 * 曜日=水,月,火
        - 0.8244 * 室温
        + 3.18 * アルコール=少々,なし
        + 86.0543

Number of Rules : 2
```

日数をはずす

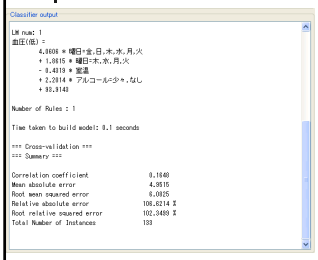


1. 日数のチェックボックスにチェック

2. 属性を remove するためクリック

3.「分類」で M5Prime を Start

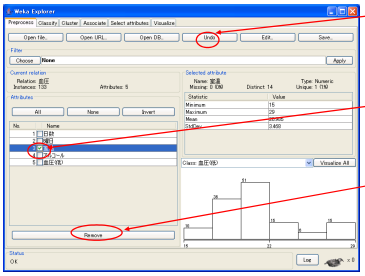
日数をはずした場合の結果



LM num: 1
血圧(低) =
4.0606 * 曜日=金,日,木,水,月,火
+ 1.8615 * 曜日=木,水,月,火
- 0.4319 * 室温
+ 2.2014 * アルコール=少々,なし
+ 93.9143

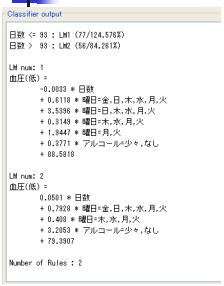
Correlation coefficient 0.1648

室温をはずす



1. Undo をクリックすると日数が戻ってくる
2. 室温にチェックをつける
3. Removeする

室温をはずした場合の結果



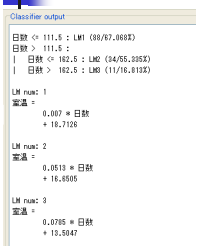
Correlation coefficient 0.2719

日数 <= 93 : LM1 (77/124.576%)
日数 > 93 : LM2 (56/84.261%)

LM1:血圧(低) =
-0.0033 * 日数
+ 0.6118 * 曜日=金,日,木,水,月,火
+ 3.5396 * 曜日=日,木,水,月,火
+ 0.3149 * 曜日=木,水,月,火
+ 1.9447 * 曜日=月,火
+ 0.3771 * アルコール=少々,なし
+ 88.5818

LM2:血圧(低) =
0.0501 * 日数
+ 0.7928 * 曜日=金,日,木,水,月,火
+ 0.408 * 曜日=木,水,月,火
+ 3.2053 * アルコール=少々,なし
+ 79.3907

日数と室温との関係



日数 <= 111.5 : LM1 (88/67.068%)
日数 > 111.5 :
| 日数 <= 162.5 : LM2 (34/55.335%)
| 日数 > 162.5 : LM3 (11/16.813%)

LM1 室温 = 0.007 * 日数 + 18.7126
LM2 室温 = 0.0513 * 日数 + 16.6505
LM3 室温 = 0.0785 * 日数 + 13.5047

Correlation coefficient 0.8465

日数と室温をはずすと

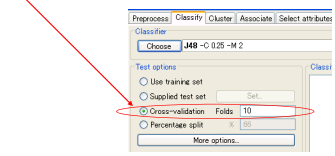
残りの属性(曜日と前日のアルコール摂取量)ではうまく説明できないことがわかる

「血圧」の総合的な結論

- 日数がたつにつれ、血圧が上昇している
- しかし、それは日数がたったからか、気温が上昇したからかはわからない
- 土曜日に低い傾向はあるが、確信できず
- 前日のアルコール摂取量で低い傾向はあるが、確信度はもっと低い

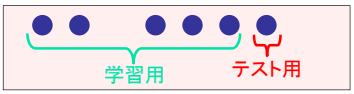
結果のテストの仕方

- 学習した結果はどの程度正しいのか、確認をする必要がある。
- Weka では標準的に 10-fold cross validation を行うようになっている。



k 重クロスバリデーション
k-fold cross validation

訓練データを k 群に分け、$(k-1)$ 群で学習し、残りで予測誤差を計測する。これを全ての k 種類の組み合わせに対して行なう



万能ではないが、多くの場合に結構うまくいく

予測誤差の計測値を、ここでは、汎化誤差と呼ぶことにする

テストデータによるテスト