

## Wekaの基礎

櫻井彰人  
慶應義塾大学理工学部

## Weka



- 今回使用するソフトウェア
- 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、そうはいつでも、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
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	気温	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

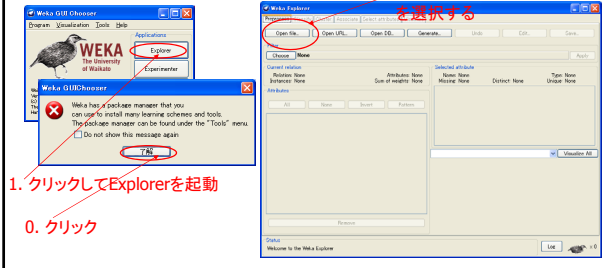
## Wekaバージョンに関する注意

	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 3.6.5	Windows	日本語化	文字化け	no	no
	others	日本語化	正しく表示	no	no
Weka 3.7.4	Windows	英語	文字化け	文字化け	正しく表示
	others	英語	正しく表示	正しく表示	正しく表示

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、同じ場所に、Windows版にある RunWeka.bat をコピーする。  
起動はこれをクリックする。

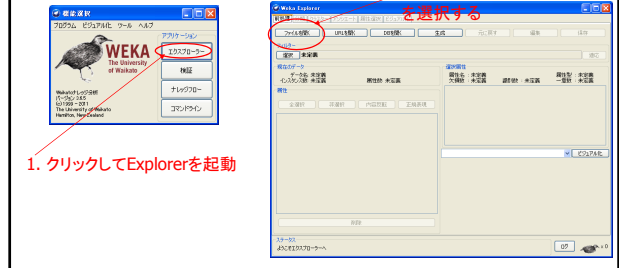
## 使ってみよう (Weka-3-7-4)

- 「すべてのプログラム」から起動



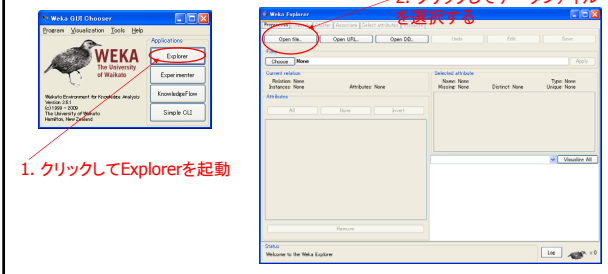
## 使ってみよう (Weka-3-6-5)

- 「すべてのプログラム」から起動



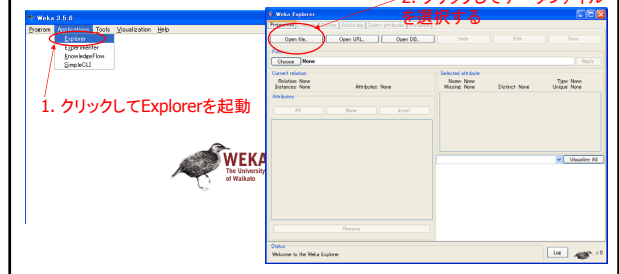
## 使ってみよう (Weka-3-6-1)

- 「すべてのプログラム」から起動



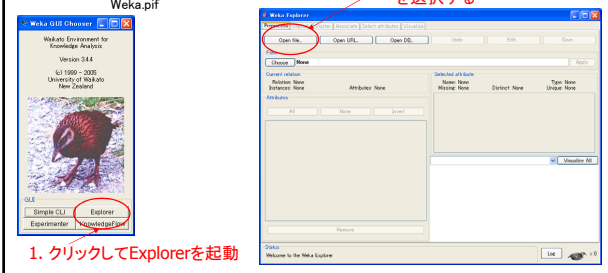
## 使ってみよう (Weka-3-5)

- 「すべてのプログラム」から起動



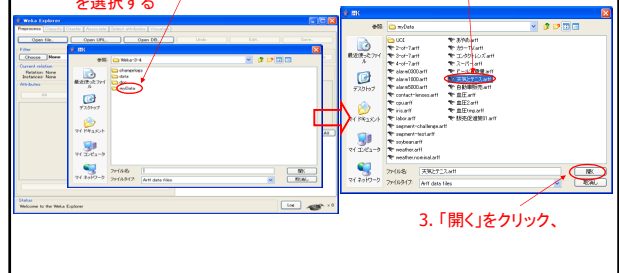
## 使ってみよう (Weka-3-4)

- Weka.pif をダブルクリック



## 対象データファイルの指定

- 1. クリックしてDataフォルダを選択する
- 2. クリックして天気とテニス.arffファイル(予めいれておく)を選択し、
- 3. 「開く」をクリック、



## 決定木の作成(計算)

1. Classify をクリック
2. Choose をクリック
3. Trees の + をクリック
4. j48 をクリック

## 結果の確認

1. 「Start」をクリック
2. 結果はこのウィンドウに表示される
3. このバーを上にもドラッグすると、最初のうが見れる
- 10重クロスバリデーションの結果の総和

CP Rate	FP Rate	Precision	Recall	F Measure	Class
4.376	4.3	8.7	8.788	9.287	10
4.4	4.202	8.4	8.488	8.9	11

## 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」をクリック
3. 「Visualize tree」の上でクリック

## 図示された木の変形

1. マウスマウスをこの角にもつくと、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」をクリック。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。文字の大きさを替えるには Select Font でクリック、木をドラッグすることもできる

## 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  
 天気予報が晴れであれば  
 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	近視	遠視	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	ハード
老眼期	近視性	なし	少量	推奨せず
老眼期	近視性	なし	正常	ソフト
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/col/special/fisher/138.pdf>)

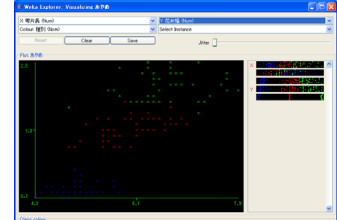
## あやめの分類問題



iris.arff  
あやめ.arff

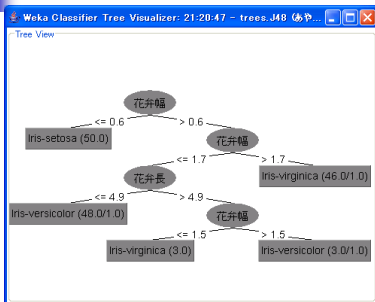
- 萼片長、萼片幅、花弁長、花弁幅とあやめ (setosa, versicolor, virginica の3種) の値が150組。

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
4.0	3.1	1.3	0.2	iris-setosa
5	3.8	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

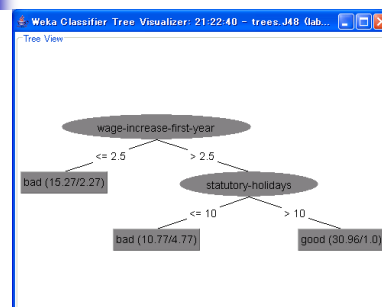
- カナダ労使間交渉の決着状況を、賃金・手当等との組みで表したのも
- 欠損値が多い(ごく普通の状況): 理論的・アルゴリズム的に困難な課題

## 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, tef, tc}	none	tef	?	...	none
労働時間/週	時間数	28	35	38	...	40
年金	{none, ret+allw, empl-cntrl}	none	?	?	...	?
stand-by pay	百分率	?	13	?	...	?
家則勤務手当	百分率	?	5	4	...	4
教育手当	{あり, なし}	あり	?	?	...	?
土曜休業	休日数	11	15	12	...	12
休職	{平均以下, 平均, 平均以上}	平均	平均以上	平均以上	...	平均
長期障害助成	{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	?	?	...	あり
健康保険助成	{なし, 半分, 完全}	なし	?	完全	...	半分
死亡	{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.affファイル(どこかにある)をクリック、

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	晴	457	通常
7/11	月	晴	484	通常
7/12	火	雨	505	通常
7/13	水	曇り	474	通常
7/14	木	雨	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	828	オートコール
7/23	土	晴	397	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.aff

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	晴	457	通常
7/11	月	晴	484	通常
7/12	火	雨	505	通常
7/13	水	曇り	474	通常
7/14	木	雨	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	828	オートコール
7/23	土	晴	397	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

## Weka による分析結果

```

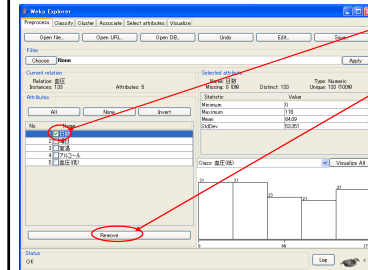
Classifier output
日数 < 93 : LM1 (77/124.576%)
日数 > 93 : LM2 (56/84.261%)

LM rule: 1
血圧(低) =
+ 0.7318 * 曜日=金,日,木,水,月,火
+ 4.1253 * 曜日=日,木,水,月,火
+ 0.2453 * 曜日=水,月,火
- 0.2591 * 室温
+ 0.335 * アルコール=少々,なし
+ 101.7857

LM rule: 2
血圧(低) =
0.1092 * 日数
+ 0.3483 * 曜日=金,日,木,水,月,火
+ 0.2107 * 曜日=水,月,火
- 0.6244 * 室温
+ 3.18 * アルコール=少々,なし
+ 88.0549

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

```

Classifier output
LM num: 1
血圧(低) =
+ 4.8958 * 曜日=金,日,木,水,月,火
- 0.4319 * 室温
+ 2.2014 * アルコール=少々,なし
+ 93.3141

Number of Rules : 1

Time taken to build model: 0.1 seconds

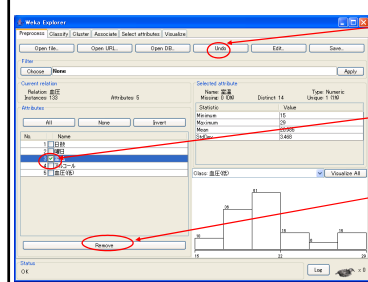
Cross-validation ==>
==== Summary ====

Correlation coefficient      0.1648
Mean absolute error        4.9515
Root mean squared error    6.0025
Relative absolute error    100.6214 %
Root relative squared error 102.3489 %
Total Number of Instances  133
    
```

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

```

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

LM rule: 1
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

LM rule: 2
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

Number of Rules : 2
    
```

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

Correlation coefficient 0.2719

## 日数と室温との関係

```

Classifier output
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%)

LM rule: 1
室温 =
0.007 * 日数
+ 18.7126

LM rule: 2
室温 =
0.0513 * 日数
+ 16.4505

LM rule: 3
室温 =
0.0785 * 日数
+ 13.5047
    
```

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 + 16.4505  
 LM3 室温 = 0.0785 \* 日数 + 13.5047

Correlation coefficient 0.8465

## 日数と室温をはずすと

```
Classifier output
LM rule: 1
血压(低) =
  3.2358 * 曜日(金,日,水,木,月,火)
  + 1.8755 * 曜日(水,木,月,火)
  + 2.4664 * アルコール摂取量(なし)
  + 85.4322

Number of Rules: 1
Time taken to build model: 0.00 seconds

*** Cross-validation ***
*** Summary ***

Correlation coefficient      -0.8888
Mean absolute error         4.261
Root mean squared error     6.1926
Relative absolute error     105.2657 %
Root relative squared error 104.2319 %
Total Number of Instances   100
```

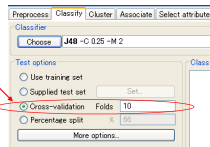
残りの属性(曜日と前日のアルコール摂取量)ではうまく説明できないことがわかる

## 「血压」の総合的な結論

- 日数がたつにつれ、血压が上昇している
- 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

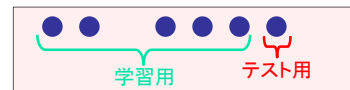
## 結果のテストの仕方

- 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、残りです測誤差を計測する。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

③ ファイル名の  
入力

② クリック

① 選択して  
クリックする

