

## 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が多いですが)
  - 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  
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
```

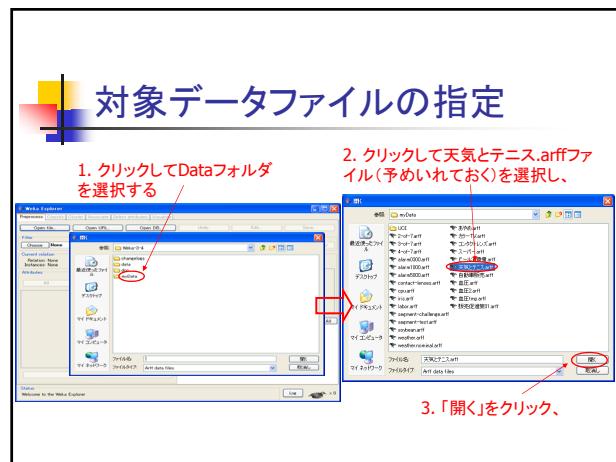
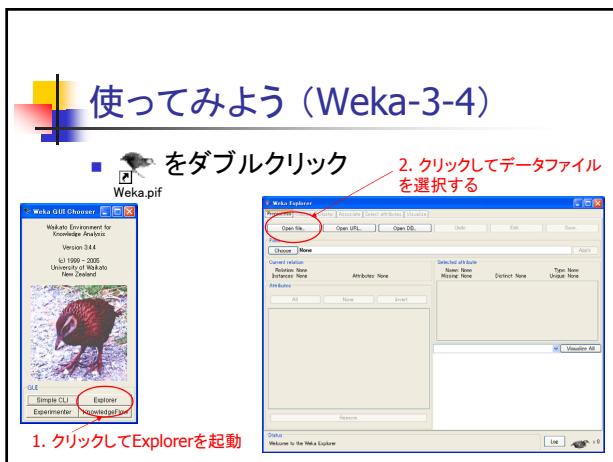
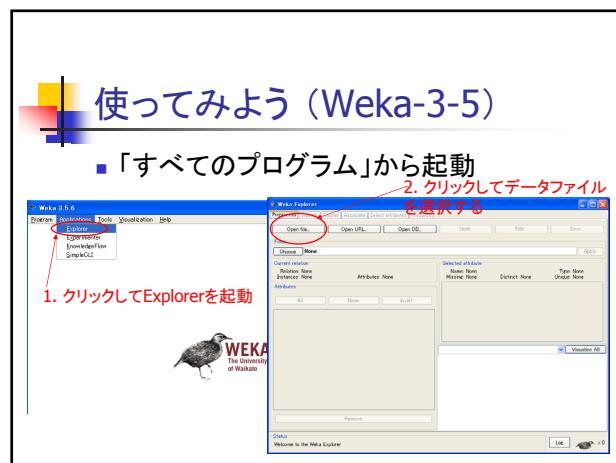
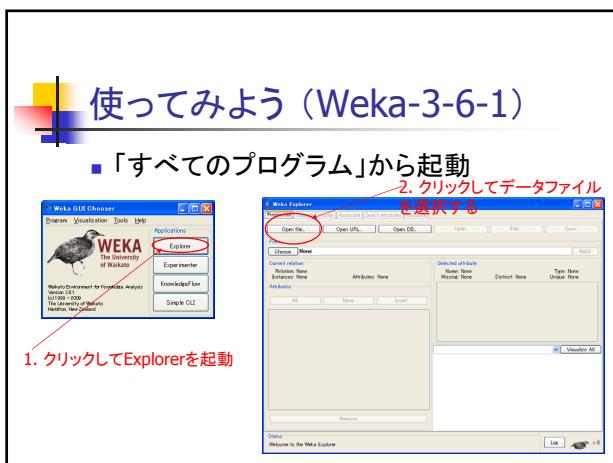
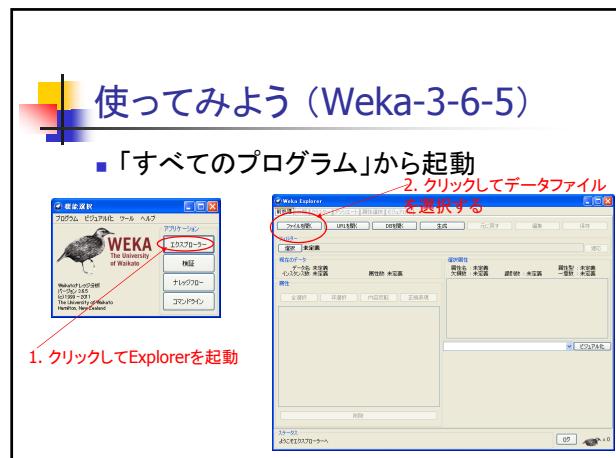
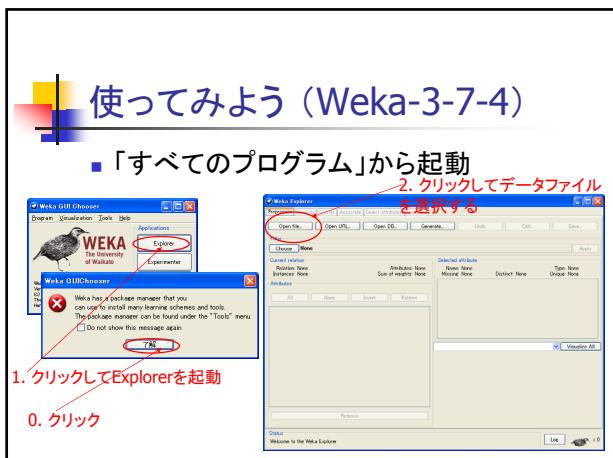
Excel の表形式で書いたもの

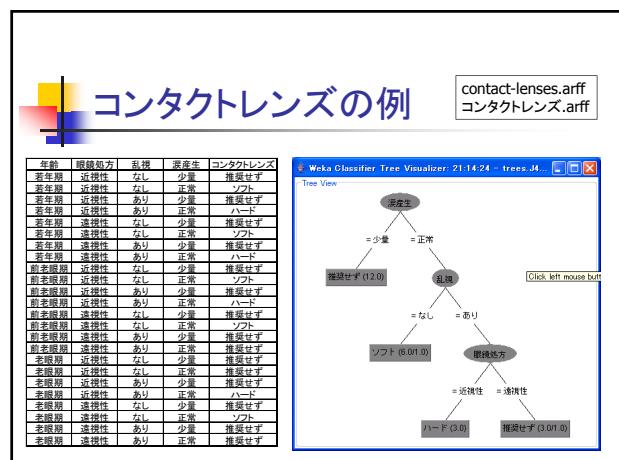
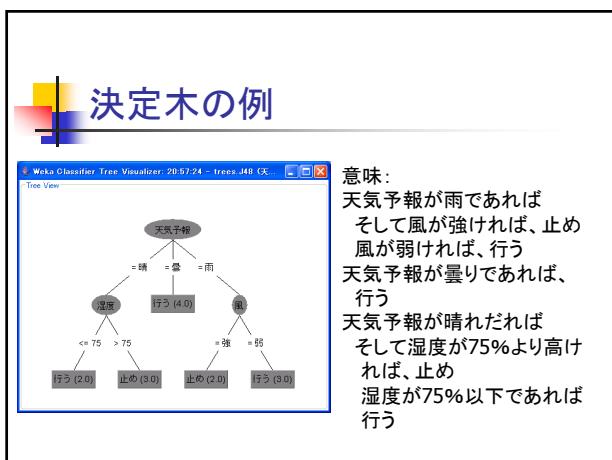
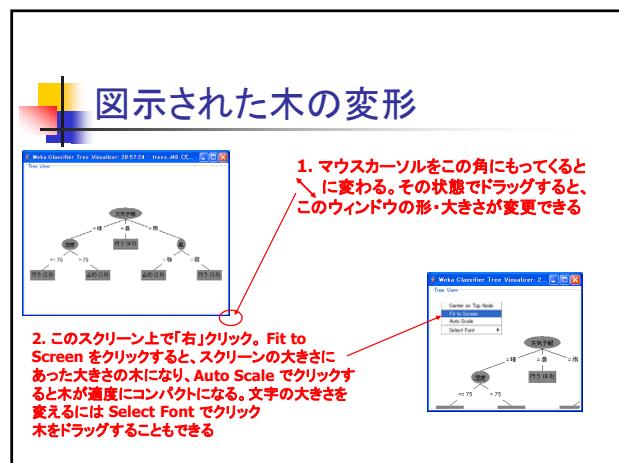
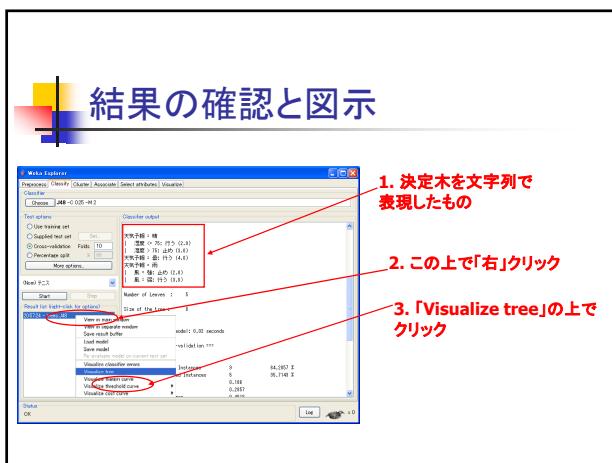
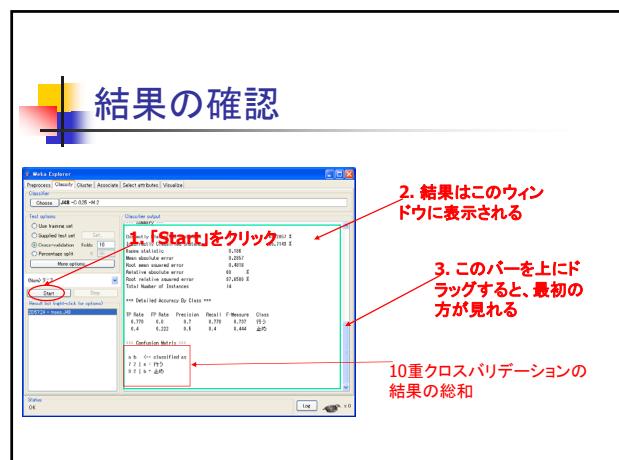
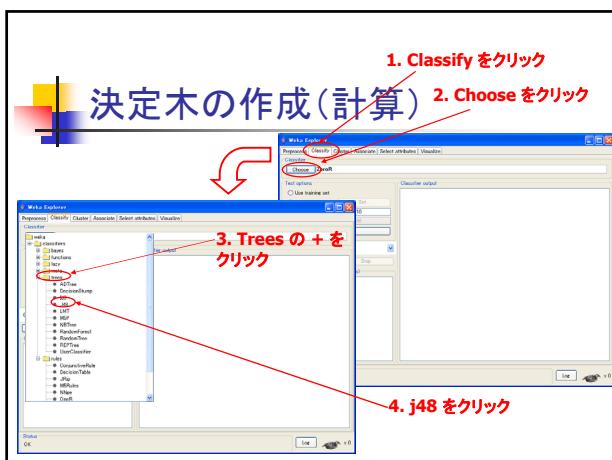
天気予報	温度	湿度	風	テニス
晴	29	85	弱	止め
晴	27	90	強	止め
曇	28	86	弱	行う
雨	21	96	弱	行う
雨	20	80	弱	行う
雨	18	70	強	止め
曇	18	65	強	行う
晴	22	95	弱	止め
晴	21	70	弱	行う
雨	24	80	弱	行う
晴	24	70	強	行う
曇	22	90	強	行う
曇	27	75	弱	行う
雨	22	81	強	止め

## 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 others	日本語化 日本語化	no no no
Weka 3.7.1	Windows others	英語 英語	文字化け 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。





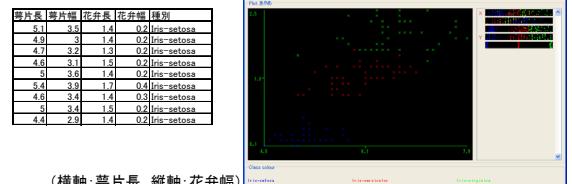
## 分類問題

- 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

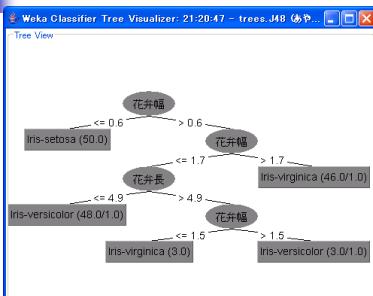
Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179-188. (<http://digital.library.adelaide.edu.au/coll/special/fisher/138.pdf>)

## あやめの分類問題

- 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。



## 分類結果



## 労使間交渉の決着状況

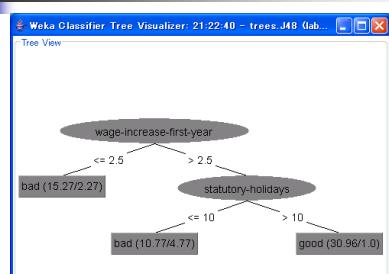
- カナダ労使間交渉の決着状況を、賃金・手当等との組みで表したもの
- 欠損値が多い(ごく普通の状況): 理論的・アルゴリズム的に困難な課題

## 労使間交渉データ

属性	例	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	?	?	?		?
生活費賃貸	(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?	?		?
家賃勤務手当	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
休暇	(半分以下, 平均, 平均以上)	平均	平均以上	平均以上		平均
長期傷害賠償	(あり,なし)	なし	?	?		あり
傷病休業賠償	(なし,半分, 完全)	なし	?	?		完全
死別助成	(なし,半分)	なし	?	?		半分
健康保険助成	(なし,半分, 完全)	なし	?	?		半分
勤労	(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

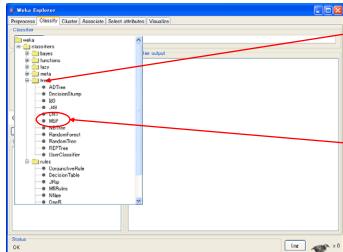
- これまででは、if ... then ... の then のあとがカテゴリ変数(クラス、分類)であった
- 数値のときを、次に扱う
- 回帰と類似であるが、説明変数にカテゴリ変数があること、一次式(直線)で説明できない場合を扱うことが特徴

## 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、

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	441	通常
7	6	水	雨	604	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	479	通常
7	11	月	雨	584	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	447	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	雨	544	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	雨	543	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



## 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

標準化された測定データ									
性別	年齢	身長	体重	心拍数	血圧	心拍数	血圧	心拍数	血圧
0	0	150	50	60	120/80	60	120/80	60	120/80
0	1	150	50	60	120/80	60	120/80	60	120/80
0	2	150	50	60	120/80	60	120/80	60	120/80
0	3	150	50	60	120/80	60	120/80	60	120/80
0	4	150	50	60	120/80	60	120/80	60	120/80
0	5	150	50	60	120/80	60	120/80	60	120/80
0	6	150	50	60	120/80	60	120/80	60	120/80
0	7	150	50	60	120/80	60	120/80	60	120/80
0	8	150	50	60	120/80	60	120/80	60	120/80
0	9	150	50	60	120/80	60	120/80	60	120/80
0	10	150	50	60	120/80	60	120/80	60	120/80
0	11	150	50	60	120/80	60	120/80	60	120/80
0	12	150	50	60	120/80	60	120/80	60	120/80
0	13	150	50	60	120/80	60	120/80	60	120/80
0	14	150	50	60	120/80	60	120/80	60	120/80
0	15	150	50	60	120/80	60	120/80	60	120/80
0	16	150	50	60	120/80	60	120/80	60	120/80
0	17	150	50	60	120/80	60	120/80	60	120/80
0	18	150	50	60	120/80	60	120/80	60	120/80
0	19	150	50	60	120/80	60	120/80	60	120/80
0	20	150	50	60	120/80	60	120/80	60	120/80
0	21	150	50	60	120/80	60	120/80	60	120/80
0	22	150	50	60	120/80	60	120/80	60	120/80
0	23	150	50	60	120/80	60	120/80	60	120/80
0	24	150	50	60	120/80	60	120/80	60	120/80
0	25	150	50	60	120/80	60	120/80	60	120/80
0	26	150	50	60	120/80	60	120/80	60	120/80
0	27	150	50	60	120/80	60	120/80	60	120/80
0	28	150	50	60	120/80	60	120/80	60	120/80
0	29	150	50	60	120/80	60	120/80	60	120/80
0	30	150	50	60	120/80	60	120/80	60	120/80
0	31	150	50	60	120/80	60	120/80	60	120/80
0	32	150	50	60	120/80	60	120/80	60	120/80
0	33	150	50	60	120/80	60	120/80	60	120/80
0	34	150	50	60	120/80	60	120/80	60	120/80
0	35	150	50	60	120/80	60	120/80	60	120/80
0	36	150	50	60	120/80	60	120/80	60	120/80
0	37	150	50	60	120/80	60	120/80	60	120/80
0	38	150	50	60	120/80	60	120/80	60	120/80
0	39	150	50	60	120/80	60	120/80	60	120/80
0	40	150	50	60	120/80	60	120/80	60	120/80
0	41	150	50	60	120/80	60	120/80	60	120/80
0	42	150	50	60	120/80	60	120/80	60	120/80
0	43	150	50	60	120/80	60	120/80	60	120/80
0	44	150	50	60	120/80	60	120/80	60	120/80
0	45	150	50	60	120/80	60	120/80	60	120/80
0	46	150	50	60	120/80	60	120/80	60	120/80
0	47	150	50	60	120/80	60	120/80	60	120/80
0	48	150	50	60	120/80	60	120/80	60	120/80
0	49	150	50	60	120/80	60	120/80	60	120/80
0	50	150	50	60	120/80	60	120/80	60	120/80
0	51	150	50	60	120/80	60	120/80	60	120/80
0	52	150	50	60	120/80	60	120/80	60	120/80
0	53	150	50	60	120/80	60	120/80	60	120/80
0	54	150	50	60	120/80	60	120/80	60	120/80
0	55	150	50	60	120/80	60	120/80	60	120/80
0	56	150	50	60	120/80	60	120/80	60	120/80
0	57	150	50	60	120/80	60	120/80	60	120/80
0	58	150	50	60	120/80	60	120/80	60	120/80
0	59	150	50	60	120/80	60	120/80	60	120/80
0	60	150	50	60	120/80	60	120/80	60	120/80
0	61	150	50	60	120/80	60	120/80	60	120/80
0	62	150	50	60	120/80	60	120/80	60	120/80
0	63	150	50	60	120/80	60	120/80	60	120/80
0	64	150	50	60	120/80	60	120/80	60	120/80
0	65	150	50	60	120/80	60	120/80	60	120/80
0	66	150	50	60	120/80	60	120/80	60	120/80
0	67	150	50	60	120/80	60	120/80	60	120/80
0	68	150	50	60	120/80	60	120/80	60	120/80
0	69	150	50	60	120/80	60	120/80	60	120/80
0	70	150	50	60	120/80	60	120/80	60	120/80
0	71	150	50	60	120/80	60	120/80	60	120/80
0	72	150	50	60	120/80	60	120/80	60	120/80
0	73	150	50	60	120/80	60	120/80	60	120/80
0	74	150	50	60	120/80	60	120/80	60	120/80
0	75	150	50	60	120/80	60	120/80	60	120/80
0	76	150	50	60	120/80	60	120/80	60	120/80
0	77	150	50	60	120/80	60	120/80	60	120/80
0	78	150	50	60	120/80	60	120/80	60	120/80
0	79	150	50	60	120/80	60	120/80	60	120/80
0	80	150	50	60	120/80	60	120/80	60	120/80
0	81	150	50	60	120/80	60	120/80	60	120/80
0	82	150	50	60	120/80	60	120/80	60	120/80
0	83	150	50	60	120/80	60	120/80	60	120/80
0	84	150	50	60	120/80	60	120/80	60	120/80
0	85	150	50	60	120/80	60	120/80	60	120/80
0	86	150	50	60	120/80	60	120/80	60	120/80
0	87	150	50	60	120/80	60	120/80	60	120/80
0	88	150	50	60	120/80	60	120/80	60	120/80
0	89	150	50	60	120/80	60	120/80	60	120/80
0	90	150	50	60	120/80	60	120/80	60	120/80
0	91	150	50	60	120/80	60	120/80	60	120/80
0	92	150	50	60	120/80	60	120/80	60	120/80
0	93	150	50	60	120/80	60	120/80	60	120/80
0	94	150	50	60	120/80	60	120/80	60	120/80
0	95	150	50	60	120/80	60	120/80	60	120/80
0	96	150	50	60	120/80	60	120/80	60	120/80
0	97	150	50	60	120/80	60	1		

## Weka による分析結果

Classifier output

```

日数 < 93 : LM1 (77/122, 81%)
日数 > 93 : LM2 (56/84, 74%)

LM num: 1
血圧(低) =
0.0039 * 曜日=金,日,木,水,月,火
+ 0.1198 * 曜日=木,水,月,火
+ 0.3354 * 曜日=木,水,月,火
+ 0.3549 * 曜日=木,水,月,火
- 0.8981 * 室温
+ 0.398 * アルコール=少々,なし
+ 101.7957

LM num: 2
血圧(低) =
0.1092 * 日数
+ 0.5483 * 曜日=金,日,木,水,月,火
+ 0.3187 * 曜日=水,月,火
- 0.8244 * 室温
+ 3.1018 * アルコール=少々,なし
+ 88.0543

Number of Rules : 2

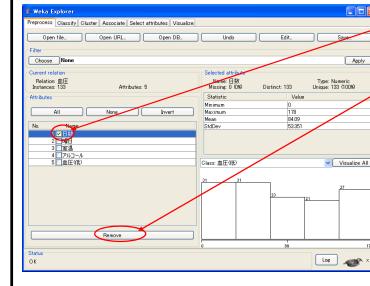
```

## 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

Classifier output

```

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

Number of Rules : 1

Time taken to build model: 0.1 seconds

```

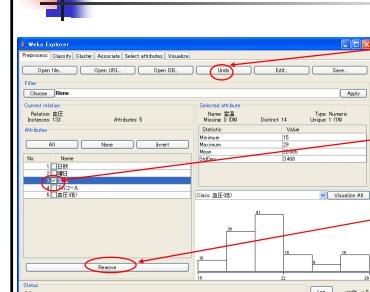
Correlation coefficient 0.1648

## 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, 57.6%)
日数 > 93 : LM2 (56/84, 26.1%)

LM num: 1
血圧(低) =
-0.0039 * 日数
+ 0.1198 * 曜日=金,日,木,水,月,火
+ 0.3354 * 曜日=木,水,月,火
+ 0.3549 * 曜日=木,水,月,火
+ 0.8981 * 曜日=木,水,月,火
+ 0.398 * アルコール=少々,なし
+ 88.5818

LM num: 2
血圧(低) =
0.0051 * 日数
+ 0.1203 * 曜日=金,日,木,水,月,火
+ 0.4008 * 曜日=木,水,月,火
+ 3.2053 * アルコール=少々,なし
+ 78.3987

Number of Rules : 2

```

Correlation coefficient 0.2719

## 日数と室温との関係

日数 <= 111.5 : LM1 (88/67, 0.68%)
日数 > 111.5 :
| 日数 <= 162.5 : LM2 (34/55, 33.5%)
| 日数 > 162.5 : LM3 (11/16, 81.3%)

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

```
Classifier output
LIM rule 1
血压(底) =
3,2589 + 昨日-1日-木-水-月-火
+ 2,4464 * 駆日-木-水-月-火
+ 85,4322

Number of Rules : 1
Time taken to build model: 0.08 seconds
*** Cross-validation ***
*** Summary ***
Correlation coefficient -0.00888
Mean absolute error 4.931
Root mean squared error 4.489
Relative absolute error 105.8637 %
Root relative squared error 104.2919 %
Total Number of Instances 138
```

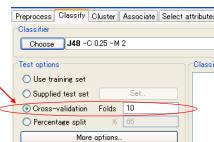
残りの属性(曜日と前日のアルコール摂取量)ではうまく説明できないことがある

## 「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

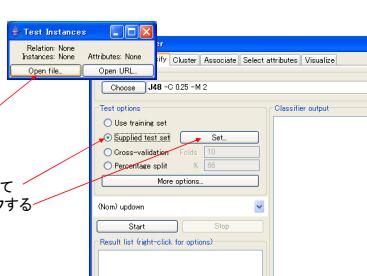
訓練データを  $k$  群に分け、 $(k-1)$  群で学習し、残りで予測誤差を計測する。これを全ての  $k$  種類の組み合わせに対して行なう



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

- 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