COM実験（2008/10/9）

櫻井彰人

実験の目的

- データの（客観的、自動的）分析を試みる、その基礎を知る

レポートについて1

- レポートに報告して戴きたいことは次の点です。
- スライド中、実験1及び実験2の「実験手順」となっている部分に書かれている手順に従い、実験を行い、その結果（自分が行ったこと、得られた結果）とそれに関する考察
  - 実験1「文字認識」
  - 実験2「歌詞の分類」
  - 実験3「ドル円レートの予測」は、ご参考
- 感想

レポートについて2

- 言うまでもないことですが、他人のレポート・著作物等を写してはいけません。自分の独力で作成してください。
- 他人の著作物からの引用には、出典を明記してください。
- 締め切りは、10月29日です。

では、本論に

データには裏がある

- 申し訳ない、ちょっと表現が悪い。
- 正しくは、「データには構造がある」と言うべき。
- データというのは、数字の並び、または、文字の並び。一列に並んだもの。だから構造はない。
- そして、紙の上に書かれたり、コンピュータのメモリに入れられているだけ
- しかし、それが何か実世界のもの（自然物、人工物）によって生成されたものなら、それらの間に何らかの関係があろう。

データの裏を知りたい、知ろう

- となれば、データの背後関係を知りたくなろう。
- 単なる興味の問題ではなく、世の中を生きていく上で（少々大げさか）必要なこと

- どうしてそのデータは出てきたのか？
- データにはどういう関係があるのか？
- （そのデータに関して）将来どうなるのか？

例えば

- 最近、「app store」という言葉を耳にしたのだが、これは今後流行るのだろうか？

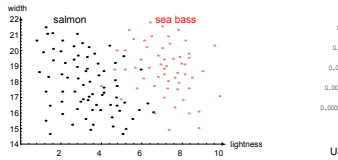
- いやいや、これほど難しい問いに答えられるわけはない。が、例えば、

- 最近、体重の変化がこんな（日々の体重がずら～っと並んだ表）具合だが、おかしくないか？病院に行くべきか？
- なんて気が利かないんだ、このブラウザは？私の過去の行動を見ているんだから、次やりたいことぐらい、察してくれよ。
- 写真の数が増えすぎて困った。見たい写真をいくつか教えるから、それに似た写真を集めて、時間順に並べてくれると助かるんだが。

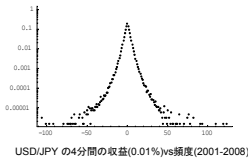
ここで、データとは

- 数字の列
- 文字の列（言葉というべき）。言語
- 写真の集まり
- 絵画の集まり
- 行動の表現（って何だろう？）の集まり
- 音の表現（って何だろう？）の集まり

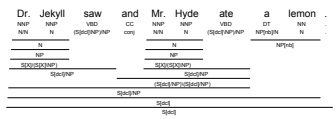
構造とは



R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*



USD/JPY の4分間の収益(0.01%)vs頻度(2001-2008)



J. Curran and S. Clark. C&C tools.

その構造1

- 数値データ（最も無味乾燥）
  - 正常と考えられる個体群、異常と考えられる個体群からデータが得られる。
  - 時間とともに変化する（時系列）。その変化の原因。
- 文字データ（少し構造があるように見える）
  - どのような文字を使うかという癖（人により違う）
  - どのような単語を使うかという癖（人により違う）
- 画像（写真、絵画等）
  - 対象（意図、目的に依存）
  - 構図、描き方、撮り方（作成者、状況等に依存）

http://cert.yahoo.co.jp/text/digicame/chap2/c2_0302.html

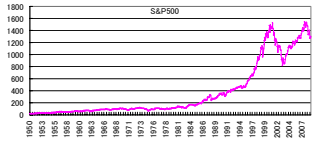
その構造2

- 音（音楽、声、虫の音、、、）
  - 道具（楽器、性別・年齢・健康、虫の種類）
  - 音楽の場合に、楽曲の種類、演奏等

- 行動
  - （買物）どういう目的。誰のために。何のために。どこで。
  - （web閲覧）目的。何（ネットワーク、PC等）を使って。

構造が分かれば

- 次の行動がとれる
  - 健康データ: 病気だと分かれば、医者に行く
  - 時系列: 次の予測
  - 文学作品: 気に入った作者が分かれば、次の作品を読む
  - web の流行: それに乗る(乗らないという選択もあり)
  - 一般に、分類し、それにあわせた行動をとる。



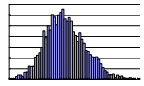
どうしたらできるか？

- コンピュータサイエンスでは、昔から、「機械学習」という分野で深く研究されてきた
  - 人工知能の一分野と考えてよい
  - なぜ「学習」か？
    - 人間の学習は、丸暗記を除くと、理解して、記憶して、次に使えるようにすること
    - この「理解する」ということがポイント
    - データの構造を知ることは、理解することの重要な一部
  - なぜ「機械」か？
    - コンピュータのこと。英語では、計算機械と言った。
  - なぜ「ロボットでない」のか？
    - ロボットにも機械学習は必要。しかし、ロボットに限らない。
  - 人間との違いは？
    - 人間ほど賢くはない。コンピュータは「実世界」を知らない。
    - しかし、大量のデータを、文句も言わず、処理してくれる

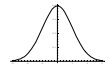
どうしたらできる？ その2

- 非常にたくさんのアルゴリズムが開発されてきた
- たくさんありすぎて、説明できない(私も全部知っているわけでは、当然、ない)。
  - というわけで、説明しません。
- 実験では、道具を使って、体感することで我慢してください。
  - 申し訳なし
- もっと知りたい場合には、是非、私の講義を聴いてください。

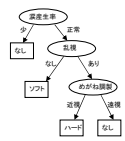
統計との違い



http://wwwsteep.hc.edu/TIMSSI/database.html より (calculus)



- 昔(20年前)はまったく別のdiscipline
- しかし、今は融合が進んでいる。
- しかし、違いはある
- 統計:
  - 機械学習に比べると簡単な(確率的)構造を考える。そのかわり、検定することが頭にある。主に数値データ
- 機械学習:
  - どんなモデルでも考える。検定したいが、できないことが結構ある。
    記号データも平気で扱う。



データマイニングとの違い

- 実は、ない。
- 研究者もかなり重なっている

- あえて言えば:
  - 機械学習:アルゴリズムに重点
    - 精度、速さ、軽さ、応用の広さ、表現力
  - データマイニング:大量のデータに重点
    - 知識、アクション、データ個別的でもOK

話しが抽象的になっているので、

- 簡単な実例を一つ二つ。

データの例

Make	Model	Year	Head inj. c.	Chest decel.	L. Leg	R. Leg	D/P	Protection
Acura	Integra	87	599	35	791	262	Driver	manual belts
Acura	Integra RS	90	585		1545	1301	Driver	Motorized belts
Acura	Legend LS	88	435	50	926	708	Driver	d airbag
Audi	80	89	600	49	168	1871	Driver	manual belts
Audi	100	89	185	35	998	894	Driver	d airbag
BMW	325i	90	1036	56	865		Driver	d airbag
Buick	Century	91	815	47	1340	315	Driver	passive belts
Buick	Elect. Park	88	1467	54				
Buick	Le Sabre	90		35				
Buick	Regal	88	880	58				
Cadillac								

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
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa

通算	曜日	室温	前夕のアルコール量	血圧(mmHg) (LOW)	(HIGH)
0	火	18	なし	107	153
1	水	20	少々	79	120
2	木	20	少々	92	133
3	金	20	少々	87	130
4	日	20	少々	86	134
5	月	20	適度	90	134
6	火	18	少々	87	134
7	水	18	少々	104	149
8	木	20	少々	83	130
9	金	20	適度	94	141

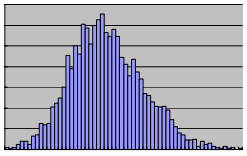
Date	Open	High	Low	Close	Volume	Adj. Cl	YearWeek
27/11/2000	53.6875	54.5156	51.0312	51.25	40198100	51.250	200049
28/11/2000	51.9375	53.1875	50.625	51	52037000	51.000	200049
29/11/2000	51.3125	53	50.3125	51.6875	55316000	51.688	200049
30/11/2000	50.1875	50.9375	45.1875	47.875	10840500	47.875	200048
01/12/2000	49.1875	51.625	47.25	48.5	70468000	48.500	200049
04/12/2000	49.0625	49.5625	45	45.8125	9501200	45.813	200050
05/12/2000	47.75	52.125	47.3125	52.125	90848900	52.125	200050
06/12/2000	52	53.5625	51.2656	51.4375	71419200	51.438	200050
07/12/2000	50.3125	51	49	49.9375	46448400	49.938	200050
08/12/2000	51.9375	53.25	51	52.375	55400200	52.375	200050
11/12/2000	52.875	55.75	52.625	54.8125	78621500	54.813	200051
12/12/2000	54.75	55.125	53.3125	54.375	39485300	54.375	200051
13/12/2000	55.1875	55.25	50.8125	51.125	54330600	51.125	200051
14/12/2000	51.0625	52.5625	50.875	50.9375	46244400	50.938	200051
15/12/2000	50.0625	50.1875	47.125	48.1719	100237900	48.172	200051
18/12/2000	49	50.125	42.3125	42.9375	126032400	42.938	200052
19/12/2000	43	46	41.5	41.75	89018800	41.750	200052

統計的な表現方法

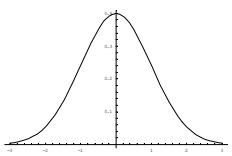
- 一つ一つのデータには規則性はないのに、集団でみると(出現頻度等に)規則性が見られる。
  - 一様分布
    - 公平なサイコロ、コイン投げ
  - 正規分布
    - たくさんの独立な要因がからむとき
  - Zipfの法則、80/20の法則、1/f ゆらぎ

正規分布

- 多数の独立な要因の結果発生する
  - コイン100回投げの表の回数の1000回実験
  - 例:(本当は違うのだが)試験の成績



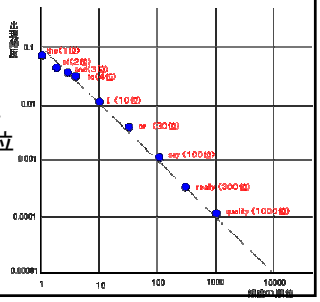
http://wwwcsteep.bc.edu/TIMSS1/database.html より (calculus)



Plot[(1/Sqrt[2 Pi]) Exp[-x^2/2], {x, -3, 3}]

Zipfの法則

- 「単語の出現頻度がその順位に反比例する」という経験則
  - 都市の人口とその順位
  - HPのヒット数とその順位
  - 資産とその順位



機械学習での規則の表現

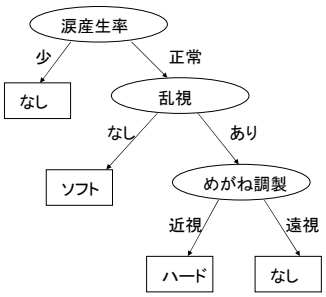
- 条件文
  - IF むにゃむにゃ THEN あれこれ
  - 信頼性は ?? %
- 決定木
  - 次のスライド
- ニューラルネットワーク
- ほかにもいろいろ

If-then 規則

If 涙産生 = 少 then 推薦 = しない
If 年齢 = 若い and 乱視 = なし and 涙産生 = 通常 then 推薦 = ソフト
If 年齢 = 老眼以前 and 乱視 = なし and 涙産生 = 通常 then 推薦 = ソフト
If 年齢 = 老眼 and めがね = 近視 and 乱視 = なし then 推薦 = なし
If めがね = 遠視 and 乱視 = なし and 涙産生 = 通常 then 推薦 = ソフト
If めがね = 近視 and 乱視 = あり and 涙産生 = 通常 then 推薦 = ハード
If 年齢 = 若い and 乱視 = あり and 涙産生 = 通常 then 推薦 = ハード
If 年齢 = 老眼以前 and めがね = 遠視 and 乱視 = あり then 推薦 = しない
If 年齢 = 老眼 and めがね = 遠視 and 乱視 = あり then 推薦 = しない

年齢	めがね	乱視	涙産生	推薦
若い	近視	なし	少	しない
若い	近視	なし	通常	ソフト
若い	近視	あり	少	しない
若い	近視	あり	通常	ハード
若い	遠視	なし	少	しない
若い	遠視	なし	通常	ソフト
若い	遠視	あり	少	しない
若い	遠視	あり	通常	ハード
老眼以前	近視	なし	少	しない
老眼以前	近視	なし	通常	ソフト
老眼以前	近視	あり	少	しない
老眼以前	近視	あり	通常	ハード
老眼以前	遠視	なし	少	しない
老眼以前	遠視	なし	通常	ソフト
老眼以前	遠視	あり	少	しない

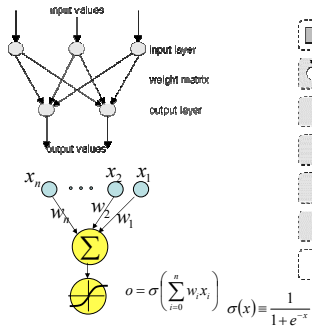
決定木

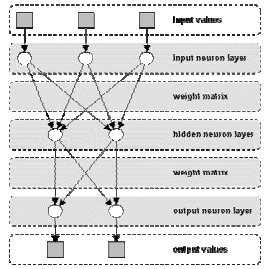


年齢	めがね	乱視	涙産生	推奨
若い	近視	なし	少	しない
若い	近視	なし	通常	ソフト
若い	近視	あり	少	しない
若い	近視	あり	通常	ハード
若い	遠視	なし	少	しない
若い	遠視	なし	通常	ソフト
若い	遠視	あり	少	しない
若い	遠視	あり	通常	ハード
老眼以前	近視	なし	少	しない
老眼以前	近視	なし	通常	ソフト
老眼以前	近視	あり	少	しない
老眼以前	近視	あり	通常	ハード
老眼以前	遠視	なし	少	しない
老眼以前	遠視	なし	通常	ソフト
老眼以前	遠視	あり	少	しない
老眼以前	遠視	あり	通常	しない
老眼	近視	なし	少	しない
老眼	近視	なし	通常	しない
老眼	近視	あり	少	しない
老眼	近視	あり	通常	ハード
老眼	遠視	なし	少	しない
老眼	遠視	なし	通常	ソフト
老眼	遠視	あり	少	しない
老眼	遠視	あり	通常	しない

ニューラルネットワーク

大量の関数合成です





機械学習での結果表現

- データマイニングならなお更
- 結果を表現する必要が(たいていは)ある。
  - 予測だけを要求されれば、その限りではない。
  - 一般に結果だけを要求されれば、その限りではない。
  - しかし、「説明責任」があるといわれるのは、世の常
  - そこで、「分かりやすく」表現する必要があることがある。
- しかし、(予測等の)「精度」と「分かりやすさ」は trade-offの関係にある
- 機械学習では両方用意している。例えば、
  - 決定木:分かりやすさ
  - SVM(support vector machine)やNN: 精度や柔軟性

まとめ

- 対象
  - 数値、記号で表現できるもの
    - e.g. 言葉、画像、音、数値、
- 目的と考え方
  - 予測、推測; 変化に対応する知識システム
    - 人間が明確には記述できない対象(予測・推測規則)
  - ノイズと対象に分離し、対象を「記述する」
    - ノイズ: ランダム、無相関
    - 対象: 構造がある
- 道具・手段
  - 統計、人工知能、機械学習!!

実際面での注意

- ツールは山ほどありますが、実用的にするには、「特徴量」をうまく選ぶことが大切。
- 「特徴量」とは、データから計算されるある量で、機械学習に用いる
  - ダメな特徴量:
    - 風邪か否かを診断するのに、身長を使う
    - ある日の株価予測: 雲の量を用いる(尤も晴雨が影響するという説あり)
- 技法よりは特徴量
  - どんな特徴量をとるかで勝負はかなり決まる
  - 特徴量: 対象から計算される値
    - 一般には、複数個組み合わせると、対象が属するグループが決まるとよい。

実験

- 機械学習ツールとして Weka を用いる
  - 別のスライドへ
- 実験1: 文字識別
  - ためしに、数字。
- 実験2: 歌詞の50音分布から、歌のジャンルの推定
  - 自分でデータが作れる。
- 実験3: ドル円レートの予測

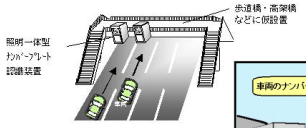
実験1：文字認識

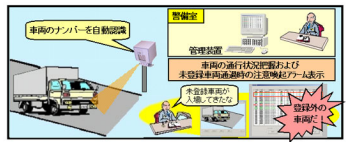
- 実世界で活躍中
  - 高速道路の料金所でナンバープレートを読む
    - Nシステムもそうらしい

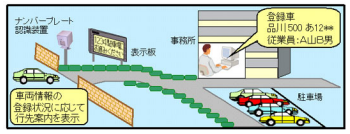
Wikipediaより

  - 郵便番号自動読み取り(「区分け」と「配達順序」が重要)。手書きの住所も読む。高速。
  - 蛇足ですが、「正解率99%　ネット認証技術、書籍のデジタル化に威力」という記事を見て下さい。

蛇足







日立情報制御ソリューションズ

文字認識は結構難しい

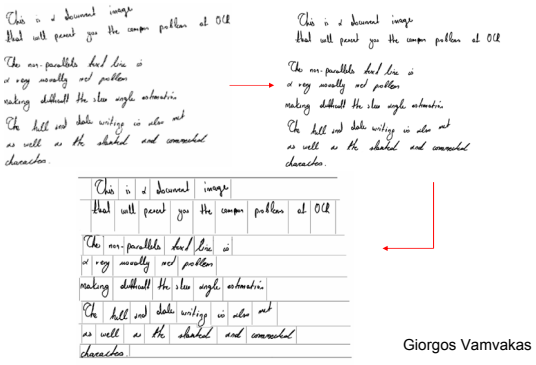
- 人間なら、崩し字でなければ、簡単だと思う
- けれども、平仮名の読み方を、日本語を勉強したことのない人に教えることを想像してみてください。

- 何が難しいか？
- 規則が書けない！（実は分からないのかも、実は「ない」のかも）

実験データの説明

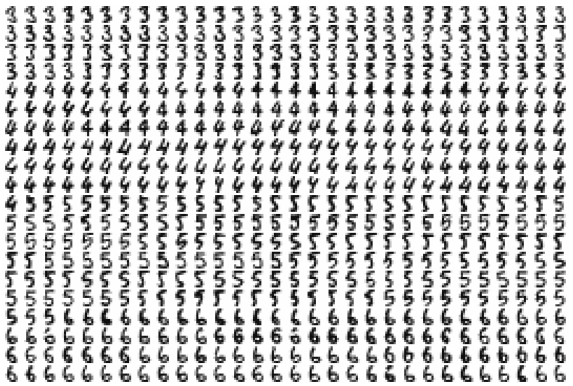
- 簡単な文字認識
  - 数字のみ。
- データの前処理（これが大変）済み
  - 分離（他の文字から分離）済み
  - 整形（大きさ、傾き、重心等）済み
- それでも、結構、難しそう。
  - 「数字」を知らない（！）人に区別の仕方を説明してみよう
- 本質「分類規則が表現できない」
  - データから得る
- データのもと:
  UCI Machine Learning Repository
  Optical Recognition of Handwritten Digits Data Set

データの前処理



Giorgos Vamvakas

データを図にして並べてみました



UCI Machine Learning Repository

データ形式は単純

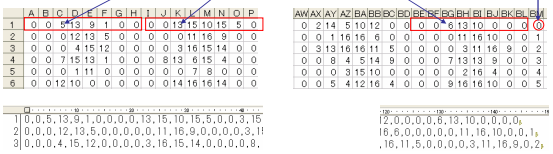
optdigits.tes.csv を開け、自分の目で確認してください

8 ピクセル

各ピクセル値は
0（白）… 16（黒）

8 ピクセル

文字
0 … 9



データ形式は単純

4. Relevant Information:
   We used preprocessing programs made available by NIST to extract normalized bitmaps of handwritten digits from a preprinted form. From a total of 43 people, 30 contributed to the training set and different 13 to the test set. 32x32 bitmaps are divided into nonoverlapping blocks of 4x4 and the number of on pixels are counted in each block. This generates an input **matrix of 8x8** where each element is an integer in the range **0..16**. This reduces dimensionality and gives invariance to small distortions.

5. Number of Instances
   optdigits.tra Training 3823
   optdigits.tes Testing 1797

6. Number of Attributes
   64 input+1 class attribute

optdigits.names.txt を開けて自分の目で確認してください

作業手順

optdigits.tra.csv

- データを入手（講義のサイトにあり）
  - ピクセル値を特徴量として用いる（最悪だけどね）
- arffファイルを作る（csvになっているので、ヘッダーをつけるだけ）。メモ帳で開こう。間違えないように。

```
@relation OptDigitsTraining
@attribute 00 real
@attribute 01 real
      .....
@attribute 06 real
@attribute 07 real
@attribute 10 real
@attribute 11 real
@attribute 12 real
      ......
      .....
@attribute 76 real
@attribute 77 real
@attribute class {0,1,2,3,4,5,6,7,8,9}
@data
以下はデータ
```

自由に

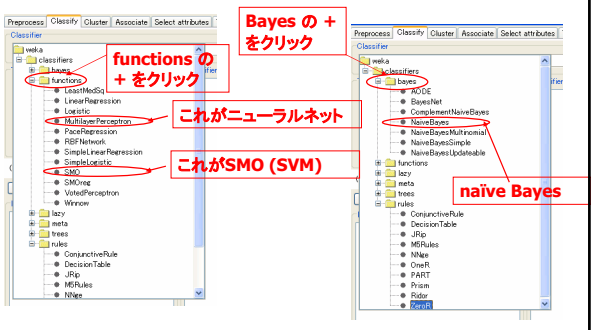
自由に。ただし、8 x 8 =64 個

実験手順：学習と検証

- 決定木を作る
  - Weka: Trees にある J48
- 10-fold cross validation での正解率を求める
- テストデータとして、optdigits.tes を用いたときの正解率を求める
- 決定木を見てみる。何か意味が見つかるか？多分見つからないと思う。何故だろう？
- 他の手法を試してみる。
  - neural network
  - SMO（support vector machineの一つ）
  - naïve Bayes
- それぞれの正解率、実行時間などを比べてみよう

データを図にしたもの:
test_images.zip,
train_images.zip

「他の手法」の選び方

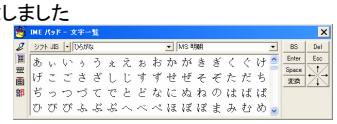


実験2: 歌詞の分類

- 楽曲のジャンルごとに、歌詞に使う言葉や言い回しは違う。
- 同様なことは、短歌でもいえる。
  - 水谷静夫先生（計量国語学の創始者. 計量国語学会は1957年(!)創立）が、その昔、白樺派とアララギ派の短歌が、そこに使われている単語を用いると、綺麗に分離されることを示した。
- 単語で分類するのは、ちょっと手間なので（準備がネ）、モーラ（音節）で分類できるか試してみよう
  - 実はモーラも少し面倒なので、「50音」を用いる。

データの説明

- 童謡と演歌それぞれ10曲の歌詞を「50音」にわけ、その頻度を数える。
- 文字コードの並びが50音ではないので並び替えたいのであるが、少々面倒
  - 「あいうえお」は纏めたい etc.
  - Excelファイルを用意しました

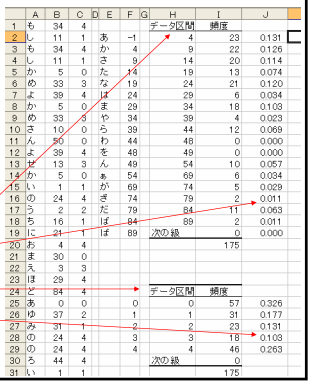


- 頻度は正規化する（頻度の総和が1となるように）。歌によって長さが異なるからである。

実験手順: データの準備1

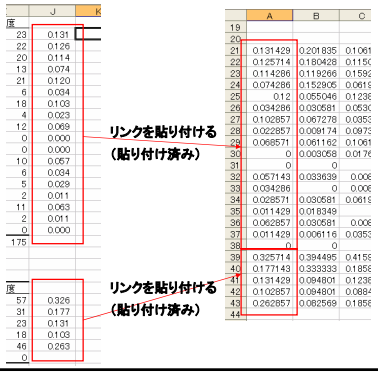
列Aに平仮名で一文字ごと歌詞を書く。
そうすると、列Bには列Aの文字を五十音のコード（自作）に変換するマクロが書かれているので、列Bには、値として、そのコードが得られる。
列Cには「あいうえお」の区別が得られる。

ヒストグラムを作るために、頻度を数える。
列E,Fはヒストグラム作成用のデータである。
Excel メニューで、「ツール」⇒「分析データ」⇒「ヒストグラム」とし、ヒストグラムを作成する。
正規化した頻度が得られる
「あいうえお」を数えて、そのヒストグラムも作成する。
正規化した頻度が得られる



	A	B	C	D	E	F	G	H	I	J
1	も	34	4					データ区間	頻度	
2	し	11	1	あ	-1			4	23	0.131
3	も	34	4	か	4			9	22	0.126
4	し	11	1	さ	9			14	20	0.114
5	か	5	0	た	14			19	13	0.074
6	め	33	3	な	19			24	21	0.120
7	よ	39	4	は	24			29	6	0.034
8	か	5	0	ま	29			34	18	0.103
9	め	33	3	や	34			39	4	0.023
10	さ	10	0	ら	39			44	12	0.069
11	ん	50	0	わ	44			48	0	0.000
12	よ	39	4	を	48			49	0	0.000
13	せ	13	3	ん	49			54	10	0.057
14	か	5	0	あ	54			69	6	0.034
15	い	1	1	か	69			74	5	0.029
16	の	24	4	さ	74			79	2	0.011
17	う	2	2	た	79			84	11	0.063
18	ち	16	1	は	84			89	2	0.011
19	に	21	1	ば	89			次の級	0	0.000
20	お	4	4						175	
21	ま	30	0							
22	え	3	3							
23	ほ	29	4							
24	ど	84	4					データ区間	頻度	
25	あ	0	0		0			0	57	0.326
26	ゆ	37	2		1			1	31	0.177
27	み	31	1		2			2	23	0.131
28	の	24	4		3			3	18	0.103
29	の	24	4		4			4	46	0.263
30	ろ	44	4					次の級	0	
31	い	1	1						175	

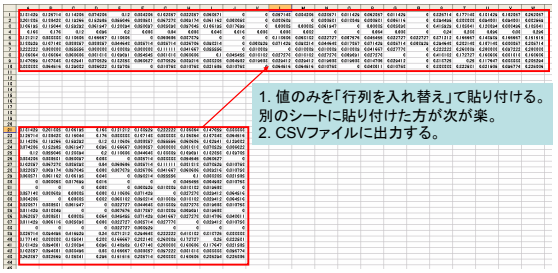
実験手順: データの準備2



リンクを貼り付ける
（貼り付け済み）

リンクを貼り付ける
（貼り付け済み）

実験手順: データの準備3



1. 値のみを「行列を入れ替え」て貼り付ける。別のシートに貼り付けた方が次が楽。
2. CSVファイルに出力する。

実験手順: データの準備4

- 童謡10曲、演歌10曲について、データを作成する。
- arff ファイルとする。

ここまでは作成済み
童謡.xls
演歌.xls

- 右のような、ヘッダを作る
- 各データ（一行に一曲、全20行）の最後に、童謡なら「,0,0」、演歌なら「,1,1」をつける。
- 同じものだが、片方は数値、片方はカテゴリ（分類クラス）である

「,0」
「,1」

削除

```
@relation 'songs'
@attribute あ real
@attribute か real
@attribute さ real
@attribute た real
@attribute な real
@attribute は real
@attribute ま real
@attribute や real
@attribute ら real
@attribute わ real
@attribute を real
@attribute ん real
@attribute ぁ real
@attribute が real
@attribute ざ real
@attribute だ real
@attribute ば real
@attribute ぱ real
@attribute A real
@attribute I real
@attribute U real
@attribute E real
@attribute O real
@attribute class {0,1}
@data
```

実験手順: 実験と検証

- Weka を用いてみる
  - J48, SMO, naïve Bayes など
- 正解率が高いかな？
- 決定木がまともにできているかどうか、わかりますか？
- では、皆さんご自分で、
  - 童謡・演歌に10曲ずつ追加するか、
  - 新たなジャンルで、最低10曲ずつデータを作成して

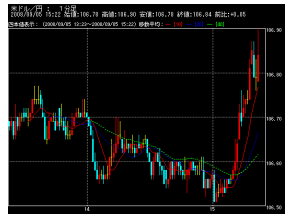
実験を行ってください。

実験3 ドル円レートの予測

- FX: 外国為替証拠金取引
  - 証拠金（保証金）を業者に預託し、主に差金決済による通貨の売買を行なう取引
- FXで利益を上げることができるのだろうか？
  - 仲介業者の取り分（スプレッド）は小さい。宝くじとは違う
  - しかし、典型的なギャンブル。ゼロサムゲーム。勝者がいれば、敗者がいる。そして、敗者が圧倒的に多い（80-20の法則、冪法則）
    と思う。
- 値動きは、基本的には、ランダムウォークのはず
  - すなわち、予測不能のはず。
- しかし、少し、試してみよう

実験データの説明

- 米ドル（USD）を日本円（JPY）で売買する
- 価格(?)の単位は 0.01 円
- 分単位の値動きを見る
  - 分足という



東京金融取引所

実験データの説明

- USD/JPYの分足を用いる
  - まず、(いつでもよいのだが)2008年9月1日にしよう
  - Forexite というサイトのものを用いる。時刻は GMT+1 (Central European Time). データの正確性の保証はない。
  - 24時間中の、一分毎、Open（始値）、High（高値）、Low（安値）、Close（終値）が時系列に記されている。
  - ある「分」の「終値 – 始値」（収益）を予測する
  - 難しい: 何を、予測の根拠に用いるか？ i.e. どんな特徴量を用いるか
  - （まずは）5分前からの各「分」の収益を用いよう

010908.zip

010908.zip 中の 010908.txt

```
1 <TICKER>,<DTYYYYMMDD>,<TIME>,<OPEN>,<HIGH>,<LOW>,<CLOSE>
2 EURUSD,20080901,000000,1.4698,1.4698,1.4697,1.4698
3 EURUSD,20080901,000100,1.4699,1.4700,1.4698,1.4699

4320 USDCHF,20080901,235800,1.1017,1.1017,1.1017,1.1017
4321 USDCHF,20080901,235900,1.1017,1.1024,1.1017,1.1021
4322 USDJPY,20080901,000000,108.31,108.31,108.31,108.31   ここから
4323 USDJPY,20080901,000100,108.30,108.31,108.29,108.29
4324 USDJPY,20080901,000200,108.28,108.30,108.28,108.29
4325 USDJPY,20080901,000300,108.30,108.30,108.30,108.30
```

作業手順

- 2008年9月1日のデータを得る。
- Excelファイルとし、各分について5分前からの収益を計算する。当「分」の収益も求める。
- 予測問題を簡単にするために、up or down を示す値をつける (上昇していたら +1, 下降していたら –1)
  - このファイルは用意してある。

010908.zip 中の 010908.txt から USDJPY を全部

USDJPY080901.xls

H7 =F2-C2

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	<TIME>	<OPEN>	<HIGH>	<LOW>	<CLOSE>		5 before	4 before	3 before	2 before	1 before	predict	updown
2	0	108.31	108.31	108.31	108.31								
3	100	108.3	108.31	108.29	108.29								
4	200	108.28	108.3	108.28	108.29								
5	300	108.3	108.3	108.3	108.3								
6	400	108.29	108.31	108.29	108.29								
7	500	108.3	108.3	108.25	108.25		0	-0.01	0.01	0	0	-0.05	-1
8	600	108.25	108.25	108.25	108.25		-0.01	0.01	0	0	-0.05	0	0
9	700	108.25	108.26	108.24	108.24		0.01	0	0	-0.05	0	-0.01	-1

実験手順: ファイルの準備

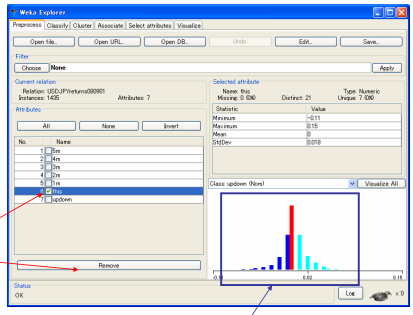
- 収益の部分のみを取り出し、csv ファイルを作る。
- arff のヘッダーをつけ、Weka用の arff ファイルとする。メモ帳なりエディタなりを使う方が間違いが少ない

```
1 @relation USDJPYreturns080901
2 @attribute 5m numeric
3 @attribute 4m numeric
4 @attribute 3m numeric
5 @attribute 2m numeric
6 @attribute 1m numeric
7 @attribute this numeric
8 @attribute updown {-1,0,1}
9 @data
10 0,-0.01,0.01,0,0,-0.05,-1
11 -0.01,0.01,0,0,-0.05,0,0
```

実験手順: 味見

- Weka を使う
  - 使うときに、当「分」の収益という属性は "Preprocess" で remove して下さい（次スライド）
  - trees にある J48: 決定木
  - functions にある neural network
  - functions にあるSMO: support vector machine の一つ
  - Bayes にある naïve Bayes
- どれも正解率は 1/3 ぐらいであろう。
  - 気軽に、なんだぁ一様ランダムなんだ、と思わないで下さい。データを見てみると、収益が正、零、負になる割合がほぼ 1/3 なのです。信じられますか？

Weka手順：蛇足と補足



実験手順: 他の日との比較

- 別の日のデータではどうか？
- 9月2日、9月3日、9月4日で試してみよう
  - xls, csv, arffファイルは自分で作りましょう
- どうですか？

020908.zip
030908.zip
040908.zip

- 予測するための情報が不足でしょうか？
- きっとそうでしょう。では、一分前の「分」の高値安値を含めてみましょう。
  - 9月1日のファイルが作ってあります。他の日のファイルも作り、試しましょう。
  - でもうまくいきません。

USDJPY080901A.xls

実験手順

- では、5分足（5分間一区切り）を試してみよう。
  - 1分足では、他の人の動きをみて動く（相関が発生する）ということが少ないので、動きはランダムになり、従って、予測できない。
  - しかし、5分程度みれば、データ間に相関が発生し、従って、予測可能となる可能性がある。
  - 本当か？

実際には逆で、収益の時間相関は20分ぐらいまでは存在すると報告されている。例えば
P. Gopikrishnan, et al. Scaling of the distribution of fluctuations of financial market indices, Physical Review E vol. 60, 5305 - 5316 (1999)

実験手順

- まず、5分ごとのデータにする前に、「5分間」の始値、高値、安値、終値を求める
- 次に、5分区切りのデータを抽出する。

コピー　最小値　最大値　コピー

A	B	C	D	E	F	G	H	I	J	K	L
<YMMDD>	<TIME>	<OPEN>	<HIGH>	<LOW>	<CLOSE>		<TIME+500>	<OPEN>	<HIGH +4m>	<LOW +4m>	<CLOSE +4m>
20080902	0	108.18	108.18	108.14	108.14		0	108.18	108.18	108.07	108.08
20080902	100	108.12	108.13	108.12	108.12		100	108.12	108.14	108.07	108.08
20080902	200	108.13	108.14	108.12	108.12		200	108.13	108.14	108.07	108.07
20080902	300	108.11	108.11	108.09	108.1		300	108.11	108.11	108.07	108.08
20080902	400	108.09	108.09	108.07	108.08		400	108.09	108.09	108.07	108.08
20080902	500	108.08	108.08	108.08	108.08		0	108.08	108.09	108.07	108.09

<TIME>を500で割った余り
あとで、この値でソートすれば、
5分区切りのデータが得られる

実験手順

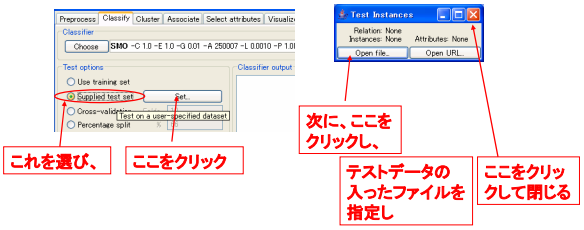
- Weka を用いてみる
  - J48, SMO, naïve Bayes など
- 少し正解率が高い
  - しかし、データ分布を見てみると、収益=0 の事例が少ないことが分かる。
  - やはり予測できないのか。
  - しかし、データ数が少ないからかもしれない。
- 9月1日~4日で試してみよう
  - 少しよいか？
  - しかし、別の日、例えば、8月6日~8日でテストをしてみたらどうだろうか？

060808.zip
070808.zip
080808.zip

実験手順

- 別のデータでテストする方法
- 学習データと同じ属性数（並び）のテストデータを用意する。
- 今回は、当「分」の収益という属性を持ったファイルしかないので、そのままではテスト用のデータファイルにはならない。そこで、当「分」の収益という属性を削除したデータファイルを作る。Weka でできる。

Weka補足: テストデータの指定



Weka補足: ファイルを作る