

# データマイニングにおける受動性と能動性

## Passivity and Activity in Data Mining

鈴木 督久

Tokuhisa SUZUKI

日経リサーチ マーケティング局

Nikkei Research Inc.

**Abstract :** There is an impression of acquiring useful knowledge automatically using tool in data mining. There is such side and there is also side which is not so. This paper discusses both sides from the viewpoint of the passivity nature and the activity nature in data mining. As an example, the prediction model by neural networks is shown and the comparison with traditional statistical data analysis and data mining is also mentioned.

### 1 はじめに

データマイニングとは何か という定義に関してデータマイニングのユーザーは興味がないし、また定義する必要も低い。従って、今日なおデータマイニングの定義が一意でないことは問題ではない。興味ある課題は、許容できるコストの範囲内で、ビジネスに有用な知識をデータベースから獲得(KDD)することができるか否かに集中している。ビジネスに有用な知識とは、利益を産む知識のことである。

ビジネスマンの間のデータマイニングに関する会話を聞いていると、データマイニングという用語は、ほとんど多変量解析(MVA)という意味で使われているような印象さえある。また統計学者であれば、探索的データ解析(EDA)が大規模データを分析対象とする態度と、どこが違うのかという印象を持つであろう。精神はEDAに通じるところがあるし、方法はMVAの範囲に含まれているからである。

もちろん違いは、ある。量が質を変えるということもあるが、少なくともビジネスマンは大規模なEDAやMVAとは違う概念としてデータマイニングと用語しているようである。ビジネスマンが「データマイニング」と発語する時、日本における毎度のビジネス英単語の目新しさとともに、意識的であるにせよ無意識的であるにせよ、「統計学」という厳粛な体系から解放された自由な気分が漂っている。そ

こでは多変量正規分布を要求されることはないし、データが汚れていることを咎められる恐縮もない。ただ、儲かる知識を得ることだけが目標である。

従来からビジネス分野でも「データ解析」はあったにもかかわらず、そういわずに「データマイニング」と呼んだだけで、「マイニング」は「解析」よりも大衆的に歓迎された。何故だろうか。データ解析には統計専門家が必要だが、データマイニングには業務専門家とマイニングツールが揃えば、宝が発見される雰囲気は漂っていたからである。もちろん、データをツールに入れれば宝が出てくるような虫のいい話はデータマイニングにおいてさえも、ない。「データ」は消極的に存在しているかも知れないが、「マイニング」は積極的でなければ宝はやってこない。

このことを、データマイニングにおける受動性と能動性という観点から実例とともに整理してみたい。具体的にはニューラルネットによる予測モデルをPOSデータに適用した事例を示す。伝統的な統計的データ解析との比較にも言及する。

### 参考文献

- [1] 岩崎学：データマイニングと知識発見 - 統計学の視点から - 行動計量学 26 (1999), 46-58.
- [2] 豊田秀樹：金鉱を掘り当てる統計学 データマイニング入門 . 講談社 . (2001) .

人工知能学会 第17回 AIシンポジウム  
データマイニングにおける受動性と能動性

鈴木督久  
日経リサーチ  
マーケティング局

1

## データマイニング (KDD) とは何か

定義よりもキーワード

- ビジネスに役立つ知識の抽出 (採掘・発見)
- コスト・時間・有用性のトレードオフ
- データベース・データ解析・大規模データ
- マーケティング・統計学・情報科学
- 方法と応用, ツールと人間, データと知識

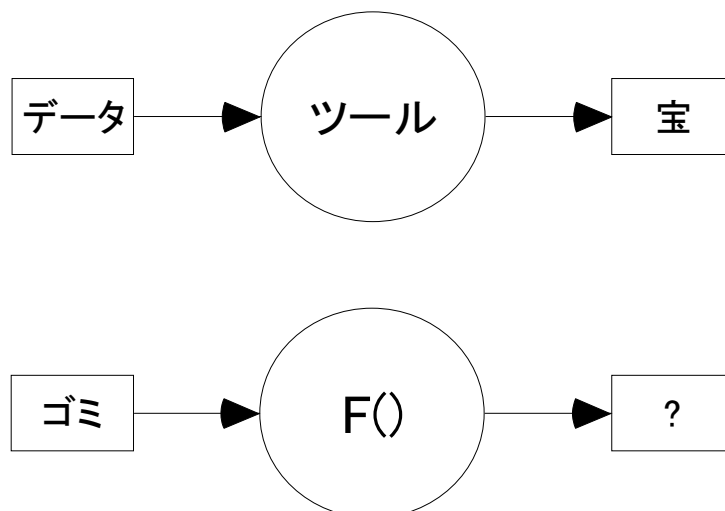
2

## コンセプト

- 狩猟と農耕
- 地味と派手
- 探索的と検証的(探偵と裁判官)
- 静的と動的
- 受動性と能動性

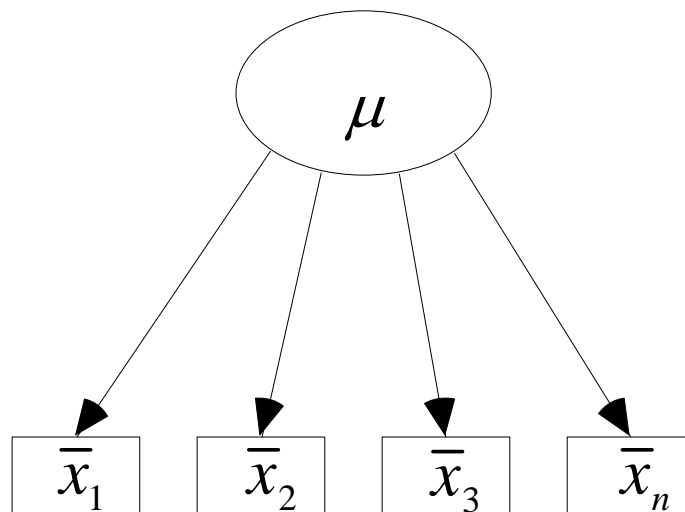
3

## ツールと人間, データと知識



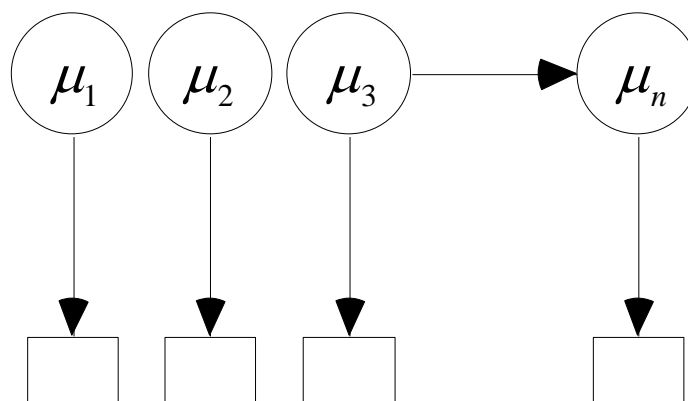
4

## 静的と動的(母数と標本値)



5

## 母集団が動的に変化する状況



6

## データの性質

	能動的	検証的	高価	少量
1.実験データ	↕	↕	↕	↕
2.調査データ				
3.実績データ				
	受動的	探索的	安価	大量

7

## Data mining : その受動性と能動性

受動的	能動的
<ul style="list-style-type: none"><li>● Data</li><li>● 業務の収集・集積</li><li>● 変数変換</li><li>● 変数選択</li></ul>	<ul style="list-style-type: none"><li>● Mining</li><li>● 実験の計画・調査の設計</li><li>● 隠れ層とユニットの個数</li><li>● 変数生成</li><li>● クロスバリデーション</li></ul>

8

## 日経のPOSデータ

- NEEDS-SCAN(日本経済新聞社のPOS情報サービス)
- スーパー・コンビニ46チェーン235店舗・1日80万人来客
- 1日800万商品購入, 年間6000億円
- 日本チェーンストア協会加盟7376店舗の食・雑・医・化＝10.5兆の5.7%に相当
- 流通システム開発センターのRDS/POSも導入(中堅420店舗:年商3000億円)してデータベース化
- データの質
- 商品マスター

## 分析の目的

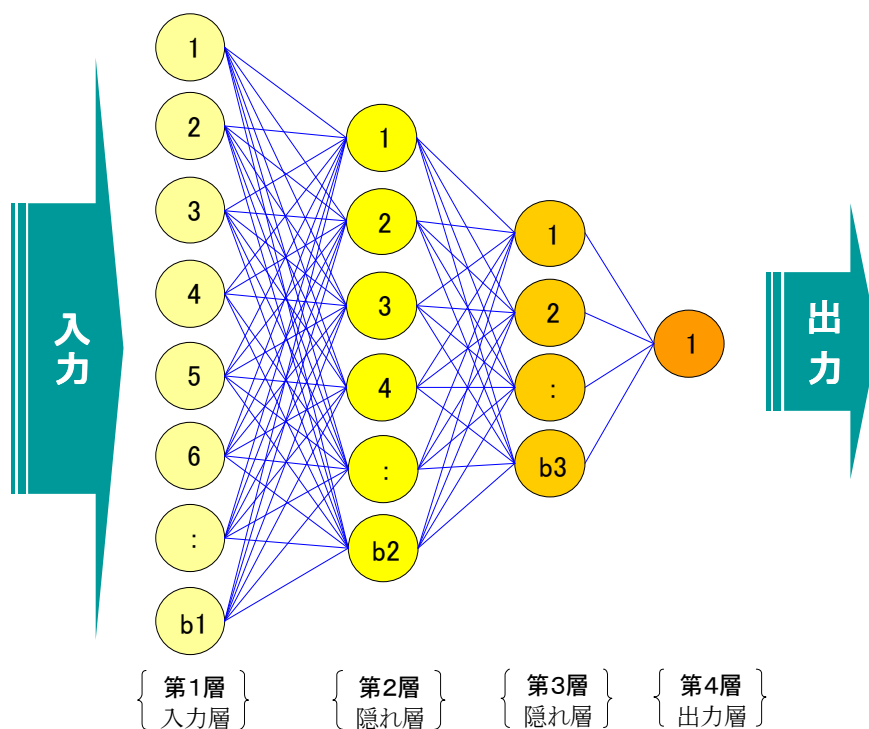
- 明日の牛乳が何本売れるかを予測したい
- 毎日の予測業務の熟練者に知識のモデル化
- POSデータ(&気象データ)からマイニング

## 予測の戦略

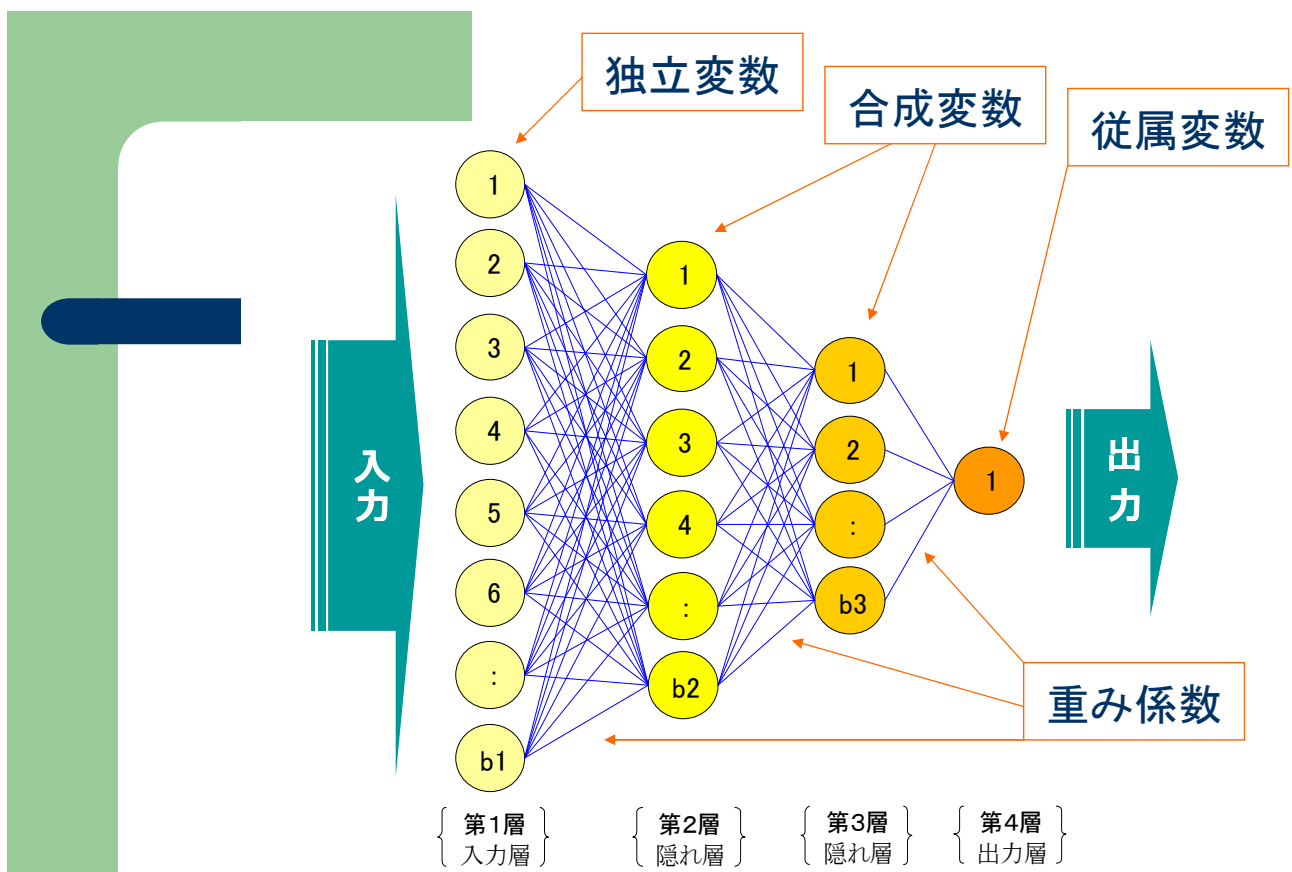
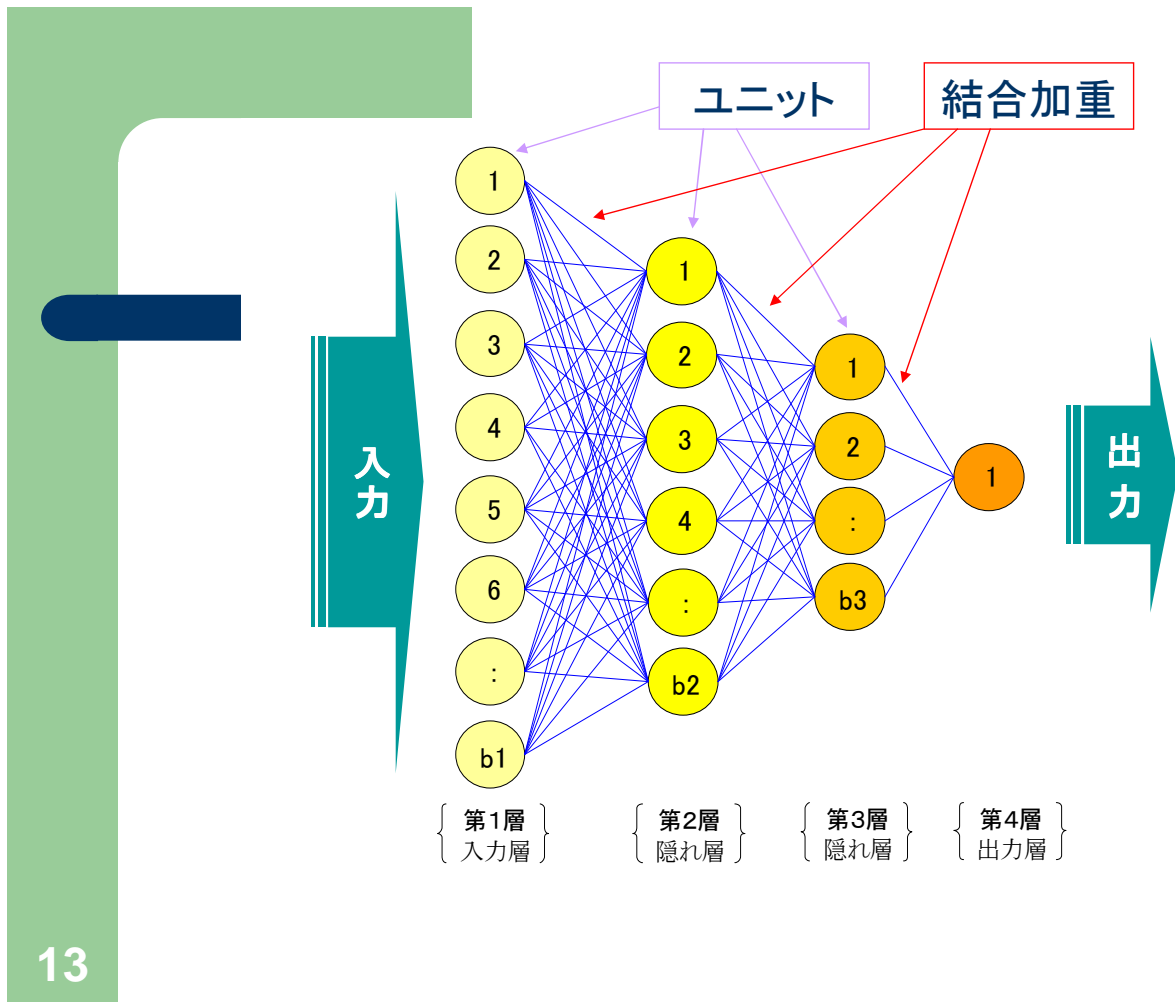
- 店舗別モデルか地域全体モデルか
- モデル構成用のデータの期間
- モデルの利用期間(有効期限)
  
- 重回帰分析, ニューラルネットワーク

11

## 階層型ネットワークモデル

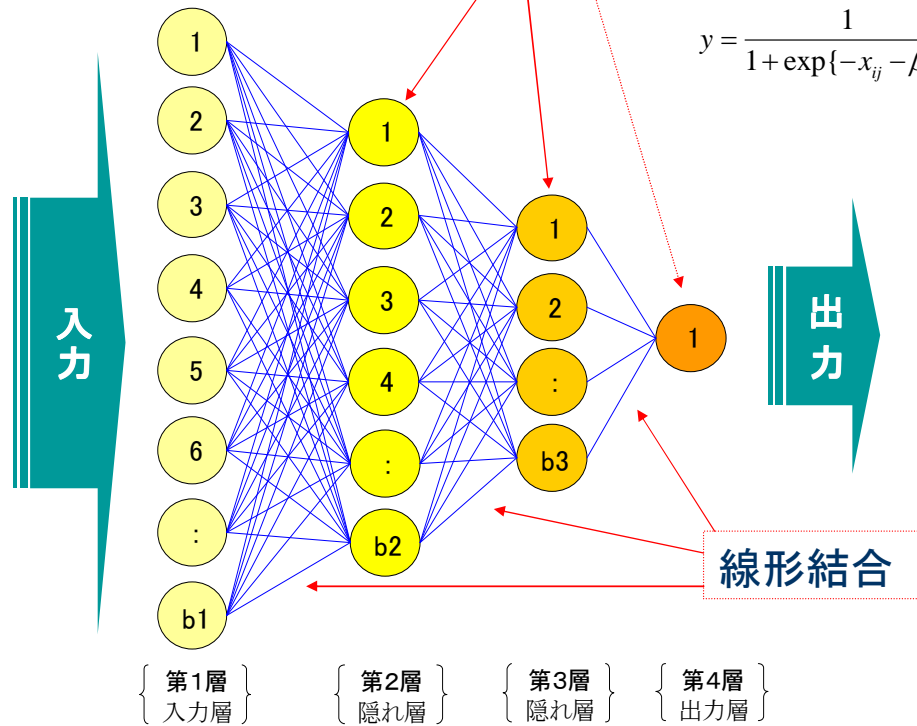


12



シグモイド関数 (ex. ロジスティック関数)

$$y = \frac{1}{1 + \exp\{-x_{ij} - \beta_{ij}\}}$$



## 用語

学習	→	母数の推定
信号	→	観測値
ユニット	→	変数
結合加重	→	重み係数
入力層	→	独立変数
出力層	→	従属変数
隠れ層	→	合成変数
教師信号	→	従属変数值

## 予備解析

- ツールかプログラムか
- 店舗ごとの吟味
- 商品ごとの吟味
- 従属変数の吟味
- 外れ値の検討
- 予測変数の吟味

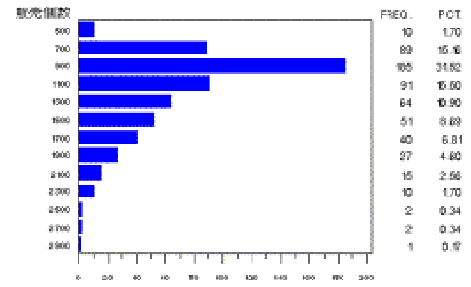
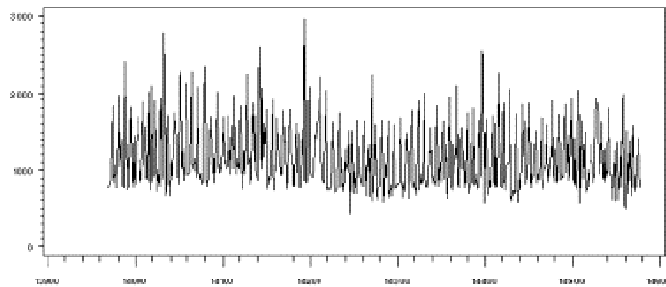
17

## 関東の5店舗・20ヶ月／営業日と商品数

店舗	A	B	C	D	E
営業日数	587	588	584	571	588
休店日数	22	21	25	38	21
牛の乳商品種類	43	23	23	21	35

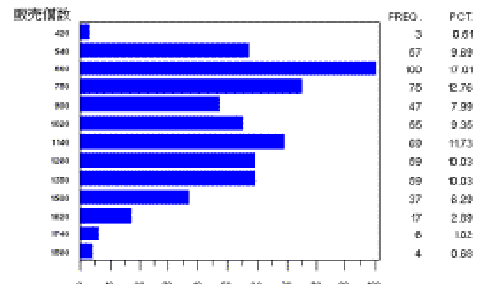
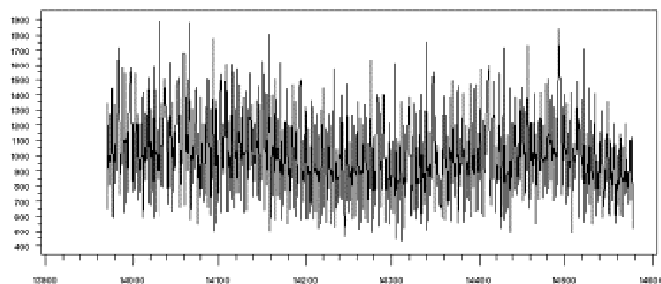
18

## 販売本数の推移とヒストグラム(A店)



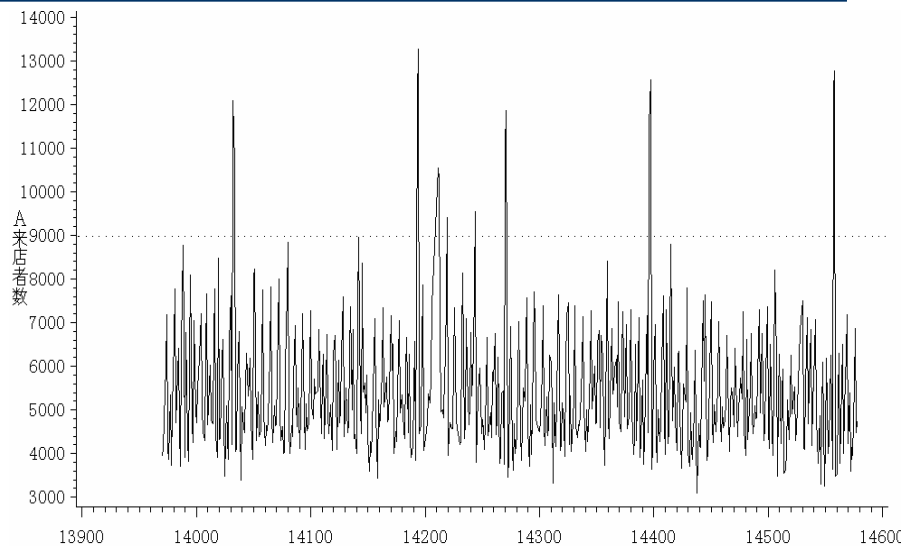
19

## 販売本数の推移とヒストグラム(B店)



20

## 来店客数(A店)



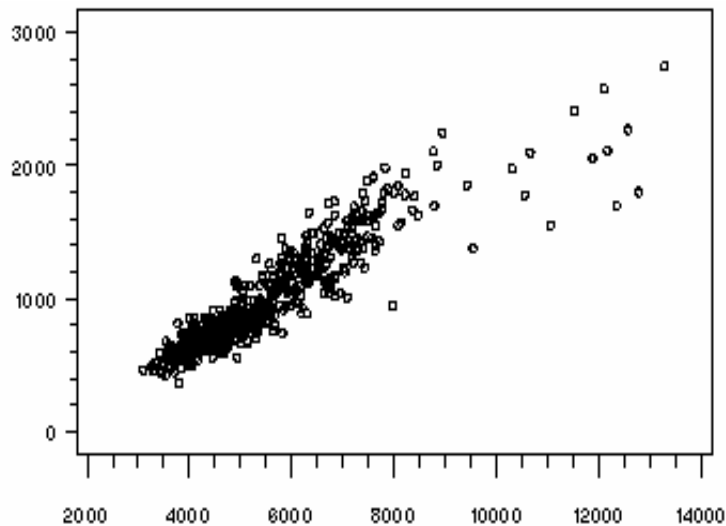
21

## 外れ値(特異日)は特殊な日

CBS	DATE	WEEK	WEEKDAY	N_STORE	TENKI	OUTLIER
1	1998-06-02	火	3	12095	くもり	*
2	1998-06-03	水	4	10661	雨	*
3	1998-11-10	火	3	12165	くもり	*
4	1998-11-11	水	4	13281	晴れ	*
5	1998-11-28	土	7	10554	晴れ	*
6	1998-11-29	日	1	10902	晴れ	*
7	1998-12-06	日	1	9400	晴れ	*
8	1998-12-31	木	5	9539	晴れ	*
9	1999-01-26	火	3	11051	晴れ	*
10	1999-01-27	水	4	11873	くもり	*
11	1999-06-01	tue	3	11524	晴れ	*
12	1999-06-02	wed	4	12574	晴れ	*
13	1999-11-09	tue	3	12338	晴れ	*
14	1999-11-10	wed	4	12774	晴れ	*

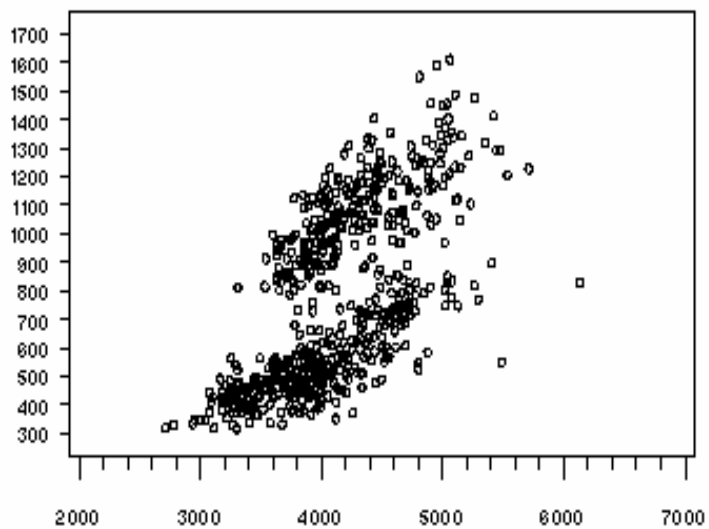
22

## 来店客数と販売個数(A店)



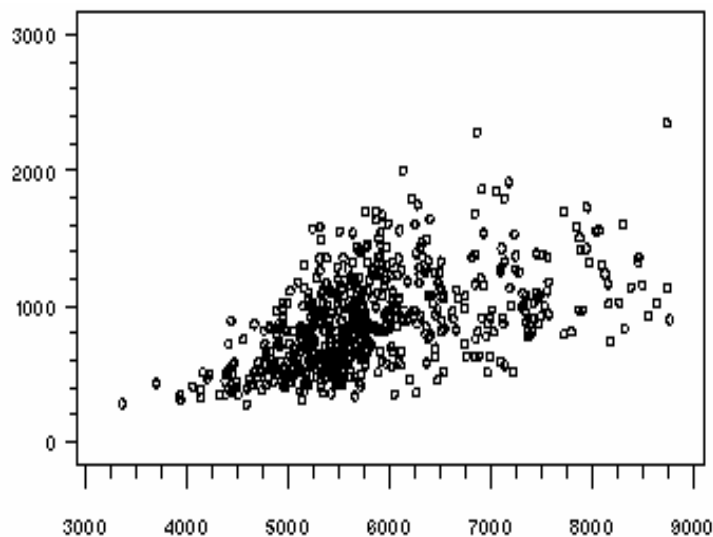
23

## 来店客数と販売個数(B店)



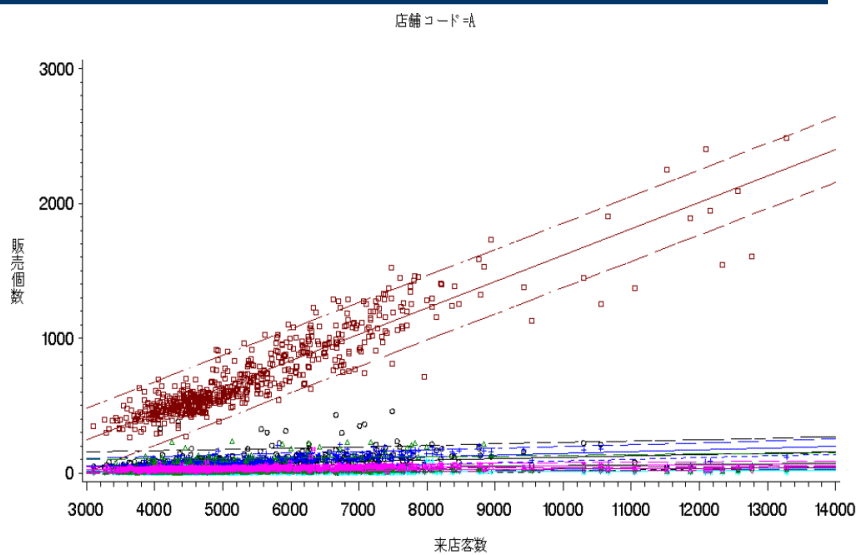
24

## 来店客数と販売個数(C店)



25

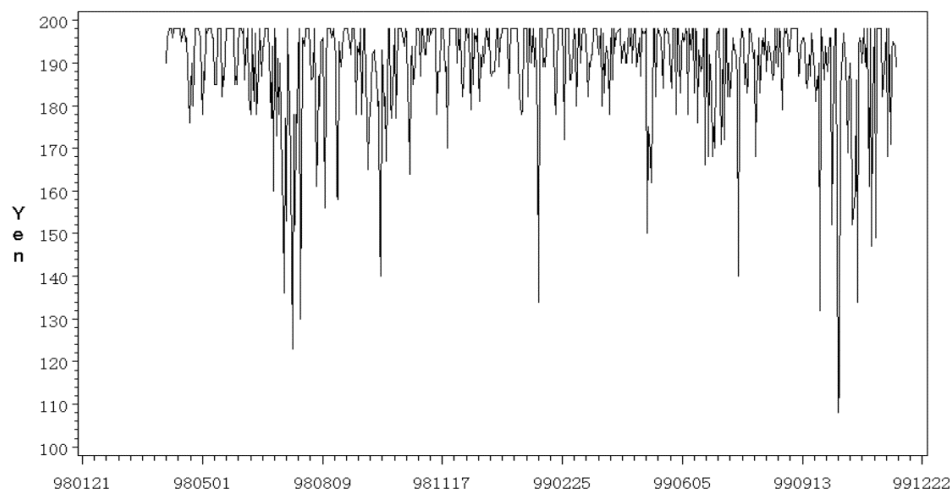
## 来店客数と販売個数(層別)A



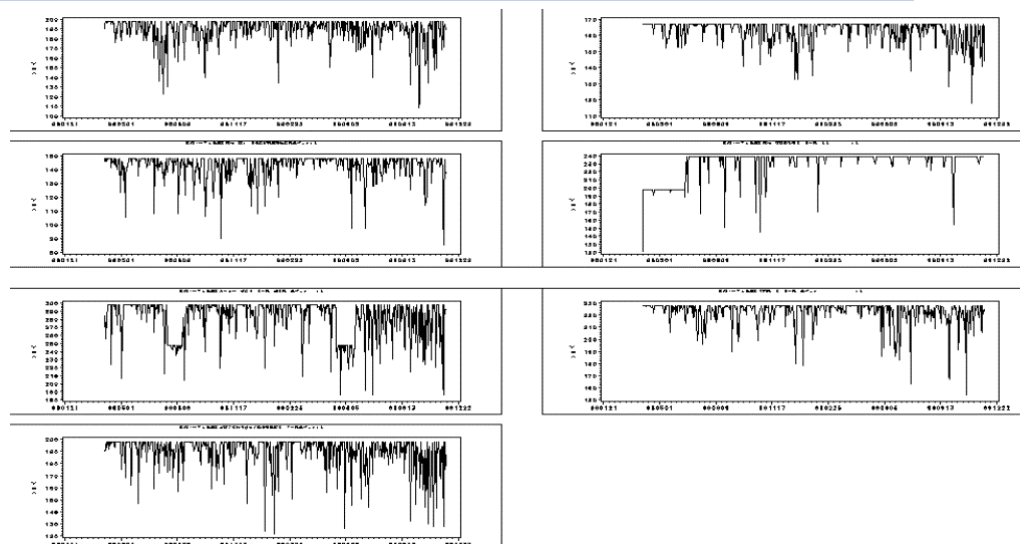
26

## 価格の日次変動(A店A商品)

店舗コード=A NAMEX=PB-a XX 北海道3.7牛乳 紙パック1L



## 価格の日次変動(A店. 各商品)



## 予備解析後の知見

- 販売「本数」を従属変数とする
- 店舗全体の本数を予測（特定商品でなく）
- 1品商品のみを対象とする
- 継続的に販売実績のある商品を対象とする
- 店舗ごとに予測モデルを構成し、地域全体の予測にあたって、各予測値を積み上げる
- 外れ値（特異日）はモデル構成では除外する

29

## NN における modeling

- Training, Validation, Test dataset の配分
- クロスバリデーションと試行錯誤
- 隠れ層のユニット数の探索
- 隠れ層の個数の探索（通常は1つだが）
- 入力層のユニットの準備（新しい変数も）
- 結合加重（重み）の解釈はしない
- 変数変換は必要ない（線形化, 対称化変換）
- 変数選択の必要はない（有望な変数を投入）

30

## 独立変数(75変数) = 入力信号

- |                                |         |
|--------------------------------|---------|
| (1) 曜日 (2 値変数化)                | : 7 変数  |
| (2) 3 日前までの販売個数 (店舗全体と上位 4 商品) | : 15 変数 |
| (3) 3 日前までの販売価格 (上位 4 商品)      | : 12 変数 |
| (4) 3 日前までの来店客数                | : 3 変数  |
| (5) 3 日前までの降水量                 | : 3 変数  |
| (6) 3 日前までの日照時間                | : 3 変数  |
| (7) 当日および 3 日前までの不快指数          | : 4 変数  |
| (8) 当日および 3 日前までの最高気温          | : 4 変数  |
| (9) 当日および 3 日前までの最低気温          | : 4 変数  |
| (10) 当日および 3 日前までの平均気温         | : 4 変数  |
| (11) 当日および 3 日前までの湿度           | : 4 変数  |
| (12) 当日および 3 日前までの天気 (2 値変数化)  | : 12 変数 |

31

## Neural Model

- Model 構成用データ: 12ヶ月
  - Training : 70% Validation : 30%
- Model 検証用データ: 次の1ヶ月
- 4層モデル(中間層2個)
  - 第1(入力)層 = 75
  - 第2(中間)層 = 30 ± 10
  - 第3(中間)層 = 15 ± 10
  - 第4(出力)層 = 1
- シグモイド関数はデフォルト

32