

数理情報工学特論第一  
【機械学習とデータマイニング】  
1章：概論（1）

かしま ひさし  
鹿島 久嗣  
（数理 6 研）

kashima@mist.i.~



## この講義の構成：

---

- 機械学習とデータマイニングについての講義を行います
  - 中川先生の講義と内容的に被る部分があります
- 最初の2回で概論、残りの回で各論を行います
  - 途中、ゲスト講演者を呼ぶかもしれません
- 評価：
  - レポート（月に一回：4月、5月、6月の末）
  - 出席
  - なお、試験は行いません
- 注意：4/30(金)は、月曜の講義を行うため、この講義がありません

## この講義の目的：（統計的）機械学習技術入門

---

- 企業のビジネスにおいて差別化要因になるデータ解析手法のひとつである機械学習技術について知ってもらう
  - 「知っておくと得する（≡つぶしがきく）」という気持ちになってもらう
- 実際の問題に対し、機械学習問題として捉え、機械学習手法を使って解決できるよう（な気）になってもらう
  - 機械学習の問題設定
  - 機械学習の数理モデル
  - 機械学習のアルゴリズム
  - 機械学習の評価方法

---

# 機械学習とは

# 機械学習とは、データ分析技術の一流派のようなものです

- 機械学習とは
  - 「人間のもつ”学習能力”を機械（計算機）にも持たせる」ことを目指す研究分野
  - もともとは人工知能の一分野として始まる
    - 論理推論がベース
  - 現在では、「統計的」機械学習が主流（≡機械学習）
    - 遺伝子情報処理、自然言語処理、他、ビジネス分野での成功
- 現在では、データ解析技術一般を指すほかの言葉とあまり変わらない
  - 統計／データマイニング／パターン認識など。  
（多少のニュアンスの違いはあるが、基本的に好みの問題）

## 学習問題の区分：

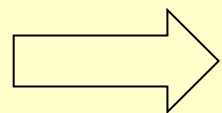
### 教師付き学習と教師無し学習に分類されます

---

- まずは、学習の問題の本質的な部分を、数学的に扱えるように、定式化する必要がある
- 学習者を、入出力のあるシステムであると捉え、学習者に対する入力と、それに対する出力の関係をモデル化する
  - 入力：視覚などからの信号（実数値ベクトルで表現）
  - 出力：入力を表す概念、入力に対してとる行動
- どうやら2つの重要な基本問題があるらしいということになった
  - 教師無し学習：入力を見ているうちに、どんなものが現れやすいかなどのパターンが分かってくる
  - 教師付き学習：入力に対する出力を試行錯誤するうちに、どういう入力のおかげでどういう出力をすればよいか分かってくる

# 入力の表現：学習者に入ってくる入力情報は通常、実数値ベクトル（特徴ベクトル）として表現します

- 入力信号を、その特徴量を列挙した  $D$ 次元の実数値ベクトル  $\mathbf{x}$  として表現する
  - $\mathbf{x}$  を「特徴ベクトル」と呼ぶ
  - その領域を「特徴空間」と呼ぶ



$\mathbf{x} :=$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$$

赤みの  
度合い

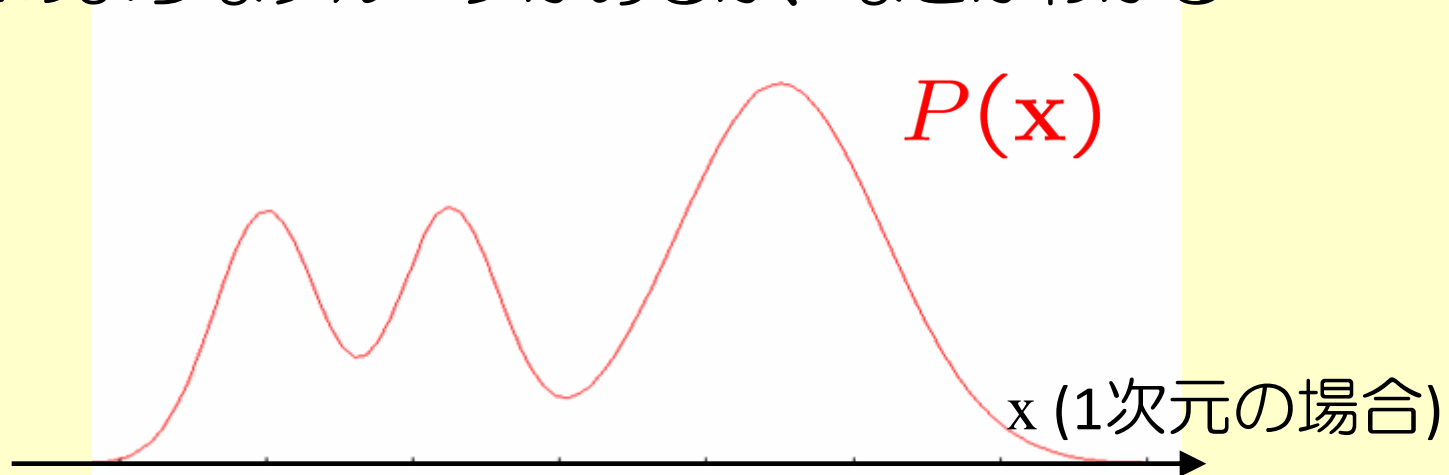
ビタミンC  
含有量

「今日の気温」  
などでもよい

- 特徴ベクトル  $\mathbf{x}$  はどのようにデザインしたらよいか？
  - 完全にドメイン依存、一般的な解はなく、目的にあわせて、ユーザーがデザインする

# 教師無し学習は、 特徴空間上での確率分布の推定問題として捉えられます

- 教師無し学習：たくさんの入力信号を見ているうちに、どんなものが現れやすいかなどのパターンが分かってくる
- 入力信号  $\mathbf{x}$ ： $D$ 次元の実数値ベクトルとして表現
$$\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} \in \mathfrak{R}^D$$
- この入力信号  $\mathbf{x}$  上の確率分布  $P(\mathbf{x})$  を考える
  - $P(\mathbf{x})$  の形をみることで、どのあたりの入力信号が現れやすいか／どのようなグループがあるか、などがわかる





# 教師無し学習は、 特徴空間上での確率分布の推定問題として捉えられます

- 目的：訓練データ ( $N$  個の入力信号) から、 $P(\mathbf{x})$  を推定する

$(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots, \mathbf{x}^{(N)})$

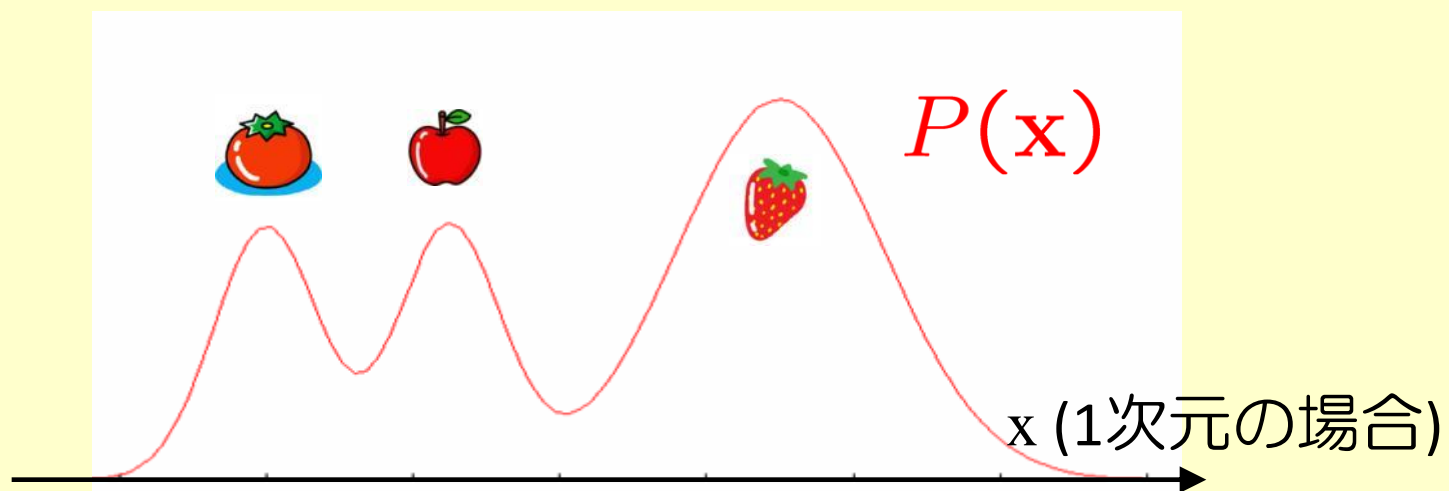
1つめの  
データ

2つめの  
データ

...

$$\mathbf{x}^{(i)} := \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_D^{(i)} \end{pmatrix} \in \mathcal{R}^D$$

- 教師無し学習は、（大げさにいえば）明示的に指定されることなしに、“概念”を形成するプロセスを表している



## 一方、教師付き学習は、条件付確率分布の推定問題です

- 条件付分布  $P(y|\mathbf{x})$  : 入力信号  $\mathbf{x}$  を条件とした、出力  $y$  の確率分布
  - 入力信号  $\mathbf{x}$  は、 $D$  次元の実数値ベクトル
  - 一方、出力  $y$  は1次元
    - データの属するカテゴリ
      - > +1 もしくは -1 の2つ ( $y \in \{+1, -1\}$ )  
(例 : 🍏 か否か)
      - > 複数のカテゴリ  $\{A, B, C, D, \dots\}$   
(例 : 🍏 か 🍅 か 🍓 か)
    - 実数値 :  $y \in \mathbb{R}$
- たとえば  $\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix}$  が 🍏 である確率はいくつか? という質問に答える

今日は  
基本コシ

# 教師付き学習は、条件付確率分布の推定問題です

- 目的：訓練データ（ $N$ 個の入力信号）から、 $P(y|\mathbf{x})$ を推定する

$$\underbrace{((\mathbf{x}^{(1)}, y^{(1)}))}_{\text{1つ目の入出力ペア}}, \underbrace{((\mathbf{x}^{(2)}, y^{(2)}))}_{\text{2つ目の入出力ペア}}, \underbrace{((\mathbf{x}^{(3)}, y^{(3)}))}_{\text{3つ目の入出力ペア}}, \dots, (\mathbf{x}^{(N)}, y^{(N)})$$

–  $\mathbf{x}^{(i)}$ :  $i$ 番目の事例の入力信号ベクトル

–  $y^{(i)}$ :  $i$ 番目の事例に対する正しい出力

(🍏ならば +1, 違うなら -1)

- 教師付き学習：与えられた入力信号に対する、あるべき出力を教師信号として、入出力の関係を学習する

# 教師なし学習モデルの典型：混合正規分布モデル

## 単一の正規分布では表現力が十分ではありません

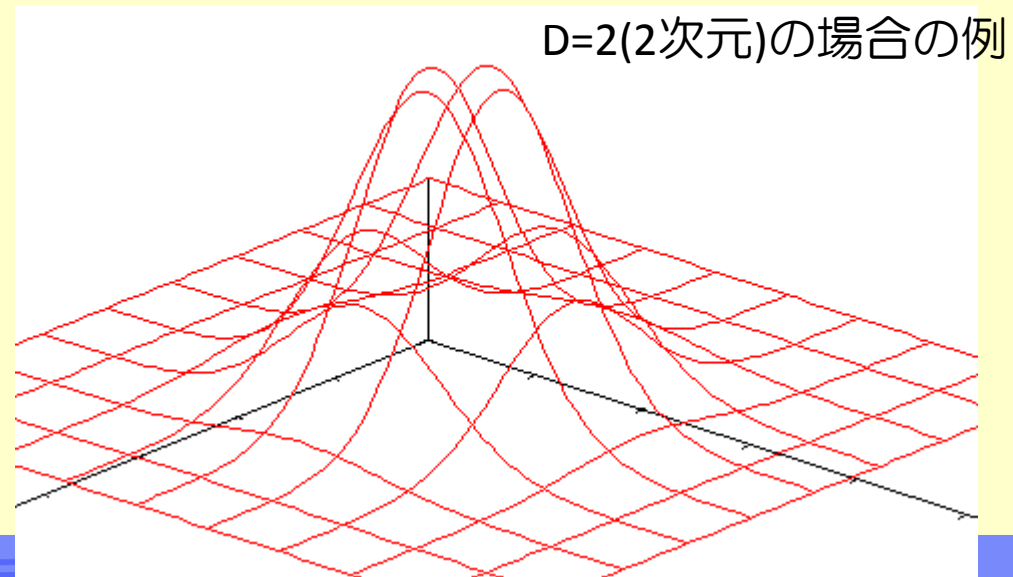
- $D$ 次元のデータの確率分布として、 $D$ 次元の多次元正規分布  $g(\mathbf{x})$  を考える

$$g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

– 1次元の正規分布  $g(x; \mu, \sigma) := \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$  の拡張

– パラメータ

- $\boldsymbol{\mu}$  : 平均 ( $D$ 次元)
  - $\boldsymbol{\Sigma}$  : 共分散行列 ( $D \times D$ )
- 単峰なので、表現力が不十分



# 教師なし学習モデルの典型：混合正規分布モデル

## 複数の正規分布によって複雑な分布を表現できます

- $K$  個の  $D$  次元正規分布  $g^{(1)}(\mathbf{x}), g^{(2)}(\mathbf{x}), \dots, g^{(K)}(\mathbf{x})$  の混合分布

$$P(\mathbf{x}) := \sum_{k=1}^K w^{(k)} g^{(k)}(\mathbf{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$$

$k$  番目の正規分布の  
重み

$k$  番目の正規分布

$$\text{ただし } \sum_{k=1}^K w^{(k)} = 1, w^{(k)} \geq 0$$

$$g^{(k)}(\mathbf{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}) := \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}^{(k)}|^{1/2}} \exp(-(\mathbf{x} - \boldsymbol{\mu}^{(k)}) \boldsymbol{\Sigma}^{(k)^{-1}} (\mathbf{x} - \boldsymbol{\mu}^{(k)}))$$

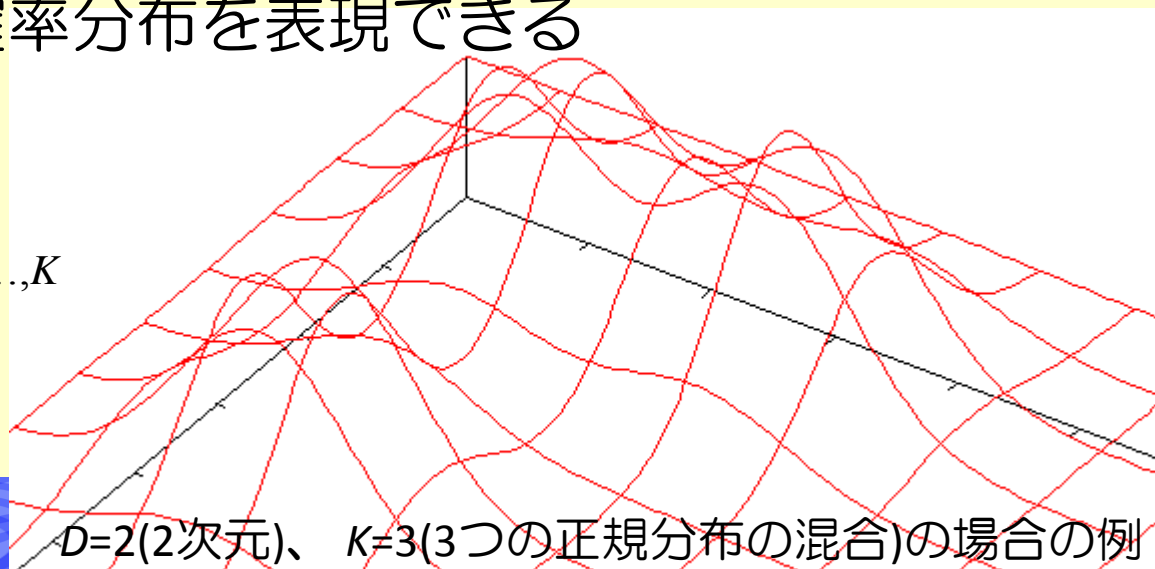
- 単一の正規分布より複雑な確率分布を表現できる

- モデルのパラメータは

- 混合比パラメータ  $\{w^{(k)}\}_{k=1, \dots, K}$

- 各正規分布のパラメータ

- $\{\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\}_{k=1, \dots, K}$



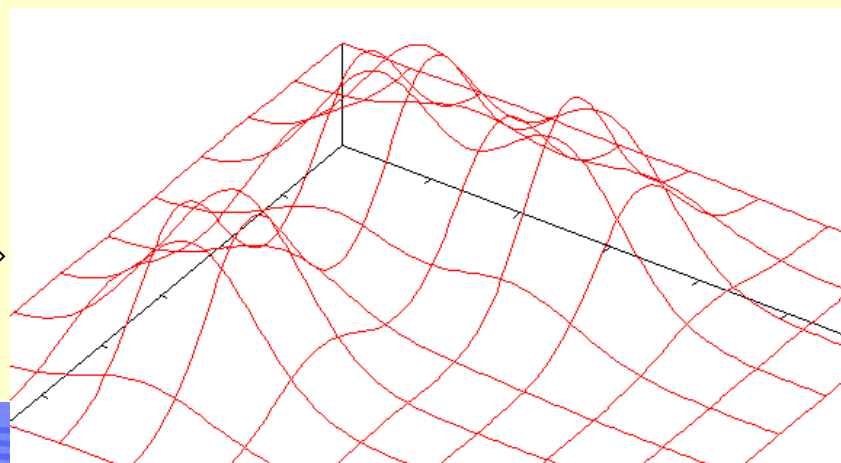
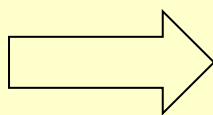
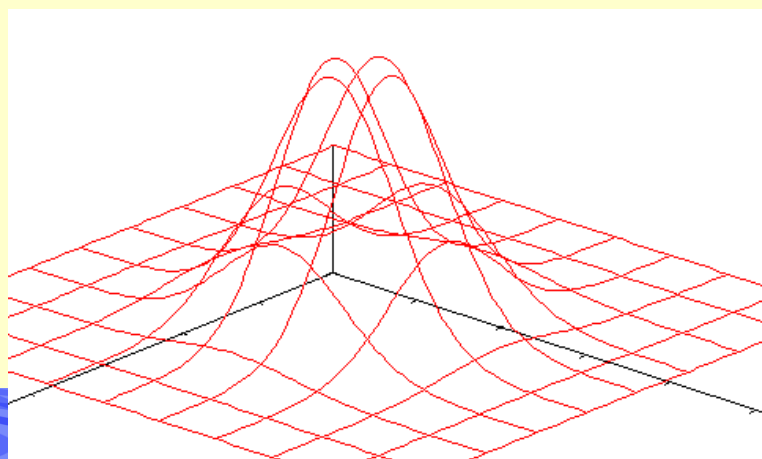
# 教師なし学習モデルの典型：混合正規分布モデル

## 2段階の生成過程として理解できます

$$P(\mathbf{x}) := \sum_{k=1}^K w^{(k)} g^{(k)}(\mathbf{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$$

■ データ $\mathbf{x}$ の生成過程を考えると

1. 確率  $(w^{(1)}, w^{(2)}, \dots, w^{(K)})$  (ただし  $\sum_{k=1}^K w^{(k)} = 1$ ) を使って、どの正規分布からデータを生成するか決める
2.  $k$  番目の正規分布  $g^{(k)}(\mathbf{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$  から $\mathbf{x}$ を生成する



# 教師付き学習モデルの典型：ロジスティック回帰モデル

## 2カテゴリ分類の標準的な線形モデルです

- 出力が2カテゴリの場合の代表的な条件付確率モデル

$$P(y = +1|\mathbf{x}; \mathbf{w}) := \sigma(\mathbf{w}^\top \mathbf{x}) = \sigma(w_1x_1 + w_2x_2 + \cdots + w_Dx_D)$$

$$P(y = -1|\mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^\top \mathbf{x})$$

- なお、 $\mathbf{w}$  はモデルを定めるパラメータベクトル

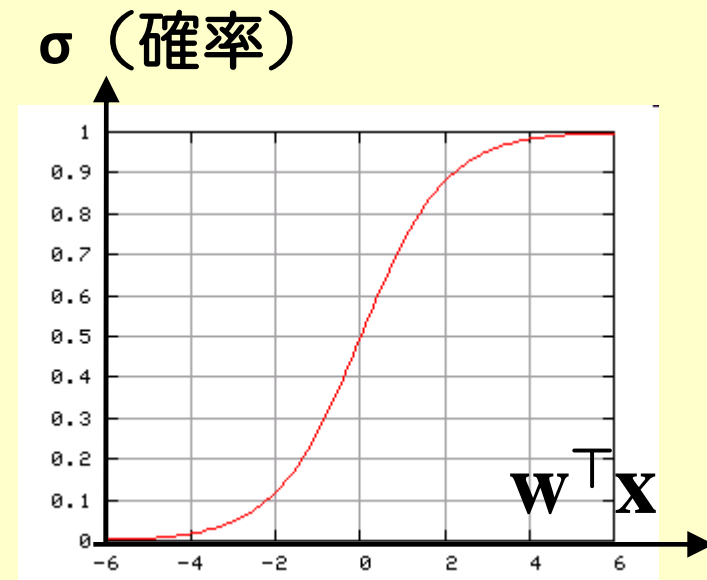
- $\mathbf{w}$  の各次元は  $\mathbf{x}$  の各次元の  $P(y = +1|\mathbf{x}; \mathbf{w})$  への寄与度

$$\mathbf{w} := \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} \quad \mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix}$$

- $\sigma$  はロジスティック関数

- 連続値を確率値  $[0, 1]$  にマップする

$$\sigma(a) := \frac{1}{1 + e^{-a}}$$

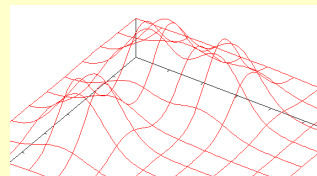


# ここまでのまとめ： 機械学習の2つの問題設定と代表的モデル

## ■ 機械学習の代表的なタスクは2つある

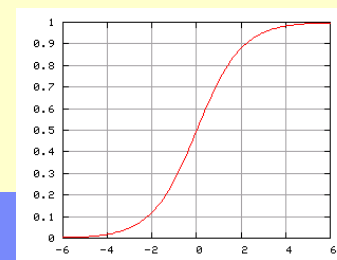
### — 教師無し学習

- 入力を見ているうちに、どんなものが現れやすいかなどのパターンが分かってくる
- 実際には、入力の確率分布の推定問題として扱われる
- 代表的なモデル：混合正規分布



### — 教師付き学習

- 入力に対する出力を試行錯誤するうちに、どういう入力のおきにどういう出力をすればよいか分かってくる
- 実際には、入力を与えられたときの出力の条件付確率分布の推定問題として扱われる
- 代表的なモデル：ロジスティック回帰





---

## 機械学習の応用

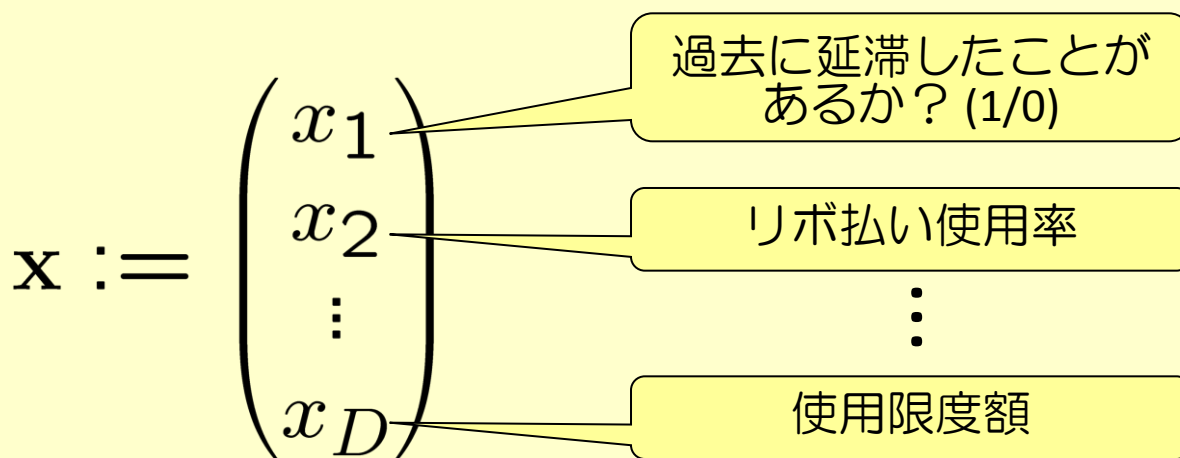
# 機械学習の応用： 実際、いろいろ役に立ちます

---

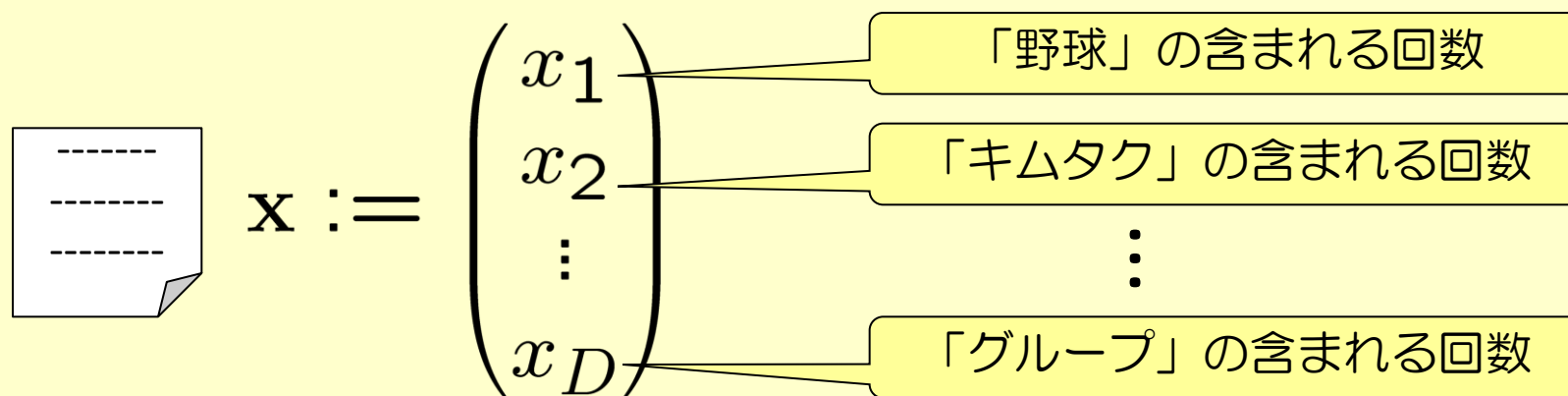
- 応用
  - 信用リスク評価（教師付き学習）
  - テキスト分類（教師付き学習）
  - 画像認識（教師付き学習）
  - 異常検知（教師無し学習）
  - クラスタリング（教師無し学習）

## 「この人にお金貸して、返ってくるんだろうか？」

- ある顧客に、融資を行ってよいか
  - 顧客  $x$  を、さまざまな特徴を並べたベクトルで表現
  - 融資を行ってよいか  $y$ 
    - 融資を行ってよい（返済してくれる）：+1
    - 融資してはいけない（貸し倒れる）：-1
  - マーケティングの文脈では、買ってくれる(+1),買ってくれない(-1)



- 自然言語の文書が、あるカテゴリーに入るかどうか
  - 文書  $x$  を、含まれる単語ベクトルで表現
  - (たとえば) ある事柄に好意的かどうか  $y$ 
    - 好意的：+1
    - 否定的：-1
  - トピック  $y$ ：「スポーツ」「政治」「経済」... (多クラス分類)

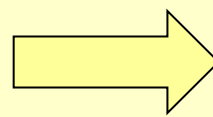


bag-of-words 表現

# 教師付き学習の応用例：画像認識

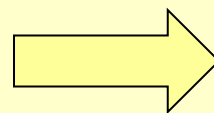
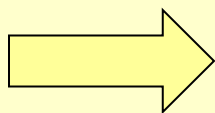
「これ、何て書いてあるの？」 「いま何考えてる？」

## ■ 手書き文字認識



{ ある文字か(+1)否か(-1)  
どの文字か? {"0","1","2",...}

## ■ BCI (Brain Computer Interface)



右(+1)? 左(-1)?

## ■ ほか、顔画像認識や、動画認識

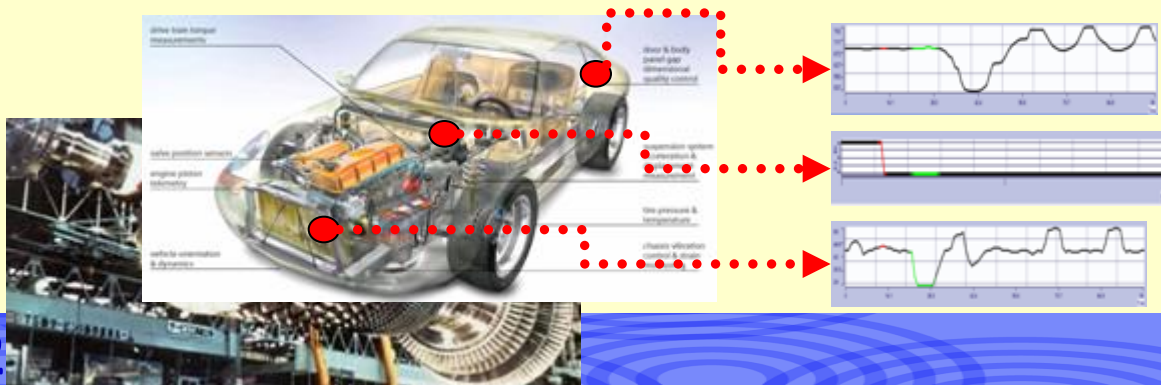
# 教師なし学習の応用例：異常検知

「ちょっと出かけてくるけど、ヤバそうだったら教えて」

- 機械システム／コンピュータシステムの異常を、なるべく早く検知したい
  - 早い段階で検出できれば、それだけコスト減
- システムに仕込まれたセンサーからの取得データを分析する
  - システムの異常／変化、不正な操作により、システムが通常とは異なった振る舞いをするようになる
  - 計測機器の異常によって、通常とは異なった計測値が得られるようになる

機械／プラントなど

センサーからの  
取得データ

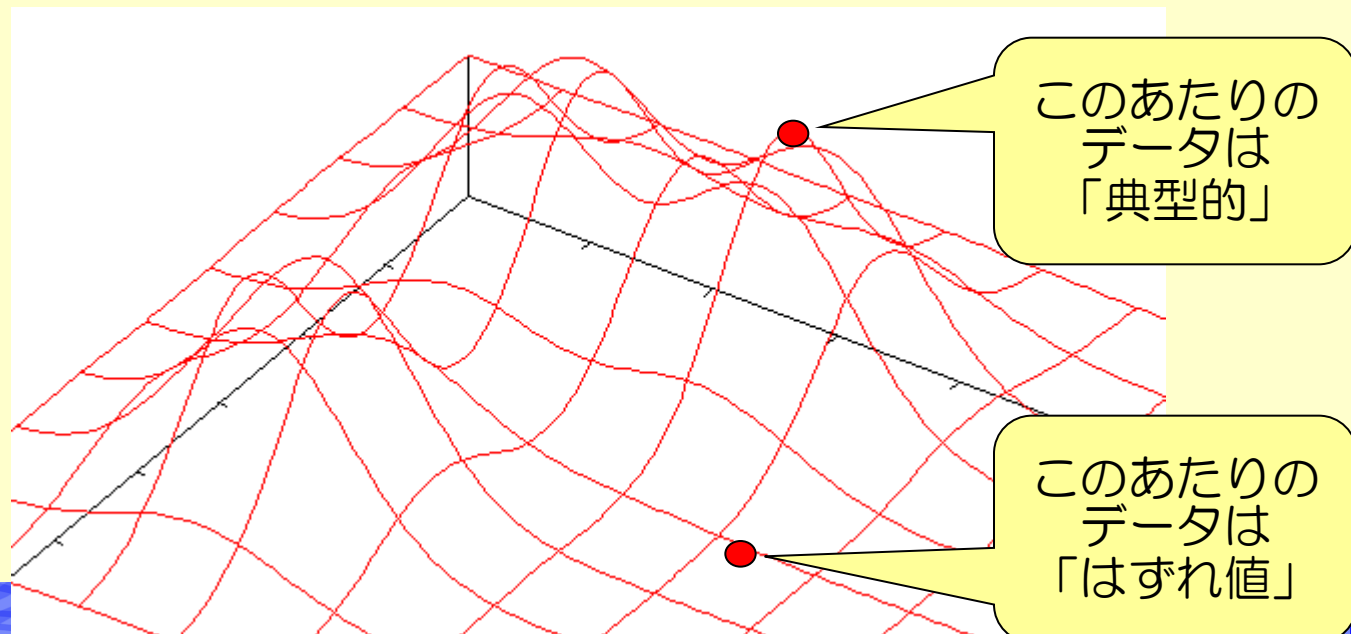


異常  
(の前触れ)

# 教師なし学習の応用例：異常検知

## 確率の低いデータ＝異常と考えます

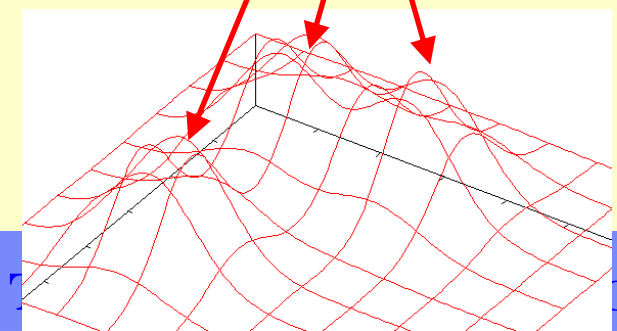
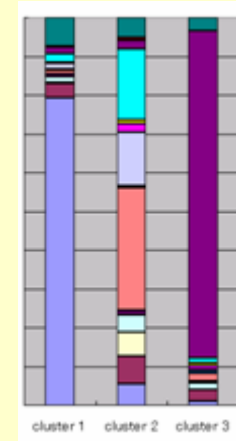
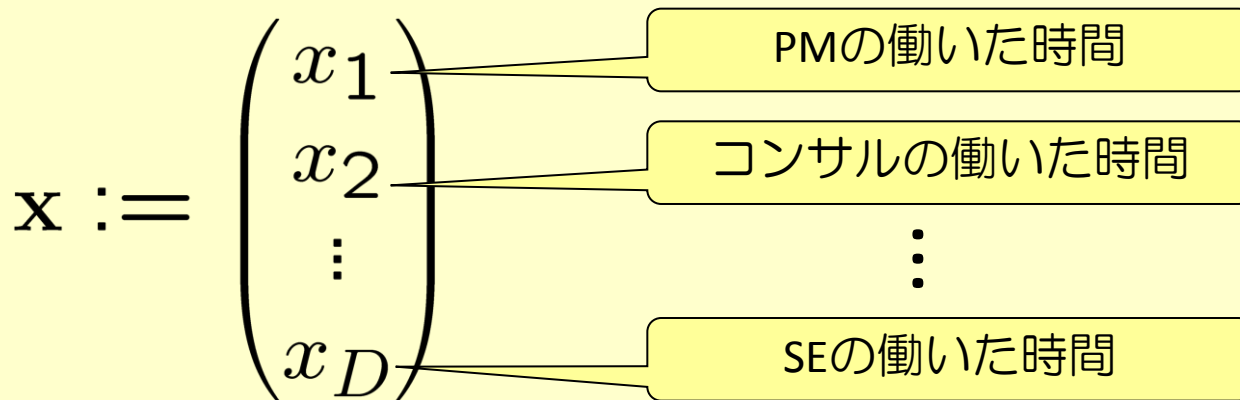
- システムの状態をベクトル  $\mathbf{x}$  で表現し、教師無し学習による確率分布  $P(\mathbf{x})$  の推定を行う
  - コンピュータ間の通信量、各コマンドやメッセージ頻度
  - 各センサーの計測値の平均、分散、センサー同士の相関
- $P(\mathbf{x})$  の小さいデータ  $\mathbf{x}$  は「めったに起こらない状態」＝システム異常、不正操作、計測機器故障などの可能性がある



# 教師なし学習の応用例：クラスタリング（人材管理での例）

## 「とりあえず、同じような塊に分けて整理しといて」

- プロジェクトには様々な職種の人間が様々な配分でかかわる
  - プロジェクト・マネージャ、コンサルタント、ソフトウェア・エンジニア、アーキテクト、...
- 実際のプロジェクトで使われた人的リソース配分を $\mathbf{x}$ として、混合分布による教師無し学習を行う
- 混合分布の各分布の中心が典型的な人材配置のテンプレートを作成
- クラスタリング：グループを発見する





## ここまでのまとめ：機械学習にはさまざまな応用がある

- 紹介した応用：信用リスク評価、テキスト分類、画像認識、異常検知、クラスタリング
- 紹介しなかった応用
  - 推薦システム
  - ユーザーモデリング
  - 需要予測（ $y$  が実数値）
- データあるところには、学習の問題がほぼ確実にある
  - 教師付き学習では1%の予測性能改善が、収益に直結する
  - 異常検出の需要は、コストのかかるシステムを抱える組織ならば常に存在する
- まだまだビジネスの現場において、機械学習（先進的なBI）が十分に入り込んでいない