Kumamoto University

SVMを用いた移動滞在判定と今後の計画について

熊本大学
交通政策分析研究室

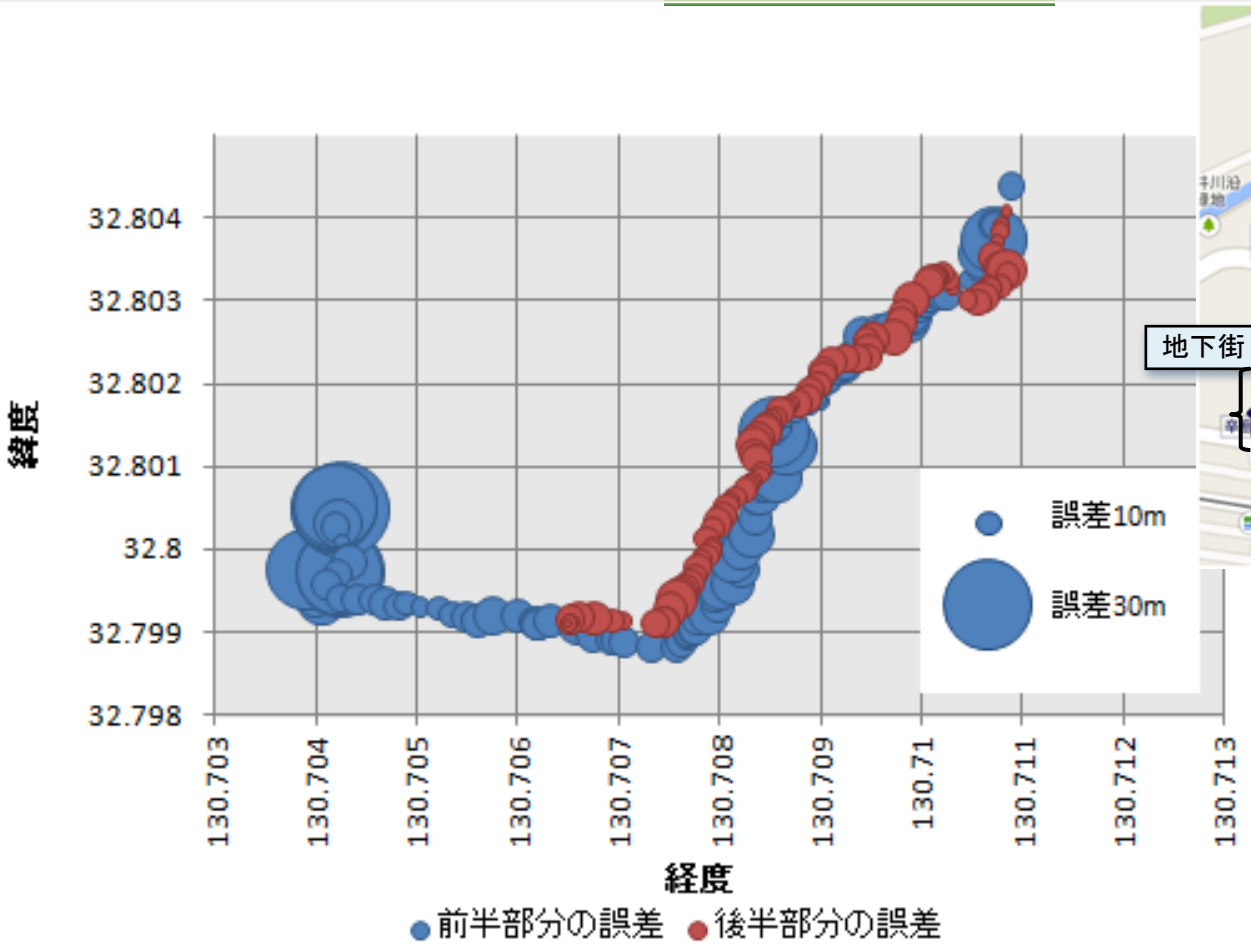
Ｍ２　井村

背景

- PT調査の補完として、熊本中心市街地のスマホ回遊行動調査
- 熊本市の再開発事業

  まちなかにおける回遊性の向上

背景



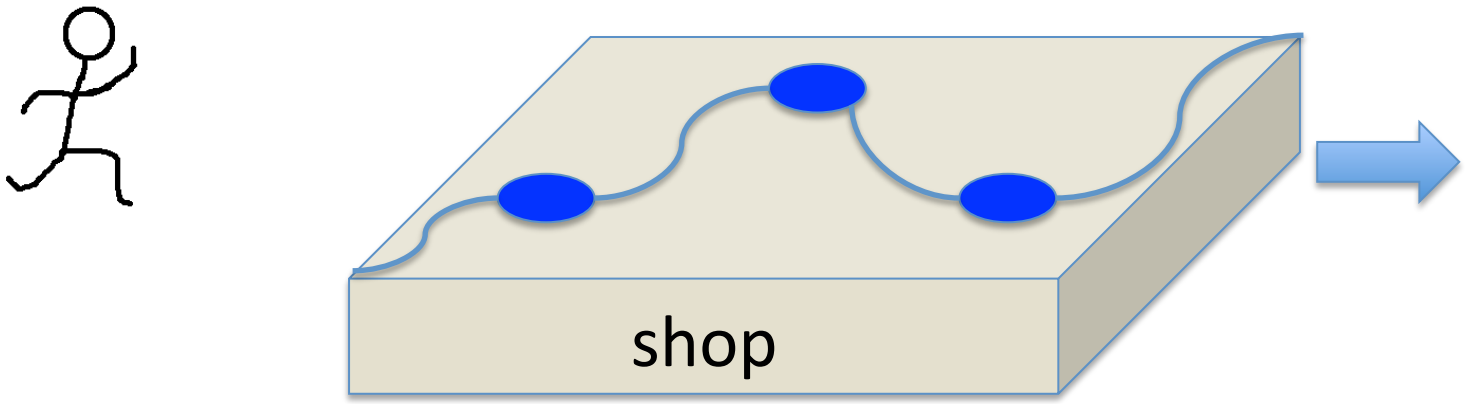
アーケード・地下での誤差大

背景

建物内、路地での行動が不明

背景

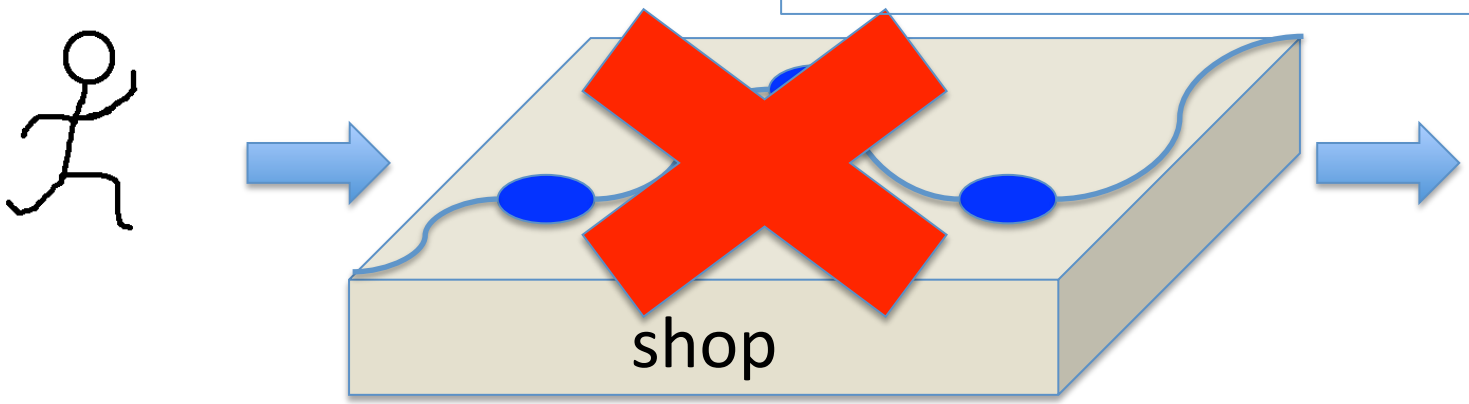
建物（店舗内）での回遊例



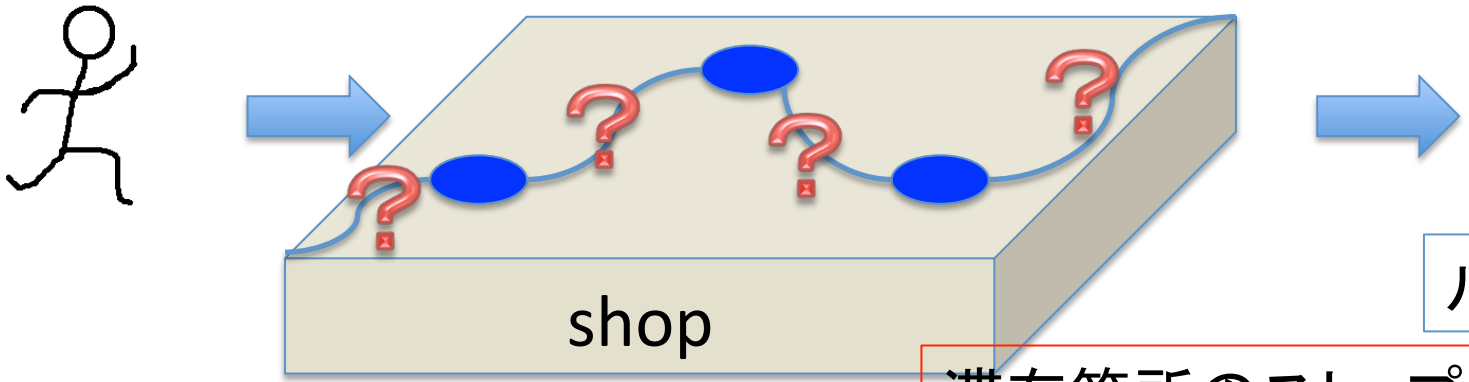
背景

従来調査

ルート、店舗内滞在ポイント数不明



PP調査から移動滞在判定



ルート不明

滞在箇所のストップポイント数として判断可能？

目的

- 判定に使うSVMについての理解を深める
- 加速度データから移動滞在の判別
  路地や店内での状態を大まかに把握
- 滞在判定から分かることを回遊モデルへ

SVM（サポートベクターマシーン）

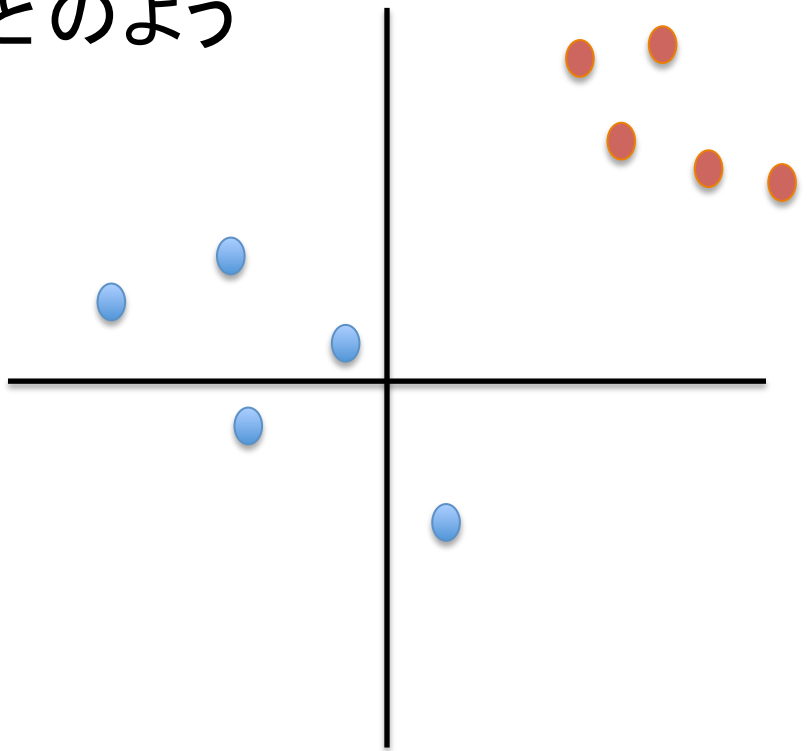
- 1995年にV.Vapnikによって統計的学習理論の枠組みで提案された学習機械

- 学習データの中の，サポートベクトル (識別境界近傍に位置する学習データ) と識別境界との距離であるマージンを最大化するように識別境界を構築し 2 クラス分類を行う

2クラス分類問題

SVMはパターン認識手法として用いられる
他のアルゴリズムと比べてどう違うのか？

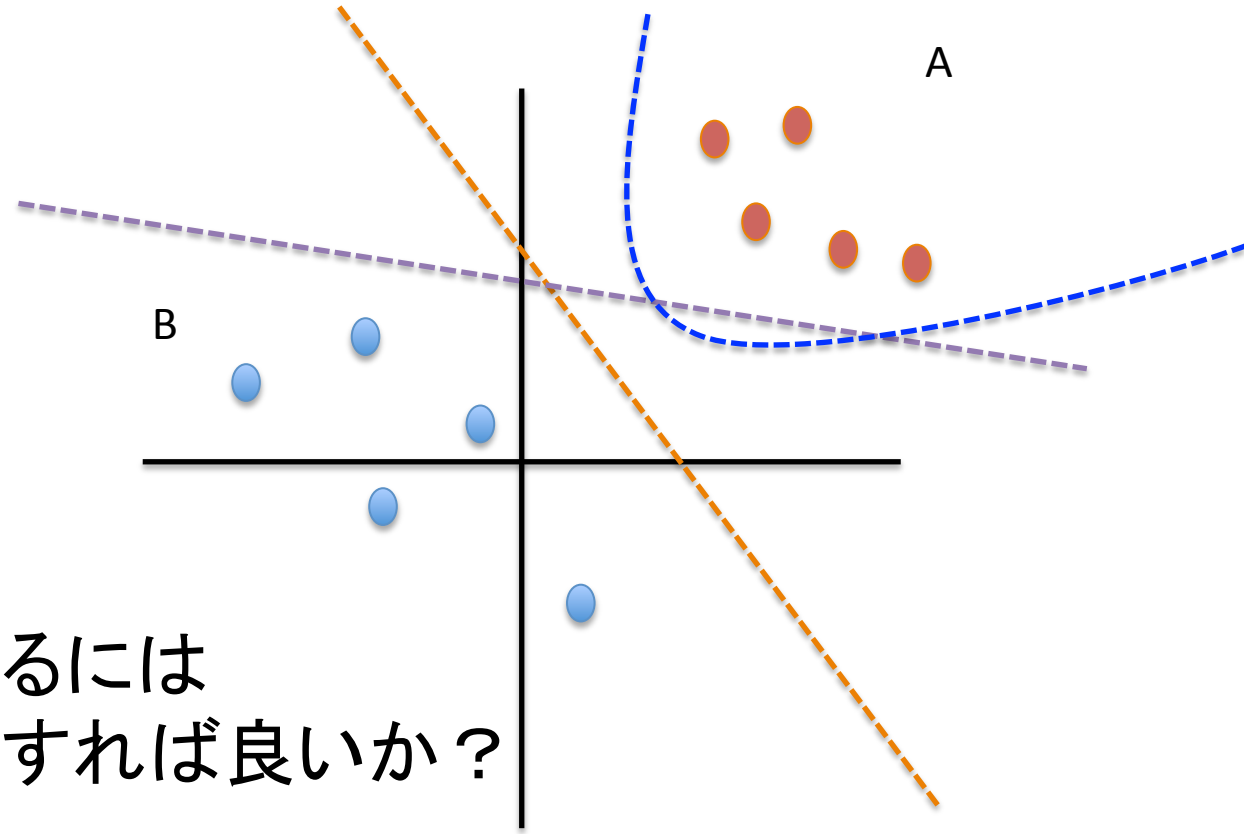
右の図のような2次元空間があった際に
2つのグループを識別するとしてどのよう
識別線を引くと良いかという問題



2クラス分類

識別線として考えられるものは無数に存在する

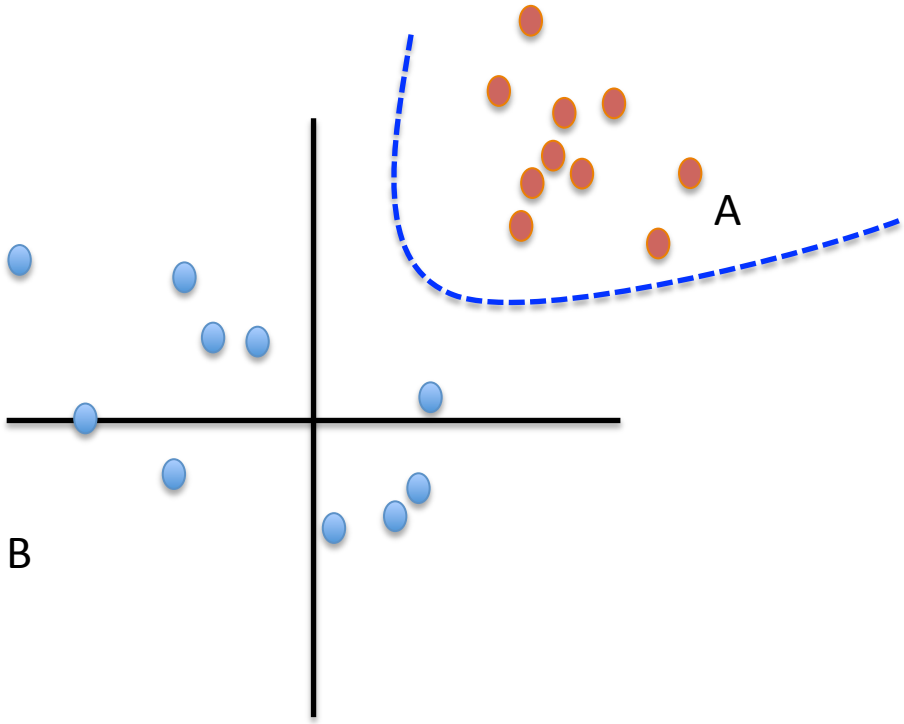
例えば…



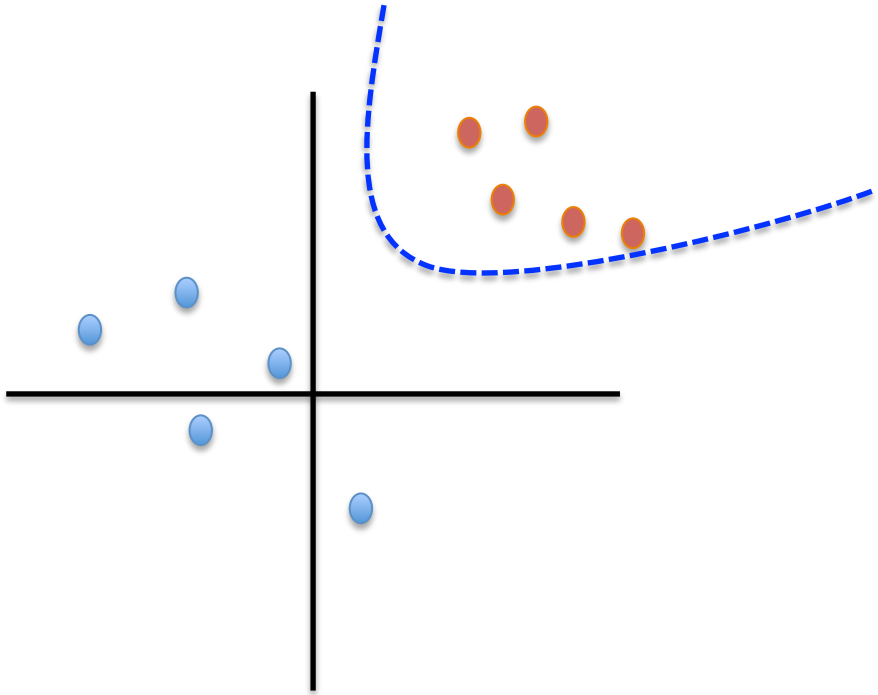
パターン認識するには
何を持って判断すれば良いか？

2クラス分類

バックプロパゲーション学習（従来手法？）だと…



教師データ

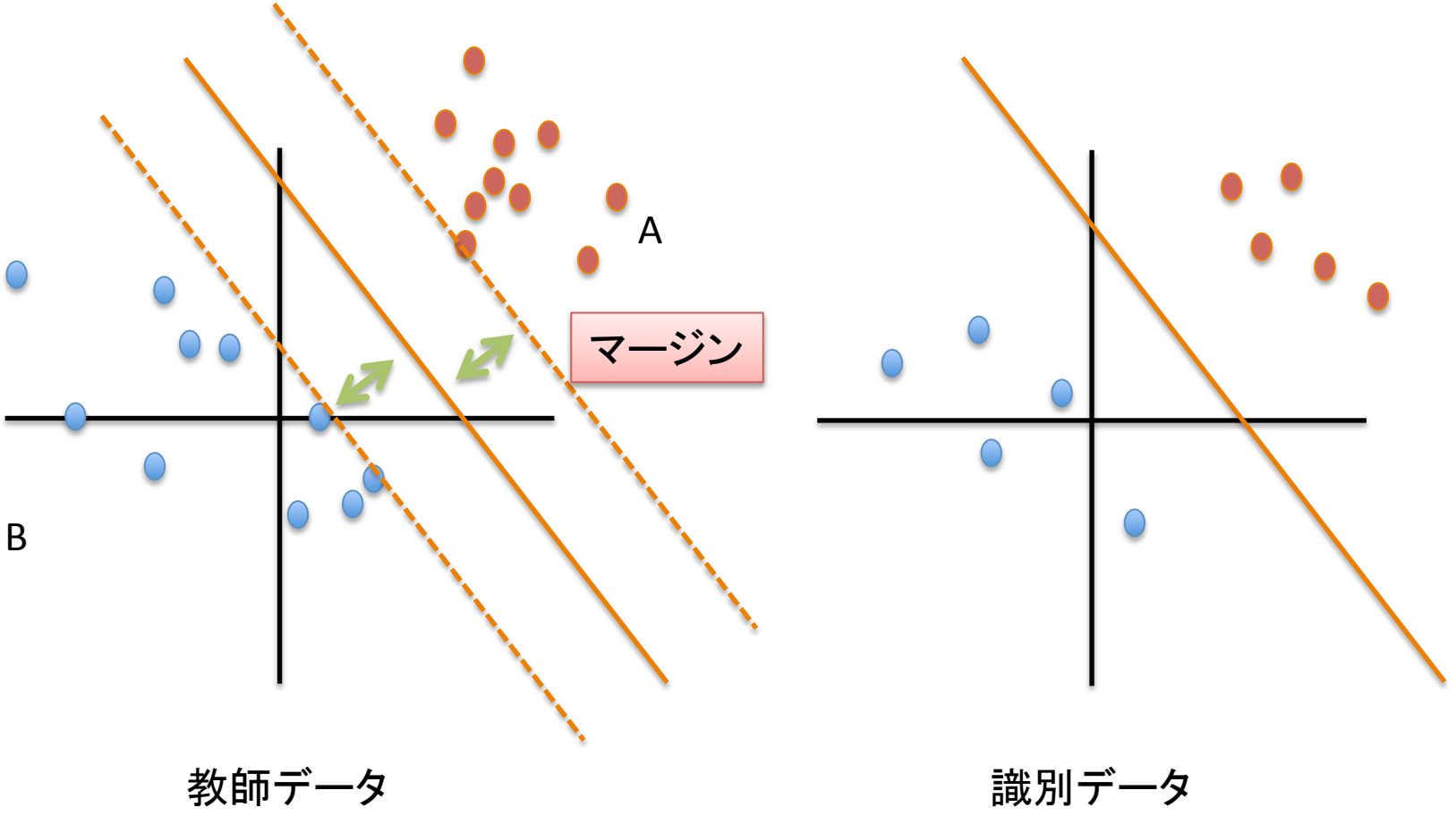


識別データ

学習した教師データに依存した結果になる　（パラメトリック）

2クラス分類

一方SVMを用いた場合



学習した教師データから他クラスに一番近いものを基準（ノンパラメトリック）

超平面集合

- SVMの中核：超平面識別関数

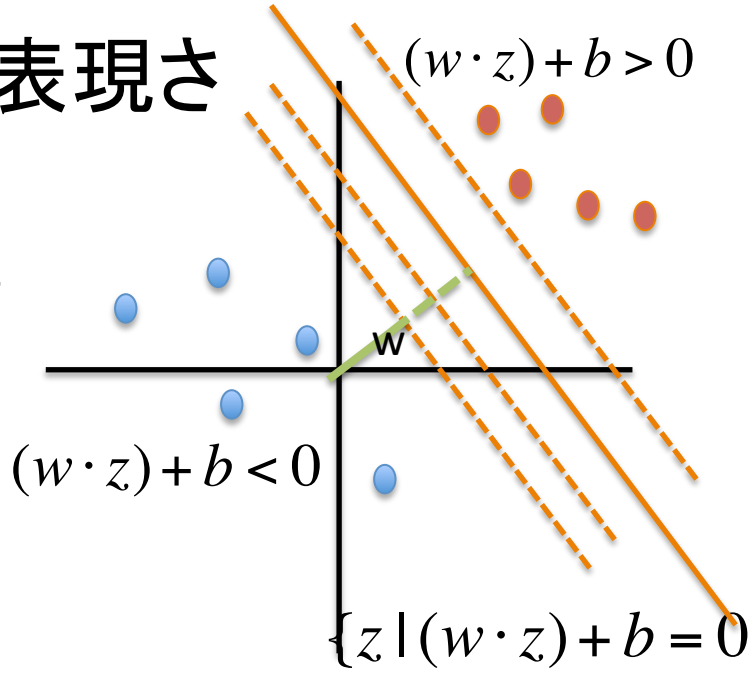
- 内積空間F及びパターンベクトル集合 $z_1,\ldots,z_r$ が与えられたとすると任意の超平面識別関数は次のように表現される

$$\{z \in F : (w \cdot z) + b = 0\}$$

F： 内積空間

w： ベクトルパラメータ

b： バイアスパラメータ



超平面集合

- 制約を加えることで識別関数となる

超平面を $(w,b) \in F \times R$ を有する関数

に一意に決める

$$\min_{i=1,\cdots,r} |(w \cdot z_i) + b| = 1$$

この制約によって w, b は距離

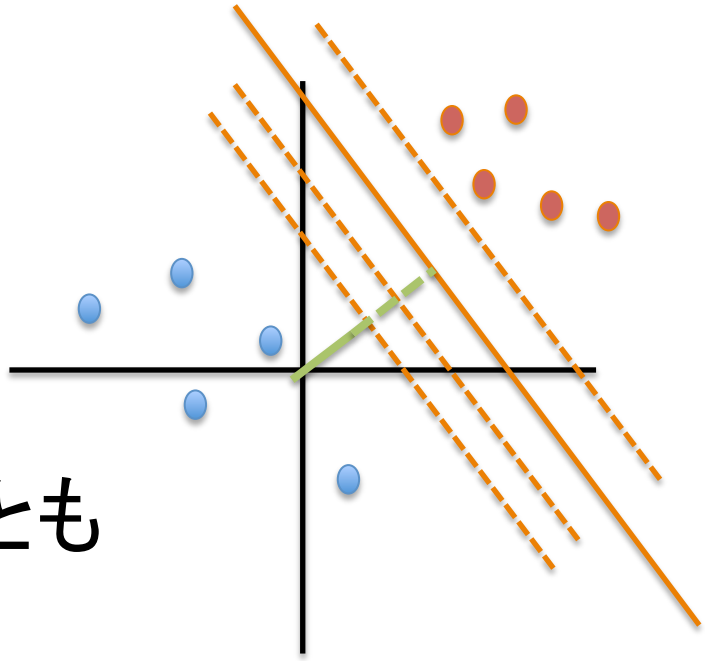
$1/\|w\|$ を持つ超平面に接近

するデータを表す

従って2クラス分類問題の場合

垂直に測ったマージンは少なくとも

$2/\|w\|$ となる



命題

- Rを点$\{z_1,\ldots,z_r\}$を含む最も小さい球体

$$B_R = (a) = \{z_1 \in F : \| z - a \| < R\}(a \in F)$$

の半径とし次式がこれらの点を定義する識別関数であるとする

$$f_{w,b} = \mathrm{sgn}((w \cdot z) + b)$$

そのとき関数集合

$$\{f_{w,b} : \| w \| \leq A\}$$

は次式を満たすVC-次元hを有する

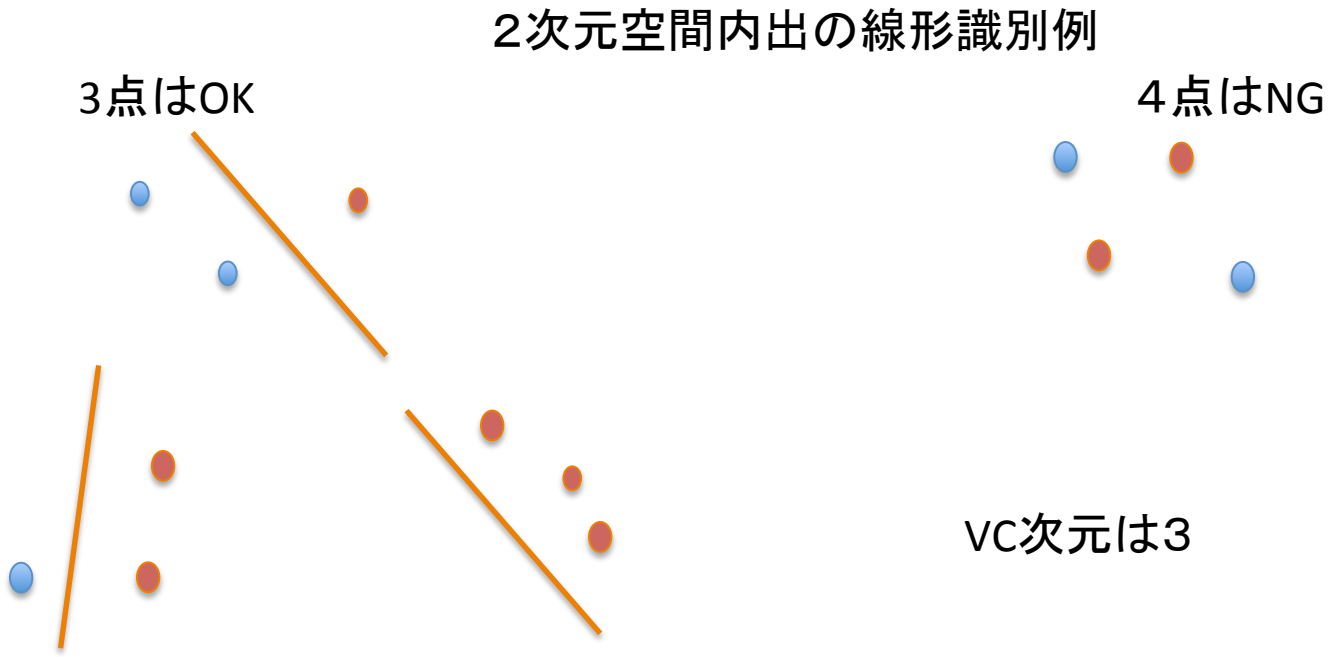
$$h < R^2 A^2 + 1$$

命題の解釈

- 条件 $\|w\| \leq A$ を省くとVC-次元が N_F+1 となる関数集合を導くことが出来る。N_F は空間 F を表す。
  よって $\|w\| \leq A$ によって N_F より小さなVC-次元を得ることが可能になり、高次元空間で識別問題を取り扱うことができる。

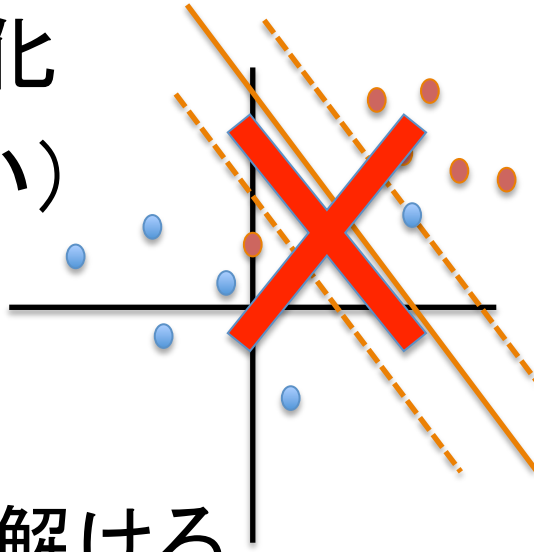
VC次元

- VC(Vapnik-Chervonenkis)次元：任意の2値ラベルを付与しても正しく分離できるデータ点の最大数
- N次元のときVC次元はN+1



ハードマージン

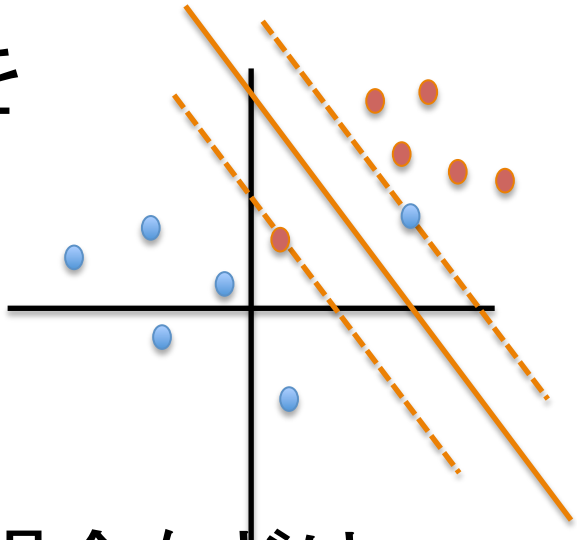
線形制約下での凸な2次関数の最適化
（制約条件を満たさないものを許さない）



- ローカルミニマムの問題がない
（局所最適＝大域的最適）
- 計算パッケージなどで比較的容易に解ける
- このままでも解くことはできるが，双対問題に変換すると，制約の部分がより簡単な式で表される

ソフトマージン

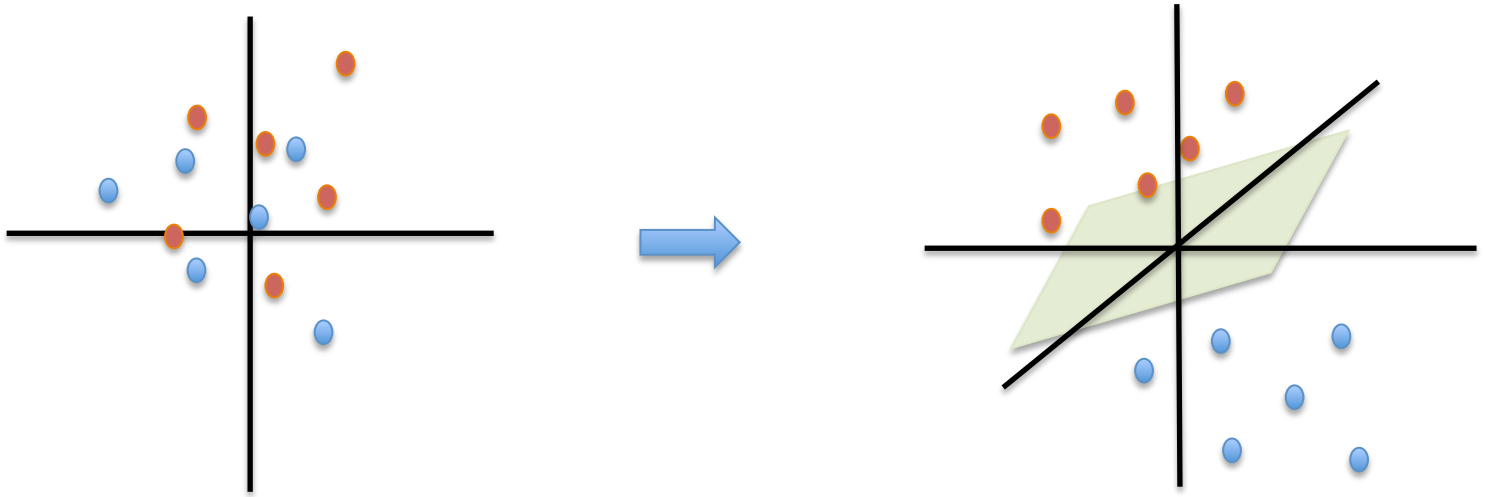
- 訓練データが線形分離可能でないとき，
HSVMは解を持たない少しの誤差ξ_iを許す
- ペナルティを与えることで，データを
上手く分けられない場合でも“ほぼ”
分けられるように超平面を定める．



- 線形分離な場合でもノイズがある場合などは
無理に線形分離な超平面を求めないほうが
よい

線形識別できない場合

線形で分離できない場合…



高次元特徴空間へ写像（カーネルトリック）

特徴空間で線形SVMを行う

カーネル関数

$$\phi : x_i \mapsto z_i \qquad (\phi(x) \cdot \phi(x_i)) \qquad (\phi(x) \cdot \phi(x_i)) = k(x, x_i)$$

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} y_i \alpha_i \cdot k(x, x_i) + b)$$

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(x_i \cdot x_j)$$

$$subject \ \ to \qquad 0 \le \alpha_i \ge \gamma \ \ , i = 1, \ldots, l,$$

$$\sum_{i=1}^{l} a_i y_i = 0$$

SVMを用いるメリット・デメリット

- メリット
- データ特徴量の次元が増加しても識別精度がよく,最適化するパラメータが少ない
- 最適化すべき パラメータが少ないので試行回数が少なく最適パラメータを求められる

デメリット

- 学習データが増えると計算量が膨大
- 基本的には 2 クラスの識別手法で 最適パラメータが異なり,多クラスを考慮に入れた識別関数の最適化・超平面作成は困難

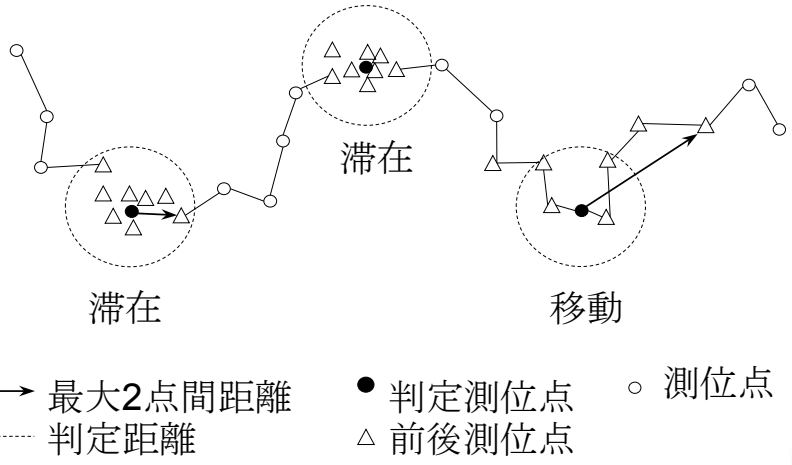
真値調査データを元に-SVMの練習-

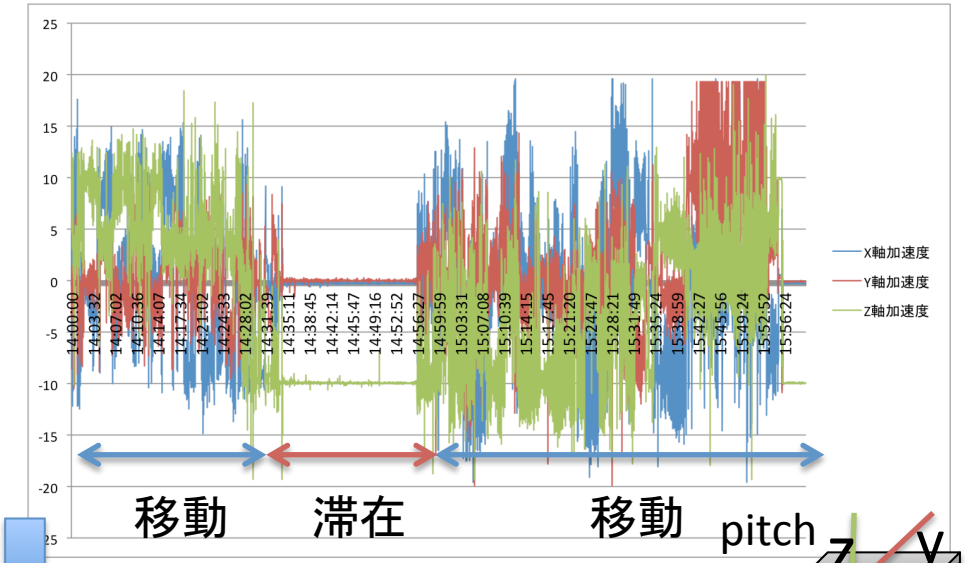
- 今泉さんの卒論時のSVMのコードを元に実際に回してみることを目的として実験的に

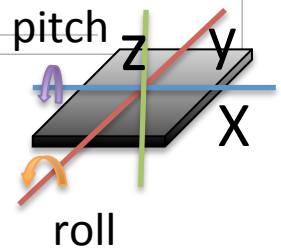
- 教師データ　真値調査時　下通り
- 判別データ　真値調査時　交通センター

- 暫定的に休憩時間を滞在、その他を移動と定義し計算

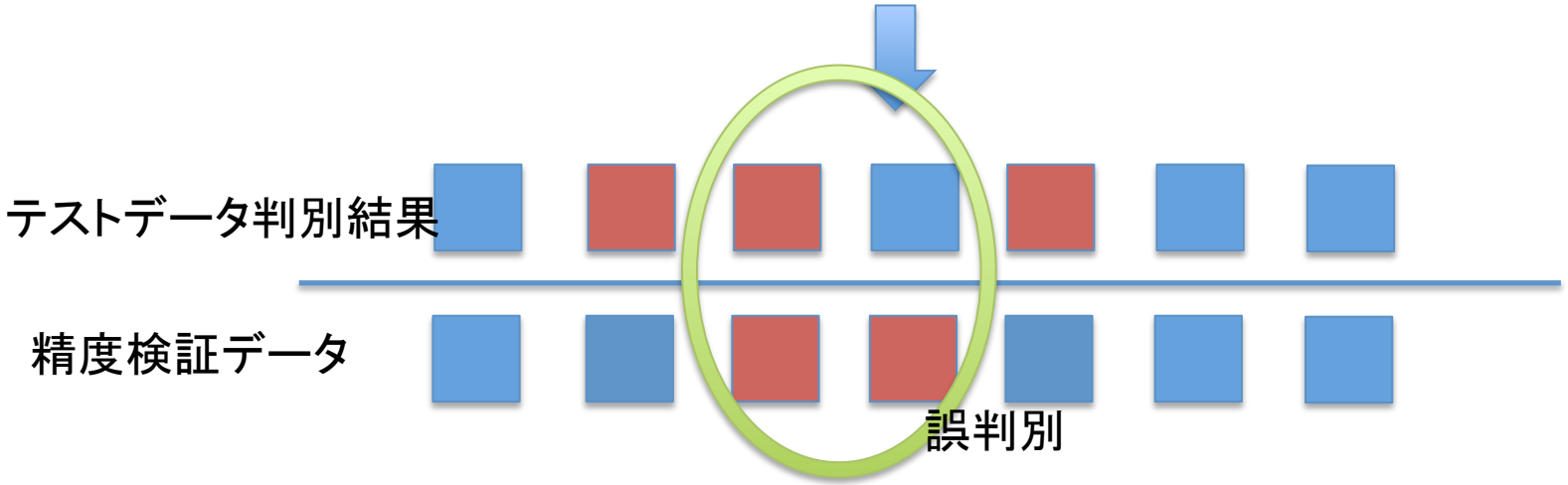
SVMを用いたPPデータ判別のイメージ





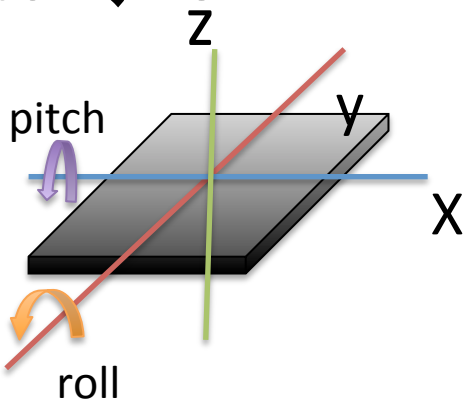


SVM



PPデータからSVM

- 今回判定に用いたパタメータとしてmms、x軸加速度、y軸加速度、z軸加速度、pitch、roll
- 合成ベクトル等や振幅などの算出は行わずにそのままのデータとしてSVMで判別



判定に用いたデータ数

場所	生データ数	移動	滞在	編集データ	移動	滞在
下通	32131	25760	6371	5356	4293	1062
交通センター	35994	30221	5773	—	—	—

判定結果

教師データ（5356サンプル）

	移動	滞在
移動	29058	1163
滞在	3263	2511

c	適合率(%)
移動	96.1
滞在	43.4

教師データ（32131サンプル）

	移動	滞在
移動	29361	860
滞在	3168	2605

	適合率(%)
移動	97.2
滞在	45.1

滞在のデータの精度が良くない

適合率向上に向けて

- 今回はコードの理解をメインに行い、コードを回すことに重点置いたためこのような結果？

- 滞在に関してはセンサーの反応が繊細であるため？位置情報とも組み合わせたデータ整理
- 加速度情報データの間引き方の検討
- 教師データサンプルを正確な状態で取得する
  移動、回遊中の状態等まで含めて再調査
- 教師データで判別する加速度データの扱い方の検討（参考　大村さん修論）
- 機械学習Adaboostと組み合わせる
  （参考　今泉さん論文）

熊本での再調査（真値調査）

- 真値調査としてまたデータを新たに計測するとの噂（詳細は不明…）

- 調査追加項目案

GPS・ルート等以外に階段やエレベータ、エスカレーター等の分類、店の中での動きを記録ウインドウショッピング、立ち止まってみた回数等？

判別するときの分類について

- 移動・滞在の２種類に絞らず増やしてみる
  早い速度での歩行、ゆっくりした歩行など？

- 階段、エレベータ、エスカレーターなどでの違い

- 携帯を操作していることが判別できれば…
  今よりも取得間隔短く、データ量増える

加速度情報を組み合わせた回遊のモデル？

- モデル考えるにあたって、どういう項目として加速度情報を組み込んでいくか？
- ストップ数の多さから、早い歩きなどから疲労の指標として
- 携帯を操作しているポイントがどこか？
  景観的（魅力が無い）、地理的（地図）

参考文献

- 杉山 将　東京工業大学

2011　サポートベクターマシンとブースティングの学習理論

- 赤穂昭太郎　産業技術総合研究所

2006　統数研公開講座「カーネル法の最前線－SVM, 非線形データ解析,構造化データ－」

- 小野田崇

2001　オペレーションリサーチ　サポートベクターマシーンの概要

メモ

- 店舗内でのトリップが多いほど帰りやすくなる？

  →生存時間？疲労度指標？

- 自分の持ってるデータでも書いてみる

- 同行者にも関連した