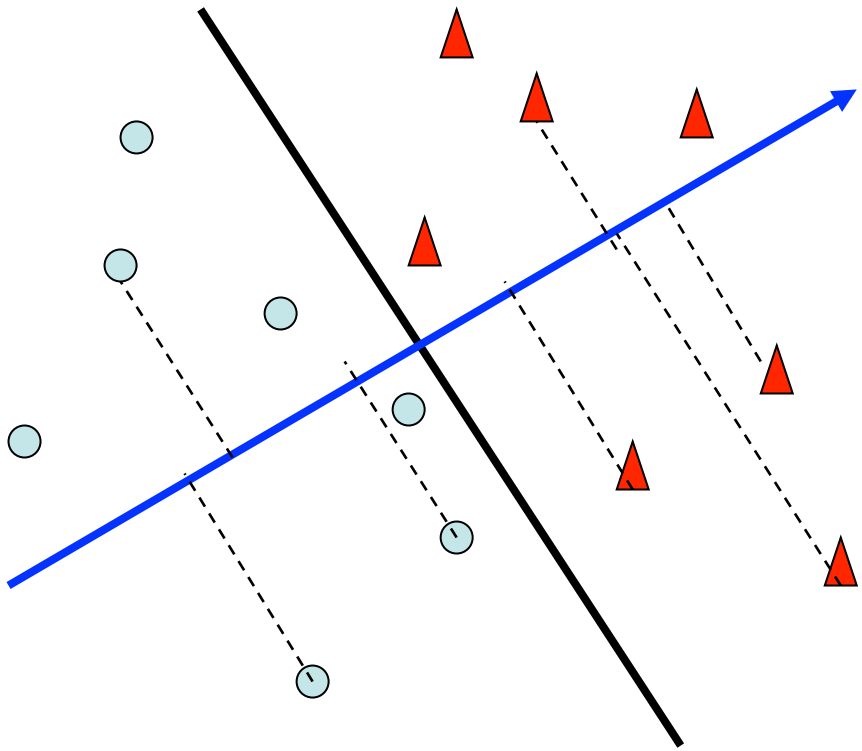
パターン認識と学習

竹内孔一

本日の内容

- 識別関数の設計
  - 2クラス分類を考え1次元空間に変換
  - 学習データは正規分布を仮定

線形識別関数を決める



w

識別面の法線

結局，この法線上に全ての事例を射影して判別

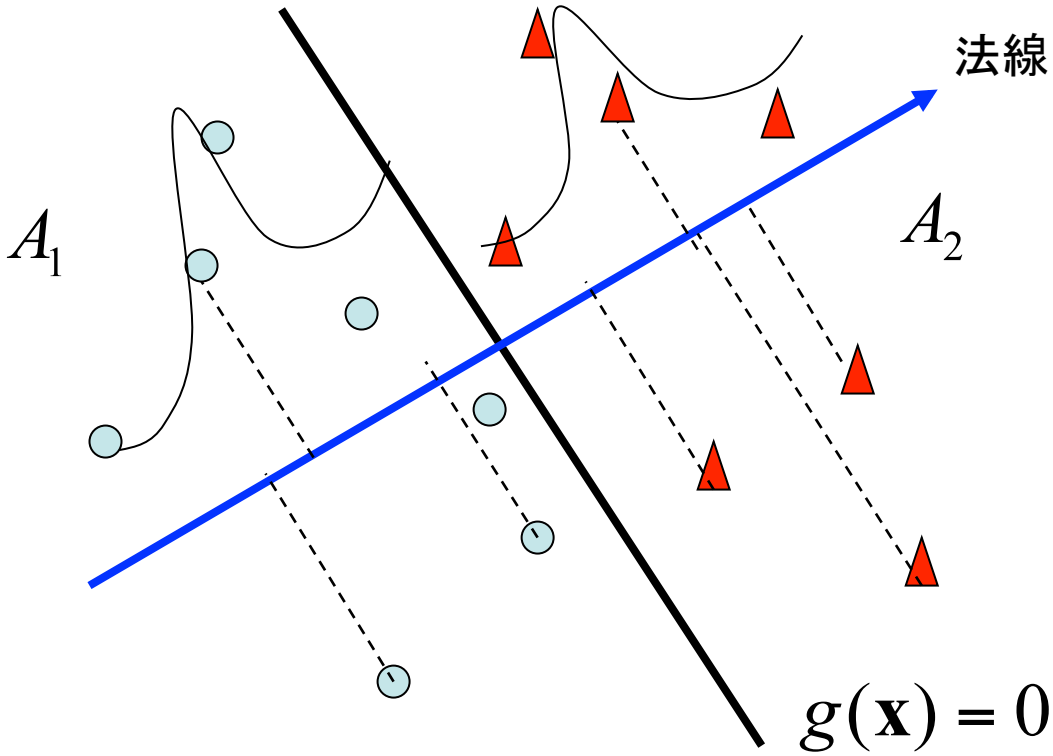
つまりd次元空間を1次元空間に変換して判別

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + \mathbf{w}_0 = 0$$ 識別平面

識別関数を決める

- 方針
  - 法線上での学習データの分布に注目
  - 平均と分散で評価関数が決まるとする

  → 重みｗと　平均，分散との関係を求める

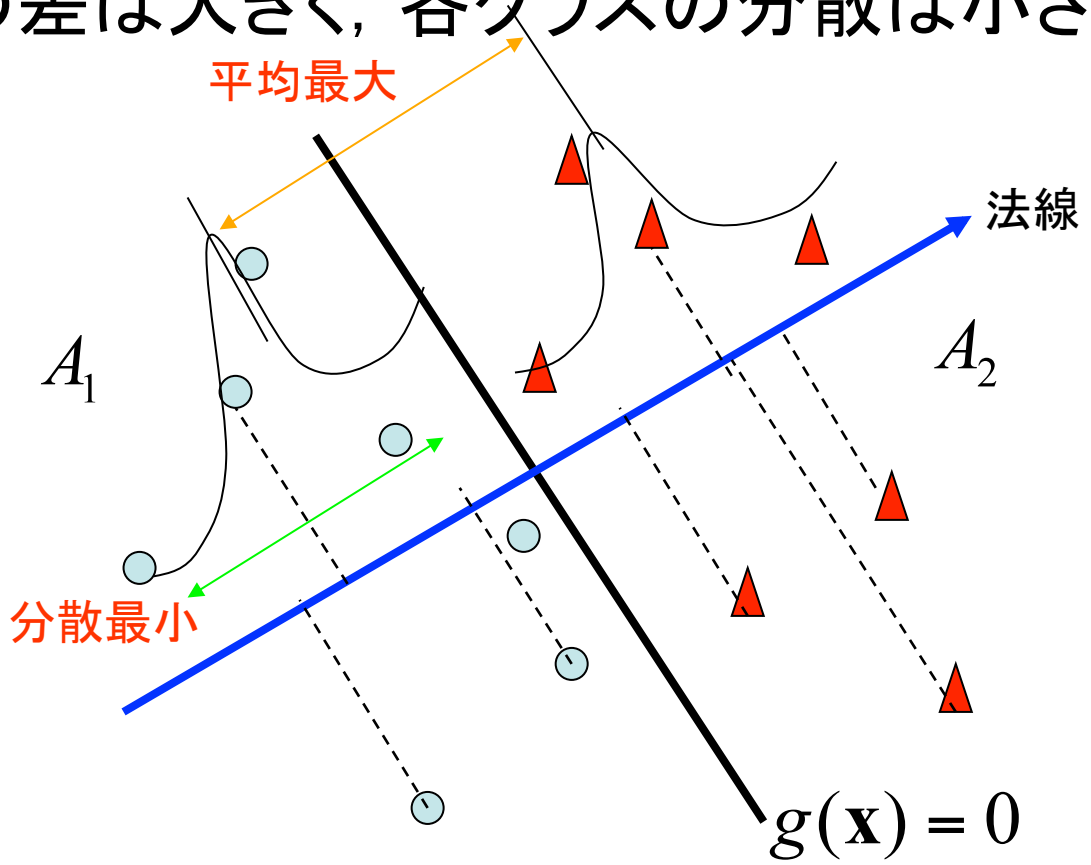


各学習データは
正規分布に従って
出力されている
と仮定

識別関数を決める

- 識別関数決定の方針
  - 重みwを決定するには

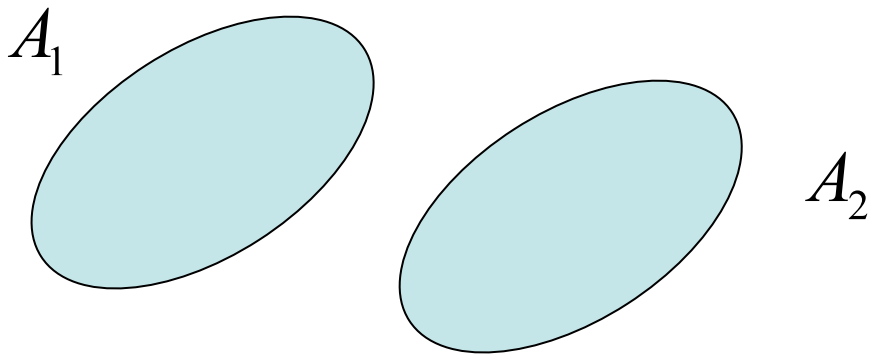
    平均の差は大きく，各クラスの分散は小さく



例題

- 2クラス分類で線形識別関数を求めたい
  - 下記のようなときどこに識別関数をおけばよいか?
  - その置き方を示唆するアイデアを答えよ

    (学習データは各クラスの正規分布に依存するとする)



特徴空間

整理

- 学習データが正規分布に従ってると仮定される場合に，線形識別関数を求めるアイデアについて述べよ

  回答: 各クラスの平均と分散によって評価関数が定義されると考える

- 2クラス分布の場合，線形識別関数を仮定すると何次元の特徴空間に変換されることになるか

  回答:1次元

一般識別関数

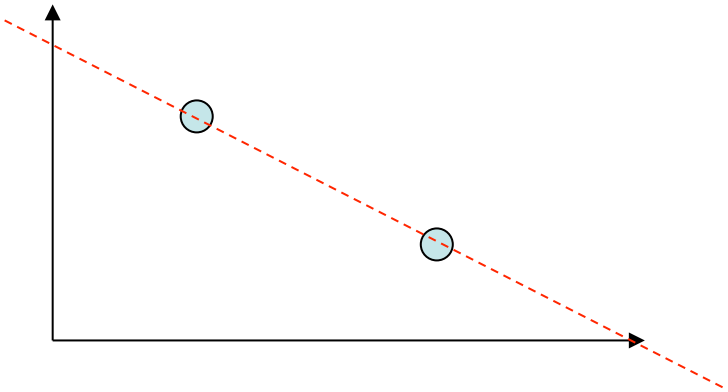
- 高次の識別関数
- 複数の関数の線形和(e個)

  → e個の特徴空間に対する線形識別関数と同じ

- 現実
  - 必要な学習データが用意できずover fittingを起こしやすい

学習データの数

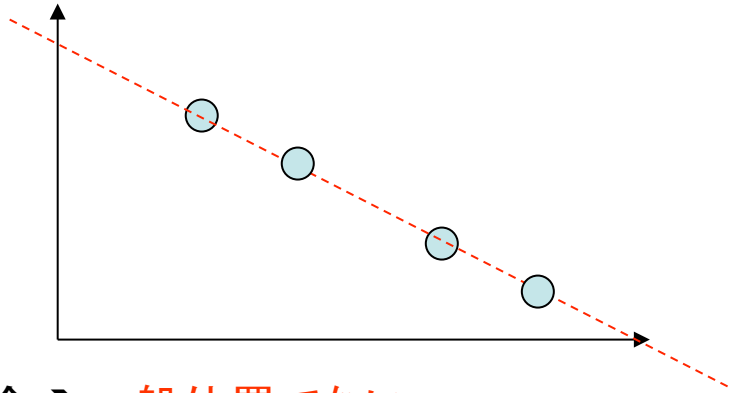
- 次元との数で議論
  - 2次元で考えてみる
  - 学習事例が2つだと1次元で考えてるのと同じ ➔不足



特徴空間

特徴空間は2次元なのに1次元上の識別と変わりがない

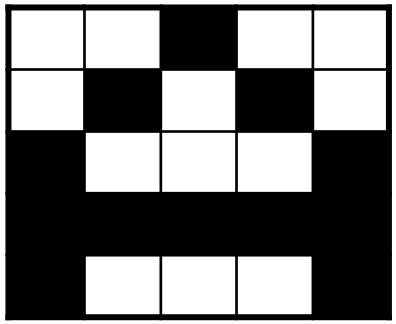
同様に学習データの位置も
特殊だと議論が一般的でない
➔ 一般位置



2次元で4つのデータがある
のに 1次元上に全てのデータがある場合➔一般位置でない

復習

- 5x5のセンサでよみとった文字を認識したいとする. どの程度正解データを用意すれば信頼できる識別関数が構築できるか?



学習データの数

- 線形分離できる可能性
  - 線形識別が見つかる可能性
  - 2(次元数+1) より学習データが少ないとほぼ100%みつかる (石井他 パターン認識と学習p. 65)
    - みつかっても使うときに問題有り
  - 2(次元数+1)より学習データが大きいとほぼ0%
    - みつかると使うときに有効

整理

- 学習データが用意した次元数より少ない場合線形識別関数が求まりやすいのはなぜか
  - 観測した学習データをどのような組にでも分解する平面が求まる確率が高くなるから
- また求まったとしても役に立たない可能性が高いのはなぜか
  - 次元数が多いため学習データに対して分離可能しても，分離するための特徴がうまく取り出せていない可能性が高いため未知のデータに対して良い予測である可能性が少ない

他のパラメータの最適化

- BP法の中間層のユニット数など実験的に決める場合
- 基本アイデア
  - 学習データを分割
    - 学習データ と テストデータ
  - 学習データで基本パラメータwを学習
    - 中間層のユニットもある値γで固定しておく
  - テストデータでの誤り量 $\mathbf{e}_\gamma$ を最小化するγを探す

やり方いろいろ
cross validationなど