# はじめてのパターン認識
# 第11章
# 識別器の組み合わせによる性能強化

大内　克之

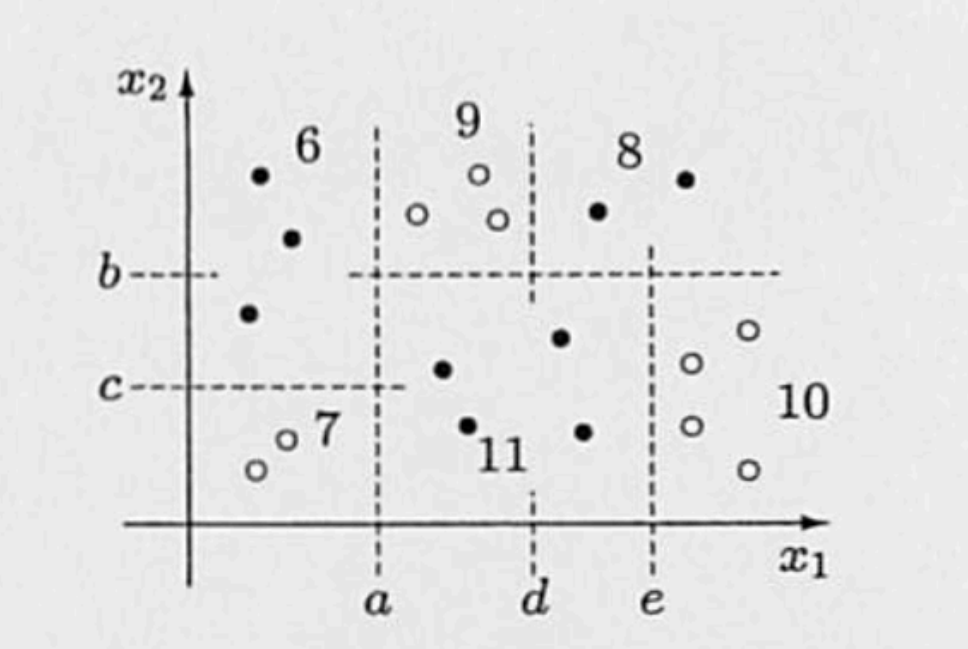
# ノーフリーランチ定理

ノーフリーランチ定理は、「全ての識別問題に対して、他の識別器より識別性能が良い識別器は存在しない」ということを主張するものである。

| 学習データ | | | | テストデータ | | | | |
|---|---|---|---|---|---|---|---|---|
| x | t | h1 | h2 | x | t1 | t2 | h1 | h2 |
| 000 | −1 | −1 | −1 | 011 | 1 | −1 | −1 | 1 |
| 001 | −1 | −1 | −1 | 100 | −1 | 1 | −1 | 1 |
| 010 | 1 | 1 | 1 | 101 | 1 | −1 | −1 | 1 |
| | | | | 110 | −1 | 1 | −1 | 1 |
| | | | | 111 | −1 | 1 | −1 | 1 |

# 決定木

単純な識別規則を組み合わせて複雑な識別境界を得る方法に決定木がある。

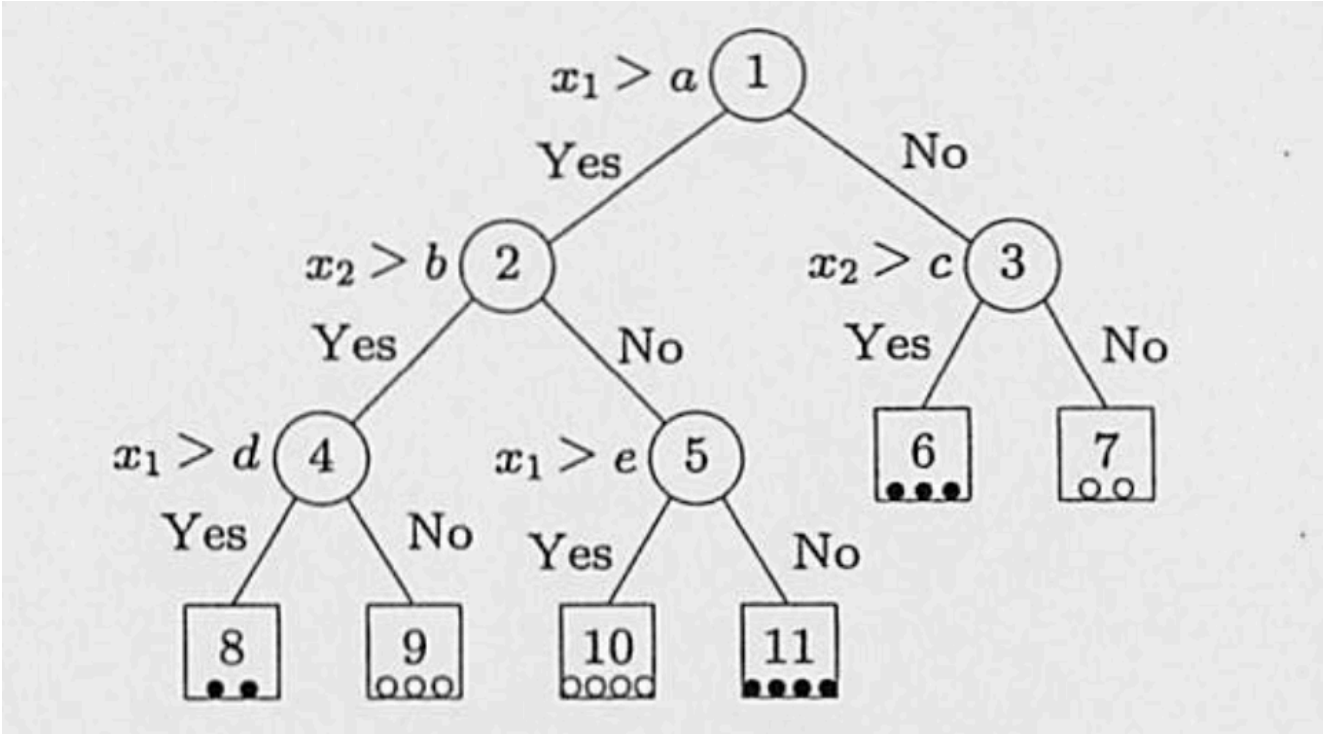



上の例では、識別関数は非線形になる。

しかし、工夫をすれば●と○を識別できる。

# 決定木

大小関係を判断する過程は決定木として表せる。



# 識別器を組み合わせる

複数の識別器を組み合わせる方法には、

- バギング
- ブースティング
- ランダムフォレスト

が挙げられる。

# バギング

学習データのブーストラップサンプルを用いて、複数の識別器の多数決で決める方法をバギングという。

この方法では、識別器がもつばらつきにはブートストラップサンプルがもつばらつきが反映されるだけなので、十分な性能強化が出来ない可能性がある。

# ブースティング

複数の弱識別器を用いて、直列的に学習する方法をブースティングという。

その中で代表的なものにアダブーストがある。

# アダブースト

アダブーストのアルゴリズムを以下に示す。

(1)重みを$w_i^1=1/N(i=1,....,N)$に初期化する。

(2)m=1,....,Mについて、以下を繰り返す。

(a)識別器$y_m(x)$を重み付き識別関数

$$E_m = \frac{\sum_{i=1}^{N} w_i^m I(y_m(\boldsymbol{x}_i) \neq t_i)}{\sum_{i=1}^{N} w_i^m}$$

が最小になるように学習する。$I(y_m(x_i)\neq t_i)$は識別関数の出力が教師データと一致したとき0,しなかったとき1となる支持関数。

(b)重み$a_m$を計算する。

$$\alpha_m = \ln\left(\frac{1-E_m}{E_m}\right)$$

(c)重み$w_i^m$を次のように更新する。

$$w_i^{m+1} = w_i^m \exp\{\alpha_m I(y_m(\boldsymbol{x}_i) \neq t_i)\}$$

(3)識別結果を

$$Y_M(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\boldsymbol{x})\right)$$

に従って出力する。

# ランダムフォレスト

バギングを改良し、あらかじめ決められた数の、相関の低い多様な決定木を選択する方法をランダムフォレストという。

学習は単純だが、SVMやアダブーストと同等、もしくはそれ以上の性能を持つ。

# ランダムフォレスト

アルゴリズムは以下の通り。

(1)m=1からMまで以下を繰り返す。

(a)N個のd次元学習データからブートストラップ$Z_m$を生成する。

(b)$Z_m$を学習データとして以下の手順により各ノードtを分割し、決定木$T_m$を成長させる。終端ノードのデータ数の下限は1。

(i)d個の特徴からランダムにd`個の特徴を選択する(d`=[√d]が推奨されているが、問題によって最適なd'は変わるので、調整パラメータである。)

(ii)d`個の中から最適な分割を与える特徴と分割点を求める。

(iii)ノードtを、分割点left(t)とright(t)に2分割する。

(2)ランダムフォレスト$\{T_m\}_{m=1}^{M}$を出力する。

(3)入力データxに対するm番目の木の識別結果を、$y_m(x) \in \{C_1,\ldots,C_K\}$とする。ランダムフォレスト$\{T_m\}_{m=1}^{M}$の識別結果を、$C_i=\mathrm{argmax}|C_j|$とする。$|C_j|$はクラス$C_j$と判断した木の数である。