情報意味論 (8) Boosting

慶應義塾大学理工学部
櫻井 彰人

競馬でもうけるには？

- 予想屋 (ではなく専門家に) 訊く
- 仮定:
  - どの専門家も、正確な法則は答えられない
  - けれども、いくつかのレース結果をみれば、ランダムよりはましな、法則を教えてくれる
- 儲かるか？

アイデア

- 専門家に経験則を聞く
- 経験則が失敗する事例を集める (困難事例)
- この困難事例について、専門家の意見を聞く
- そして…

- こうして得られた経験則をすべて統合する
- 実は、専門家でなくても弱学習者 "weak" learning algorithm でもよい

課題

- (教えを請うときには) どのレースについて訊けばよいのか?
  - 最も難しいレースに集中する
    (それまでの経験則では最も外れているレースのこと)
- これらの経験則を一つの予測規則に統合するにはどうするのか?
  - 経験則の (重み付き) 多数決をとる

Boosting

- boosting = 複数個の大雑把な経験則を高精度な予測規則に変換する一般的方法
- より技術的には:
  - 弱(weak)学習アルゴリズム (誤差 $\leq 1/2-\gamma$ なる仮設 (分類規則) を常に見出すことができる) が与えられたとき
  - boosting アルゴリズムは、誤差 $\leq \varepsilon$ なる単一の仮説を構成することができる (ことが証明できる)
  - 理論によれば、しばしば、汎化能力はよい

つもり

- boosting 入門 (AdaBoost)
- 訓練誤差の解析
- マージンの理論に基づく、汎化誤差の検討
- 拡張
- 結果例

以下のスライドは、主に、下記論文に基づく
Robert E. Schapire.
The boosting approach to machine learning: An overview.
In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, *Nonlinear Estimation and Classification*. Springer, 2003.

背景

- [Valiant'84]
  PAC (Probably Approximately Correct) 学習の提唱
- [Kearns and Valiant'88]
  boosting アルゴリズムを見出すことの提案
- [Schapire'89], [Freund'90]
  最初の、多項式時間 boosting アルゴリズム
- [Drucker, Schapire and Simard '92]
  boosting を用いた最初の実験結果

背景 (続)

- [Freund and Schapire '95]
  - AdaBoost の提案
  - 以前の boosting アルゴリズムより実用的価値が高い
- AdaBoost 使用例:
  [Drucker & Cortes '95] [Schapire & Singer '98]
  [Jackson & Cravon '96] [Maclin & Opitz '97]
  [Freund & Schapire '96] [Bauer & Kohavi '97]
  [Quinlan '96] [Schwenk & Bengio '98]
  [Breiman '96] [Dietterich'98]
- 理論とアルゴリズム:
  [Schapire,Freund,Bartlett & Lee '97] [Schapire & Singer '98]
  [Breiman '97] [Mason, Bartlett & Baxter '98]
  [Grive and Schuurmans'98] [Friedman, Hastie & Tibshirani '98]

Boosting を形式化

- 所与の訓練データ集合 $X=\{(x_1,y_1),\ldots,(x_m,y_m)\}$
- $y_i \in \{-1,+1\}$ 事例 $x_i \in X$ に対する正しいラベル
- for $t = 1,\ldots,T$:
  - 分布 D_t を $\{1,\ldots,m\}$ の上に定める
  - 弱仮説を見出す
    $$h_t : X \to \{-1,+1\}$$
    ただし D_t 上で小さい誤差 ε_t あり
    $$\varepsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$$
- 最終仮説 H_{final} を出力

AdaBoost [Freund & Schapire '97]

- D_t の作成:
  - $D_1(i) = \frac{1}{m}$
  - 所与 D_t と h_t:
    $$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$
    $$= \frac{D_t}{Z_t} \cdot \exp(-\alpha_t \cdot y_i \cdot h_t(x_i))$$
    ただし: Z_t = 正規化定数
    $$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$$
- 最終仮説: $H_{\text{final}}(x) = \text{sgn}\left(\sum_t \alpha_t h_t(x)\right)$

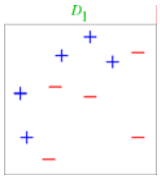
AdaBoost 主要部

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$$

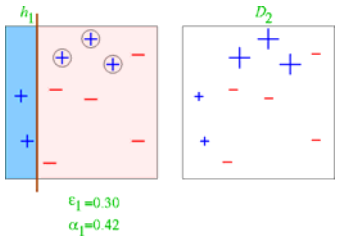
$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$H_{\text{final}}(x) = \text{sgn}\left(\sum_t \alpha_t h_t(x)\right)$$

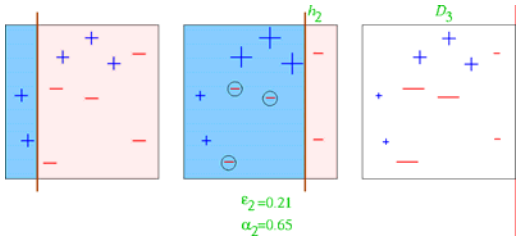
トイ



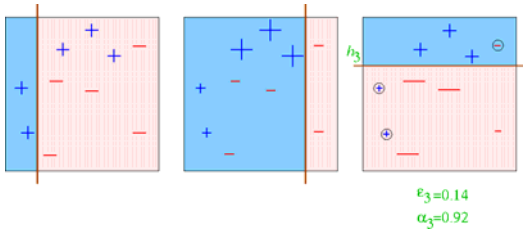
第一巡目



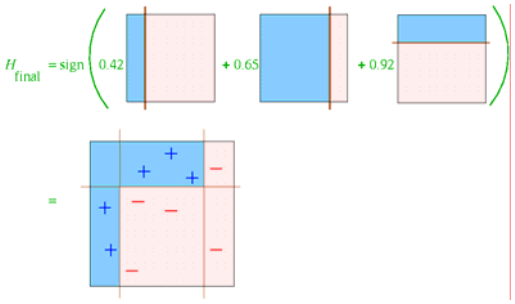
第二巡目



第三巡目



最終仮説



Boosting Applet

http://www.cse.ucsd.edu/~yfreund/adaboost/index.html

$$\sum_{i:\text{正}} D_t(i) = \varepsilon_t$$

$$\sum_{i:\text{誤}} D_t(i) = 1-\varepsilon_t$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

$$\sum_{i:\text{正}} D_t(i)\exp(-\alpha_t) = (1-\varepsilon_t)\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} = \sqrt{(1-\varepsilon_t)\varepsilon_t}$$

$$\sum_{i:\text{誤}} D_t(i)\exp(\alpha_t) = \varepsilon_t\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} = \sqrt{(1-\varepsilon_t)\varepsilon_t}$$

$$\begin{aligned} Z_t &= \sum_{i=1}^{N} D_t(i)\exp(-\alpha_t y_i h_t(x_i)) \\ &= \sum_{i:\text{正}} D_t(i)\exp(-\alpha_t) + \sum_{i:\text{誤}} D_t(i)\exp(\alpha_t) \\ &= 2\sqrt{(1-\varepsilon_t)\varepsilon_t} \end{aligned}$$

$$\begin{aligned} e_t &= \Pr_{D_{t+1}}\{h_t(x_i) \neq y_i\} \\ &= \sum_{i:\text{誤}} D_{t+1}(i) \\ &= \sum_{i:\text{誤}} \frac{D_t(i)\exp(\alpha_t)}{Z_t} \\ &= \frac{\sqrt{(1-\varepsilon_t)\varepsilon_t}}{2\sqrt{(1-\varepsilon_t)\varepsilon_t}} \\ &= 0.5 \end{aligned}$$

訓練誤差の解析

- **定理** [Freund and Schapire '97]:

  ε_t を ½-γ_t と書く

  この時 $\text{training error}(H_{\text{final}}) \le \exp\left(-2\sum_t \gamma_t^2\right)$

  従って、もし $\forall t: \gamma_t \ge \gamma > 0$ **なら**

  $$\text{training error}(H_{\text{final}}) \le e^{-2\gamma^2 T}$$

- **注**: AdaBoost **は** adaptive:
  - γ **や** T **を事前に知っている必要はない**
  - $\gamma_t >> \gamma$ **としてもよい**

$$\exp(-\alpha_t y_i h_t(x_i)) = \frac{D_{t+1}(i)}{D_t(i)} Z_t$$

$$\exp\left(-y_i \sum_t \alpha_t h_t(x_i)\right) = \frac{D_{T+1}(i)}{D_1(i)} \prod_t Z_t$$

$$= N D_{T+1}(i) \prod_t Z_t$$

$$\frac{1}{N}\left|\{i \mid H(x_i) \ne y_i\}\right| = \frac{1}{N}\sum_{i=1}^{N} I(H(x_i) \ne y_i)$$

$$\le \frac{1}{N}\sum_{i=1}^{N} \exp(-y_i f(x_i))$$

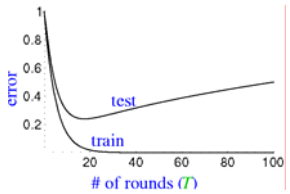
$$= \prod_t Z_t$$

$$= \prod_t \left(2\sqrt{(1-\varepsilon_t)\varepsilon_t}\right)$$

$$= \prod_t \sqrt{1-4\gamma_t^2}$$

$$\le \exp\left(-2\sum_{t=1}^{T} \gamma_t^2\right)$$

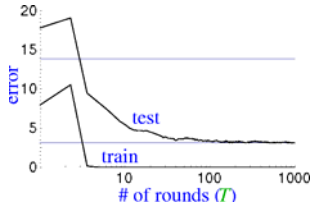
汎化誤差の解析



期待すること:

- **訓練誤差は、継続して、低下する** (0になるかも)
- H_{final} **が複雑になりすぎると、テスト誤差は、増大する** (**オッカムの剃刀**)

ある実験結果 [Schapire et al. 98]

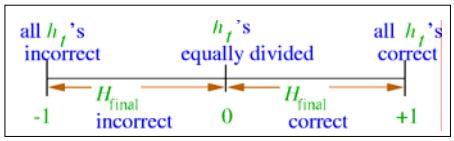


(boosting on C4.5 on "letter" dataset)

- **1,000 巡以降でもテスト誤差は増加しない**
- **訓練誤差が0となった後も、テスト誤差は減少を続ける**
- **オッカムの剃刀のいう単純な規則がよいというのは、誤り**

http://www.cs.princeton.edu/courses/archive/fall05/cos402/readings/boost-slides.pdf

マージンからみると



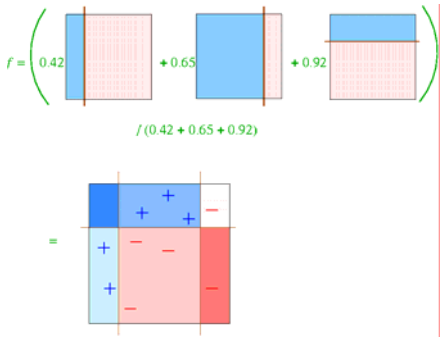
アイデア: **信頼度 (マージン) を考えよう**:

- **まず下記に注意**

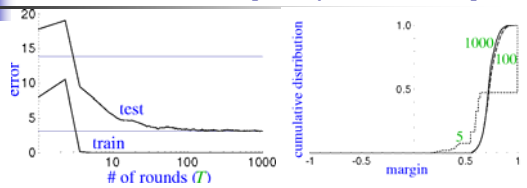
  $H_{\text{final}}(x) = \text{sgn}(f(x))$ $\qquad \dfrac{f(x)}{\sum_t |\alpha_t|} = \dfrac{\sum_t \alpha_t h_t(x)}{\sum_t |\alpha_t|} \in [-1,1]$

- **定義**: (x,y) **のマージン**: $\text{margin}_f(x,y) = \dfrac{y \cdot f(x)}{\sum_t |\alpha_t|}$

トイ



マージンの分布 [Schapire et al. 98]



epoch	5	100	1000
training error	0.0	0.0	0.0
test error	8.4	3.3	3.1
%margins≤0.5	7.7	0.0	0.0
Minimum margin	0.14	0.52	0.55

Boosting はマージンを最大化する

- 次の損失関数を最小化することが示せる

$$\sum_i e^{-y_i F(x_i)} = \sum_i e^{-y_i \sum_t \alpha_t h_t(x_i)}$$

(x_i, y_i) のマージンに比例

マージンに基づく解析

汎化誤差を訓練事例のマージンの関数で抑える:

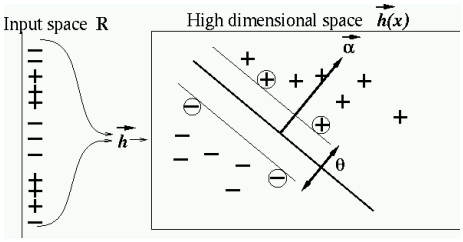
$$\text{error} = \Pr[\text{margin}_f(x,y) \le 0]$$

$$\le \hat{\Pr}[\text{margin}_f(x,y) \le \theta] + \tilde{O}\left(\sqrt{\frac{\text{VC}(H)}{m\theta^2}}\right)$$

- マージン大 ⇒ 上界が小さいくなる
- 上界は学習エポック数に依存しない
- boosting は、マージンが最小の事例に着目することにより、訓練事例のマージンを増加させる傾向にある

SVM との関係

SVM: x を高次元空間に写像して、線形分離する



SVM との関係（続）

$$H(x) = \begin{cases} +1 & \text{if } 2x^5 - 5x^2 + x > 10 \\ -1 & \text{otherwise} \end{cases}$$

$$\vec{h}(x) = (1, x, x^2, x^3, x^4, x^5)$$
$$\vec{\alpha} = (-10, 1, -5, 0, 0, 2)$$

$$H(x) = \begin{cases} +1 & \text{if } \vec{\alpha} \cdot \vec{h}(x) > 0 \\ -1 & \text{otherwise} \end{cases}$$

SVM との関係

- どちらもマージンを最大化する:

$$\theta \doteq \max_w \min_i \frac{(\vec{\alpha} \cdot \vec{h}(x_i)) y_i}{\| \vec{\alpha} \|}$$

- SVM: $\| \vec{\alpha} \|_2$ ユークリッドノルム (L_2)
- AdaBoost: $\| \vec{\alpha} \|_1$ マンハッタンノルム (L_1)

- 最適化や PAC による上界と関係がでてくる

[Freund et al '98]

拡張: 多クラス問題

- 事例ごとに2進分類問題に還元する:
  - 事例 x はクラス1に属するか否か？
  - 事例 x はクラス2に属するか否か？
  - 事例 x はクラス3に属するか否か？
  - ⋮

拡張: 信頼度と確率

- 仮説の予測 h_t: $\mathrm{sgn}(h_t(x))$
- 仮説の信頼度 h_t: $|h_t(x)|$
- 確率 H_{final}: $\Pr_f[y=+1|x] = \dfrac{e^{f(x)}}{e^{f(x)}+e^{-f(x)}}$
  - log loss 最小化

$$\sum_i \ln\left(1+e^{-2y_i f(x_i)}\right)$$

[Schapire and Singer '98], [Friedman, Hastie and Tibshirani '98]

AdaBoost の実用的価値

- かなり速い
- 単純かつ容易にプログラムできる
- チューニングパラメータは一個だけ (T)
- 事前知識不要
- 融通性: どんな分類器とも組合せ可能 (ニューラルネット, C4.5, …)
- 有効性が証明済み (弱学習器は仮定する)
  - 発想の転換: 目標は、単に、random guessing よりよい仮説を見つければよいだけ
- はずれ値も見つける

御注意

- 性能は、データと当該弱学習器に依存
- AdaBoost が失敗するのは
  - 弱学習器が複雑すぎる (過学習)
  - 弱学習器が弱すぎる ($\gamma_t \to 0$ となるのが速すぎる)

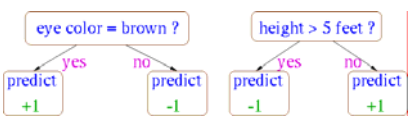
$$\text{training error}(H_{\text{final}}) \le e^{-2\sum_t \gamma_t^2}$$

    - 学習不足
    - マージンが小 → 過学習
- 経験的には, AdaBoost はノイズの影響を受けやすいように思われる

UCI ベンチマーク

比較

- C4.5 (Quinlan の決定木学習)
- Decision Stumps (切株. ノード一個)



UCI 結果

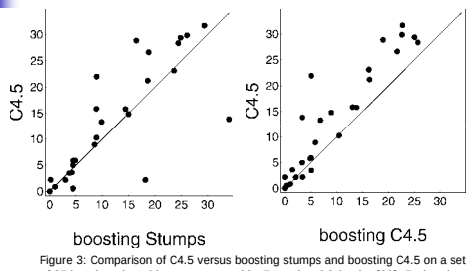
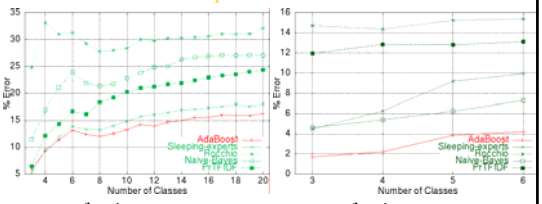


Figure 3: Comparison of C4.5 versus boosting stumps and boosting C4.5 on a set of 27 benchmark problems as reported by Freund and Schapire [30]. Each point in each scatterplot shows the test error rate of the two competing algorithms on a single benchmark. The y-coordinate of each point gives the test error rate (in percent) of C4.5 on the given benchmark, and the x-coordinate gives the error rate of boosting stumps (left plot) or boosting C4.5 (right plot). All error rates have been averaged over multiple runs.

テキスト分類

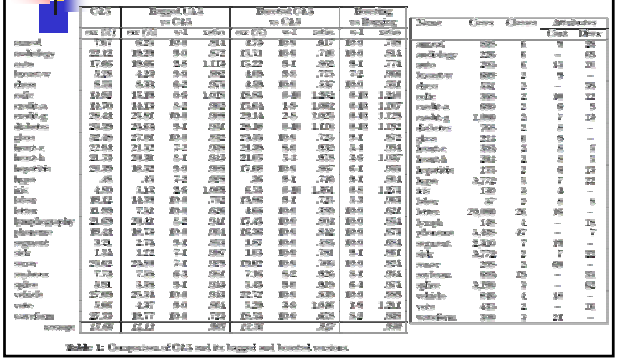
- Decision stumps: **単語や短い句の存在**.
  例:

  "If the word Clinton appears in the document predict document is about politics"



データ: AP　　**データ**: Reuters

他の比較 [Quinlan, '96]



まとめ

- boosting **は分類課題に有用**
  - **豊富な理論に裏付けられる**
  - **実験的にも、パフォーマンスの良さが確認ずみ**
  - **しばしば (いつも、ではない) 過学習しにくい**
  - **応用事例多い**
- **しかし**
  - **(得られた) 分類器は遅い**
  - **結果は、分かりにくい**
  - **ノイズに敏感なことあり**