機械学習

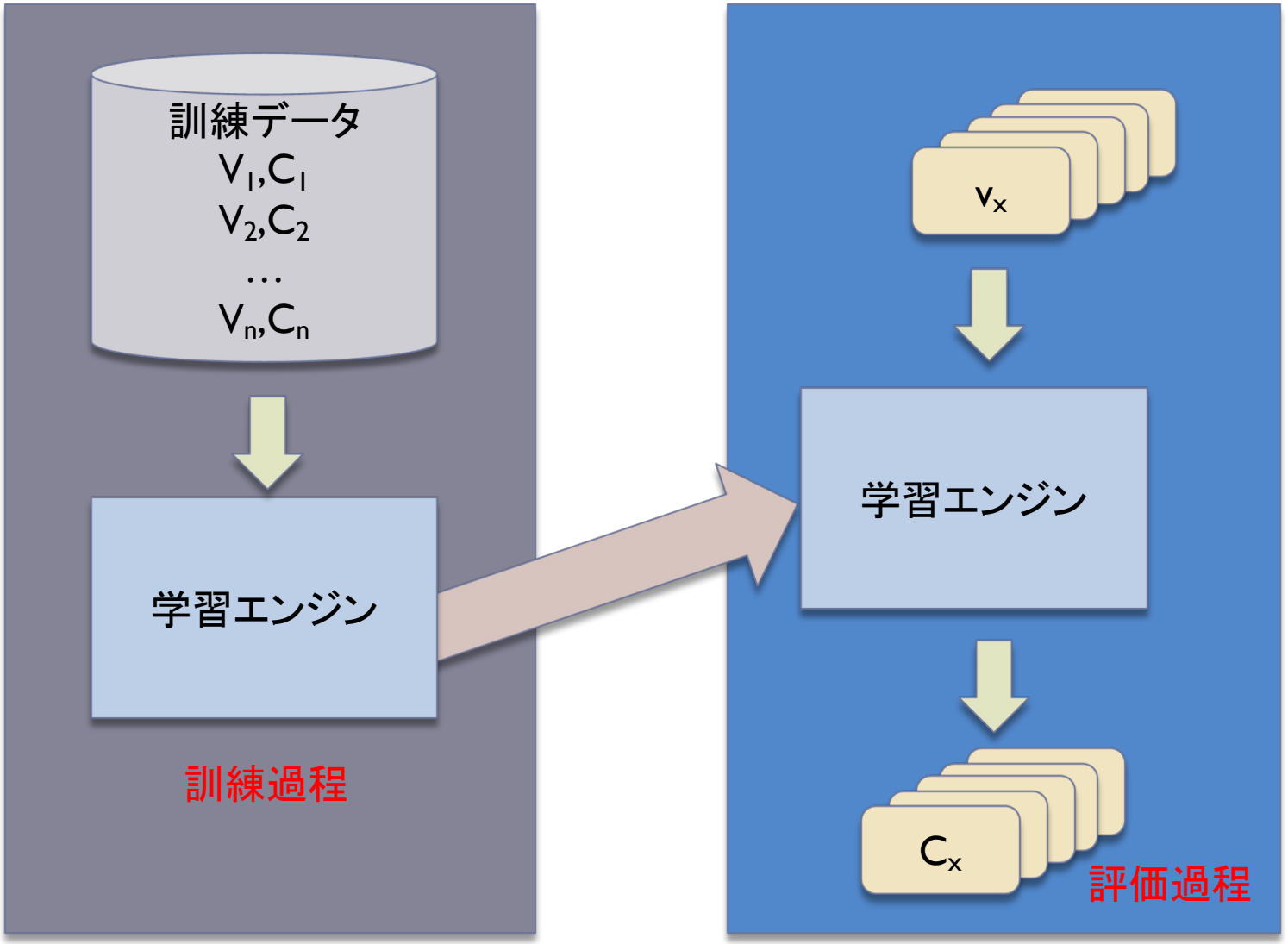
機械学習

- 教師付き学習：
  - 分類すべきクラスが示された事例データにより学習し、未知データを分類する問題。
- 属性と属性値（素性、素性値ともいう）
  - 属性：事例（事象）を特徴付ける「パラメータ」
  - 属性値：
    - 数値：整数や実数などの連続的な値
    - 離散クラス：有限の要素からなる
- 事例と特徴ベクトル
  - 事例
    - 訓練の際の観察されたデータ、または評価の際の判定すべきデータ。
    - 事例は、何らかのクラスに属し、特徴ベクトルにより表現される。
  - 特徴ベクトル
    - 各観測事例について、事例を特徴付ける属性（と値）を並べたもの。
    - 属性の個数＝ベクトルの次元
- 教師なし学習：

  事例を適当なグループに分類する問題。クラスタリングと等価。
- 強化学習：

  報酬を得るために、エージェントが現在の状態からとるべき行動を学習する問題。

（教師付き）機械学習の概要



訓練集合の例

天候	温度(° F)	湿度(%)	強風	クラス
晴れ	75	70	真	開催
晴れ	80	90	真	中止
晴れ	85	85	偽	中止
晴れ	72	95	偽	中止
晴れ	69	70	偽	開催
曇り	72	90	真	開催
曇り	83	78	偽	開催
曇り	64	65	真	開催
曇り	81	75	偽	開催
雨	71	80	真	中止
雨	65	70	真	中止
雨	75	80	偽	開催
雨	68	80	偽	開催
雨	70	96	偽	開催

- 訓練データ
  - 原理的には、区別するクラスの総数よりも多くの訓練データが必要。数十から数百、数万件。
  - アンケート、調査により十分な事例データを収集し、一部を訓練データ、残りを評価データとして用いる(交差検定)。

属性定義ファイル(決定木学習C4.5)の例

開催,中止
天候: 晴れ,曇り,雨
温度: continuous
湿度: continuous
強風: 真,偽

(最初の1行)分類するクラス

属性「天候」の値は、「晴れ」「曇り」「雨」のいずれか

属性「温度」の値は、連続値

実際の定義ファイル(実際の記述)

```
Play, Don't Play.

outlook: sunny, overcast, rain.
temperature: continuous.
humidity: continuous.
windy: true, false.
~
~
~
```

前頁の訓練データ（C4.5での実際の記述（部分））

```
sunny, 85, 85, false, Don't Play
sunny, 80, 90, true, Don't Play
overcast, 83, 78, false, Play
rain, 70, 96, false, Play
rain, 68, 80, false, Play
rain, 65, 70, true, Don't Play
overcast, 64, 65, true, Play
sunny, 72, 95, false, Don't Play
sunny, 69, 70, false, Play
rain, 75, 80, false, Play
sunny, 75, 70, true, Play
overcast, 72, 90, true, Play
overcast, 81, 75, false, Play
rain, 71, 80, true, Don't Play
~
~
.
```

	手法	精度	モデルの理解度	属性の多さ
分類	決定木	○	◎	×
	ナイーブベイズ	◎	○	○
	SVC	◎	×	◎
	ランダムフォレスト	◎	×	○
回帰	重回帰	○	○	×
	Lasso（L1正則化）, Ridge（L2正則化）	○	◎	○
	SVR	◎	×	◎
クラスタリング	階層的クラスタリング	○	◎	×
	K平均法	○	◎	○
	スペクトラルクラスタリング	◎	×	◎
	混合正規分布	◎	◎	○

コーパスとテキストマイニング、pp.18-19

決定木学習（C4.5）

- 1992年　J.Ross Quinlan
- エキスパートシステム構築の需要
  - AMEX：クレジットの許可の判定支援
  - TI:資本経費の計画の提案支援
  - Gravan医療研究所：甲状腺の検査
- 記録された膨大な分類データを調べ、特定の例を一般化することによりモデルを帰納的に作る。（他の方式についても。）

属性定義ファイル(C4.5)の例

開催,中止

天候: 晴れ,曇り,雨

温度: continuous

湿度: continuous

強風: 真,偽

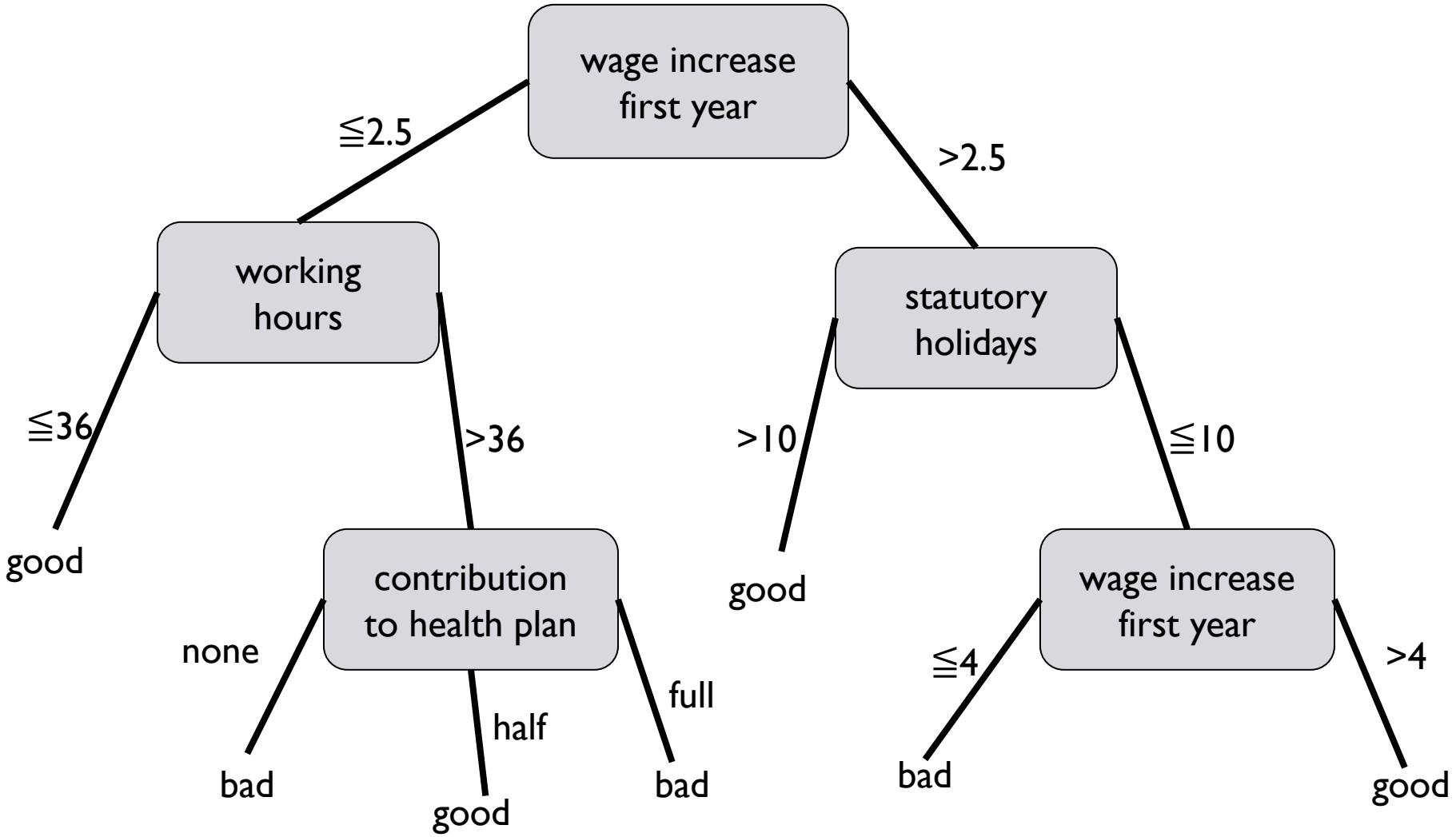
訓練集合の例

天候	温度(° F)	湿度(%)	強風	クラス
晴れ	75	70	真	開催
晴れ	80	90	真	中止
晴れ	85	85	偽	中止
晴れ	72	95	偽	中止
晴れ	69	70	偽	開催
曇り	72	90	真	開催
曇り	83	78	偽	開催
曇り	64	65	真	開催
曇り	81	75	偽	開催
雨	71	80	真	中止
雨	65	70	真	中止
雨	75	80	偽	開催
雨	68	80	偽	開催
雨	70	96	偽	開催

決定木

- ノード：候補のクラスを絞りこむためのテストが付加されている。
- ラベル：テストに応じて分岐するための条件が付加される。
- 葉：判定したクラスを示す。

決定木の例


wage increase first year
≦2.5
>2.5
working hours
statutory holidays
≦36
>36
>10
≦10
good
contribution to health plan
good
wage increase first year
none
half
full
≦4
>4
bad
good
bad
bad
good

決定木の例（式）

if 最初の年の賃金増加率 ≦ 2.5 then
　if 労働時間　≦ 36 then クラスはgood
　else if 労働時間 > 36 then
　　if 健康維持への援助はnone then クラスはbad
　　else if 健康維持への援助はhalf then クラスはgood
　　else クラスはbad
else if 最初の年の賃金増加率 > 2.5 then
　if　法定休暇 > 10 then クラスはgood
　else if 法定休暇 ≦ 10 then
　　if 最初の年の賃金増加率 ≦ 4 then クラスはbad
　　else if 最初の年の賃金増加率 > 4 then クラスはgood

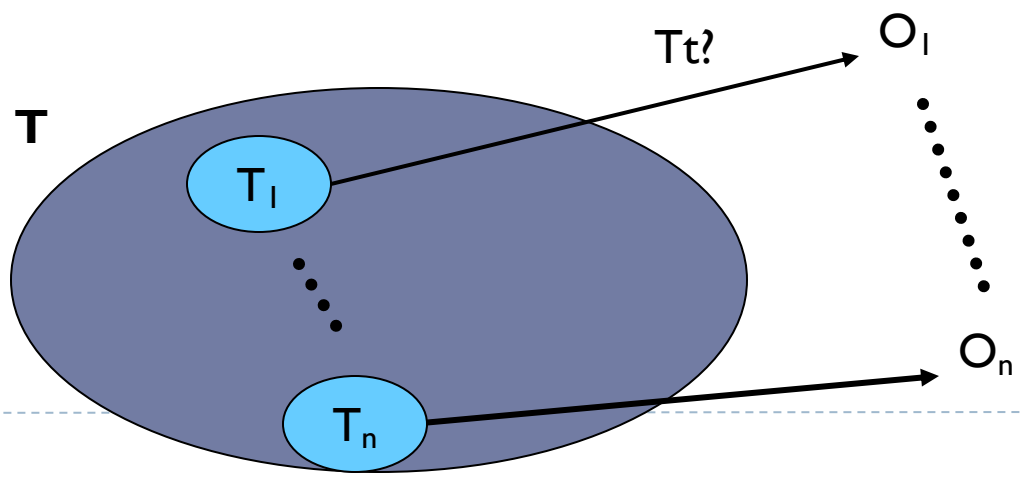
決定木の構成（分割統治法１）

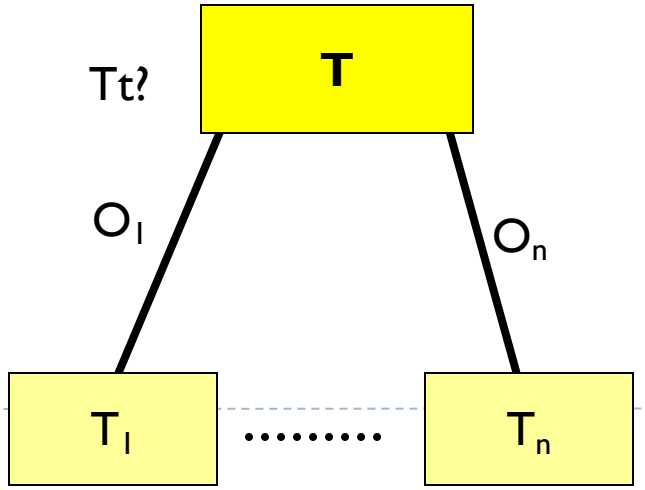
訓練事例の集合Tについて、以下を停止するまで繰り返す：

1. Tは1つ以上の事例を含み、その全ての事例が一つのクラスC_jのみに属する：
   - Tの決定木は1つの葉のみ。クラスはC_j
2. Tは全く事例を含まない場合：
   - （T以外の情報で決定する。）C4.5では親ノードで最も頻繁に現れたクラス。

決定木の構成（分割統治法２）

3. Tが種々のクラスに属する事例を含む場合：
   - ある属性についてのテストTtで異なる結果$\{O_1,O_2,\ldots,O_n\}$が得られるとき、結果がO_iになるTの部分集合をT_iとする。
   - Tを$T_1,T_2,\ldots,T_n$に分割する。
   - $T_1,T_2,\ldots,T_n$のそれぞれについて、同様に再帰的に分割していく。





事例の分割

晴れ	75	70	真	開催
晴れ	80	90	真	中止
晴れ	85	85	偽	中止
晴れ	72	95	偽	中止
晴れ	69	70	偽	開催
曇り	72	90	真	開催
曇り	83	78	偽	開催
曇り	64	65	真	開催
曇り	81	75	偽	開催
雨	71	80	真	中止
雨	65	70	真	中止
雨	75	80	偽	開催
雨	68	80	偽	開催
雨	70	96	偽	開催

対応する決定木

if 天候=晴れ then
　if 湿度≦75 then クラスは「開催」
　else if 湿度>75 then クラスは「中止」
else if 天候=曇り then クラスは「開催」
else if 天候=雨 then
　　if 強風=真 then クラスは「中止」
　　else if 強風=偽 then クラスは「開催」

効率的なテストの選び方（利得基準）

- テストの選び方：評価結果により最大の効果を持つ分割が優先されるべき。
- 定義
  - 情報量：
    - 値がn通りの違いがある情報源の情報量（ただし等確率で生起）

      ⇒　$-\log_2 1/n=\log_2 n$ビット
    - メッセージを伝えるための情報の桁数
  - freq(C_i,S):Sの中でクラスC_iに属する事例の数

利得基準（その２）

- 事例集合Sからランダムに一つの事例を選び、それがクラス C_j に属するとき、このメッセージの確率

$$\frac{freq(C_j, S)}{|S|}$$

- それが伝える情報量

$$-\log_2\left(\frac{freq(C_j, S)}{|S|}\right)$$

利得基準（その３）

- S内で頻度で重み付けしたメッセージの平均情報量（Sのエントロピー）:

$$\mathrm{info}(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right)$$

- テストXのn通りの結果によりTを分割、クラスを同定するのに必要な情報量の期待値:

$$\mathrm{info}_X(T) = \sum_{j=1}^{n} \frac{|T_i|}{|T|} \times \mathrm{info}(T_j)$$

利得基準（その４）

- 差（利得基準）:

$$gain(X) = \text{info}(T) - \text{info}_X(T)$$

- 利得比基準（参考）:

$$\text{split info}(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right)$$

（テストの結果を伝えるメッセージの情報量）

$$gain\ ratio(X) = gain(X) / split\ info(X)$$

利得基準の例題

先の例で、

- １事例のクラスを同定するために必要な情報量: info(T)
- 「天候」で分割した後クラスを同定するのに必要な情報量: $info_{天候}(T)$
- 「強風」で分割した後クラスを同定するのに必要な情報量: $info_{強風}(T)$
- 利得基準は？

  gain(天候) と　gain(強風) を比べる。

1事例のクラスを同定するために必要な情報量info(T)

$$\begin{aligned}\mathrm{info}(T) &= -\left(\frac{5}{14}\times\log_2\frac{5}{14}+\frac{9}{14}\times\log_2\frac{9}{14}\right)\\ &= -\left(-0.357\times1.485-0.643\times0.637\right)\\ &= 0.940\end{aligned}$$

「天候」で分割した後クラスを同定するのに必要な情報量$\mathrm{info}_{天候}(T)$

$$
\begin{aligned}
\mathrm{info}_{天候}(T) &= \frac{5}{14}\mathrm{info}(T_{晴れ}) + \frac{4}{14}\mathrm{info}(T_{曇り}) + \frac{5}{14}\mathrm{info}(T_{雨}) \\
&= \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\
&\quad + \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) \\
&\quad + \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\
&= 0.357\times\left(0.4\times 1.322 + 0.6\times 0.737\right)\times 2 \\
&= 0.694
\end{aligned}
$$

「強風」で分割した後クラスを同定するのに必要な情報量$\mathrm{info}_{強風}(T)$

$$
\begin{aligned}
\mathrm{info}_{強風}(T) &= \frac{6}{14}\mathrm{info}(T_{真}) + \frac{8}{14}\mathrm{info}(T_{偽}) \\
&= \frac{6}{14}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right) \\
&\quad + \frac{8}{14}\left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) \\
&= 0.429\times(0.5\times 1 + 0.5\times 1) \\
&\quad + 0.571\times(0.75\times 0.415 + 0.25\times 2) \\
&= 0.892
\end{aligned}
$$

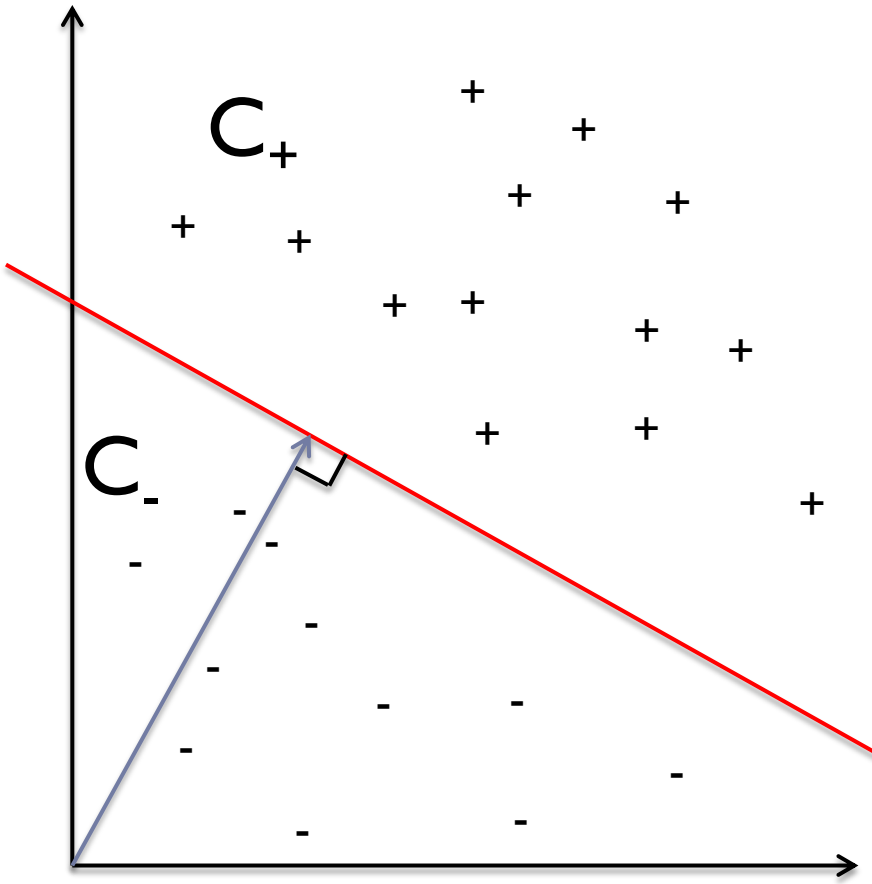
利得基準

- gain(天候)=0.940-0.694=0.246
- gain(強風)=0.940-0.892=0.048

SVM(Support Vector Machine)

- Vapnik 1995
- 線形二値分類器
- 1990年代終わりから、自然言語処理その他において最も多く使われている。

SVMによる識別器(1)



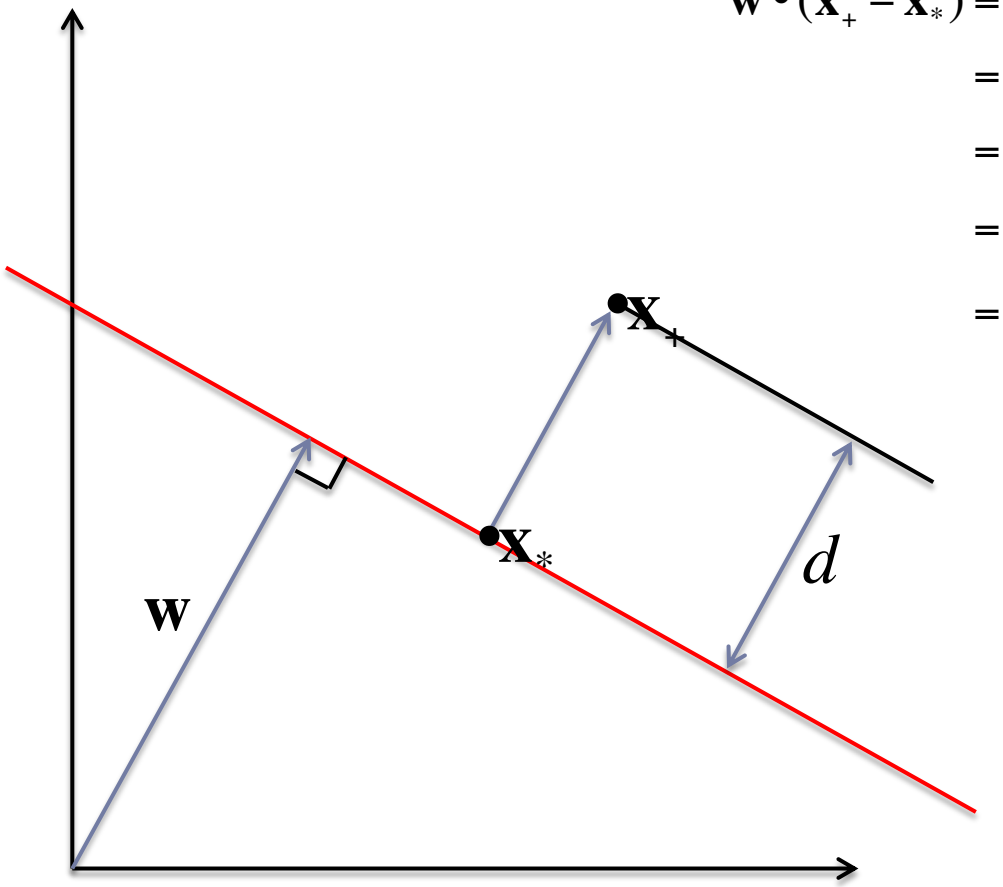
$$\mathbf{x}^{(i)} \in C_+ \quad \cdots\cdots \quad if\ \mathrm{f}(\mathbf{x}^{(i)}) > 0$$

$$\mathbf{x}^{(i)} \in C_- \quad \cdots\cdots \quad if\ \mathrm{f}(\mathbf{x}^{(i)}) < 0$$

となるようなf(x)を求めたい

$$\mathrm{f}(\mathbf{x}) = \mathbf{w} \bullet \mathbf{x} - \mathbf{b} = 0$$

SVMによる識別器(2)



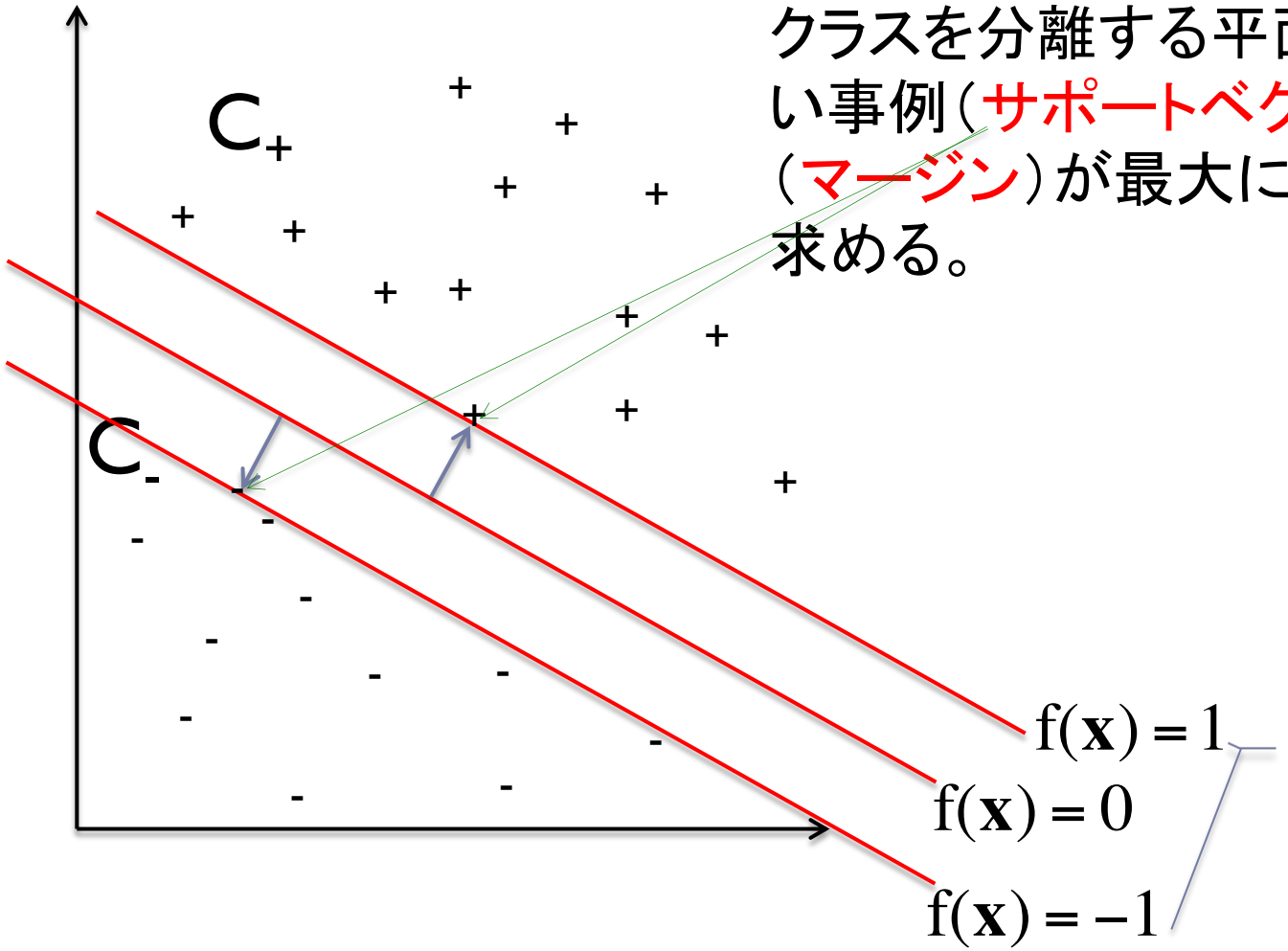
$$\begin{aligned}\mathbf{w}\bullet(\mathbf{x}_+ - \mathbf{x}_*) &= \mathbf{w}\bullet\mathbf{x}_+ - \mathbf{w}\bullet\mathbf{x}_* \\ &= \mathbf{w}\bullet\mathbf{x}_+ - \mathbf{b} \\ &= \mathrm{f}(\mathbf{x}_+) \\ &= \|\mathbf{w}\|\|\mathbf{x}_+ - \mathbf{x}_*\| \\ &= \|\mathbf{w}\| d\end{aligned}$$

$\mathrm{f}(\mathbf{x}_*) = \mathbf{w}\bullet\mathbf{x}_* - \mathbf{b} = 0$

$\mathbf{w}$ と $\mathbf{x}_+ - \mathbf{x}_*$ は平行

$$\therefore \quad d = \frac{|\mathrm{f}(\mathbf{x}_+)|}{\|\mathbf{w}\|}$$

SVMによる識別器(3)



クラスを分離する平面とそれに最も近い事例（サポートベクトル）との距離（マージン）が最大になるような平面を求める。

値が1になるように式を定数倍しても一般性は変わらない

事例からのサポートベクトルの求め方

- 訓練事例

$$D = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \cdots\}$$

$$if\ \mathbf{x}^{(i)} \in C_+ \quad \cdots\cdots \quad y^{(i)} = 1, \mathrm{f}(\mathbf{x}^{(i)}) \geq 1$$

$$if\ \mathbf{x}^{(i)} \in C_- \quad \cdots\cdots \quad y^{(i)} = -1, \mathrm{f}(\mathbf{x}^{(i)}) \leq -1$$

（等号は$x^{(i)}$がサポートベクトルのとき）

まとめると、

$$y^{(i)}\,\mathrm{f}(\mathbf{x}^{(i)}) \geq 1$$

これを解くには

- 次の最適化問題を解けばよい

$$\min. \quad \frac{1}{2}\mathbf{w}^2 \qquad \text{i.e. } d = \frac{|\mathrm{f}(\mathbf{x})|}{\|\mathbf{w}\|} \text{ の最大化}$$

$$s.t. \quad y^{(i)}\,\mathrm{f}(\mathbf{x}^{(i)}) - 1 \geq 0; \forall i$$

多クラス分類（その２）

- pairwise法
  - クラス$C_0, C_1, \ldots, C_{n-1}$から任意の2つのクラスを選ぶ全ての組み合わせ

    $$(C_0, C_1), \ldots, (C_i, C_j),\ldots$$

    に対して、

    それぞれの判定器

    $$\ldots,SVM_{ij},\ldots$$

    を作る。
  - 多数決で一番多く判定されたクラスを採用する。

SVM^{light}

- 作者 : Thorsten Joachims
- http://svmlight.joachims.org
- Windows, Solaris, Linux, Cygwin, Mac OS Xで稼働
- 学習 :

  svm_learn “事例データ” “モデル”
- 評価 :

  svm_classify “事例データ” “モデル” “結果”

SVM^{light}の訓練データの例

action.y

```
1       1:24    2:0     3:1     4:1     5:0     6:0     7:91    8:0     9:1     10:1    11:0    12:1    13:0    14:1    15:1    16:1    # そ
の人は、ずいぶん不仕合せな男なのです。
1       1:91    2:0     3:1     4:1     5:0     6:1     7:23    8:0     9:0     10:1    11:0    12:0    13:0    14:1    15:1    16:1    # ほ
んとうに、その人は、生れて来なかったほうが、よかった」と意外にはっきりした語調で言って、一つまみのパンをとり脱をのばし、あやまたず私の口にひたと押し当てました。
-1      1:23    2:0     3:0     4:1     5:0     6:0     7:11    8:0     9:0     10:1    11:0    12:0    13:1    14:0    15:0    16:0    # 私
も、もうすでに度胸がついていたのだ。
-1      1:11    2:0     3:0     4:1     5:0     6:0     7:24    8:0     9:0     10:1    11:0    12:0    13:0    14:0    15:1    16:0    # 恥
じるよりは憎ん
-1      1:24    2:0     3:0     4:1     5:0     6:0     7:50    8:0     9:0     10:1    11:0    12:1    13:1    14:1    15:0    16:1    # あ
の人の今更なが
-1      1:56    2:0     3:0     4:1     5:0     6:1     7:6     8:0     9:0     10:1    11:0    12:0    13:1    14:1    15:0    16:0    # こ
のように弟子たち皆の前で公然と私を辱かしめるのが、あの人の之までの仕来りなのだ。
-1      1:6     2:0     3:0     4:1     5:0     6:0     7:37    8:0     9:1     10:1    11:0    12:1    13:1    14:1    15:0    16:0    # 火
と水と。
1       1:37    2:0     3:1     4:1     5:0     6:1     7:60    8:0     9:1     10:1    11:0    12:1    13:1    14:0    15:0    16:3    # 永
遠に解け合う事の無い宿命が、私とあいつとの間に在るのだ。
-1      1:60    2:0     3:1     4:1     5:0     6:1     7:4     8:0     9:0     10:1    11:0    12:0    13:0    14:1    15:1    16:0    # 犬
か猫に与えるように、一つまみのパン屑を私の口に押し入れて、それがあいつのせめてもの腹いせだったのか。
-1      1:4     2:0     3:0     4:1     5:0     6:0     7:8     8:0     9:0     10:1    11:0    12:0    13:1    14:0    15:0    16:0    # は
はん。
1       1:8     2:0     3:0     4:1     5:0     6:0     7:37    8:0     9:1     10:1    11:0    12:0    13:1    14:0    15:0    16:0    # ば
かな奴だ。
-1      1:37    2:0     3:1     4:1     5:0     6:0     7:52    8:0     9:1     10:1    11:0    12:1    13:1    14:1    15:0    16:1    # 旦
那さま、あいつは私に、おまえの為すことを速かに為せと言いました。
-1      1:52    2:0     3:1     4:1     5:0     6:1     7:26    8:0     9:1     10:1    11:0    12:0    13:0    14:1    15:1    16:0    # 私
はすぐに料亭から走り出て、夕闇の道をひた走りに走り、ただいまここに参りました。
1       1:26    2:0     3:1     4:1     5:0     6:0     7:17    8:0     9:1     10:1    11:0    12:0    13:0    14:0    15:1    16:0    # そ
うして急ぎ、このとおり訴え申し上げました。
-1      1:17    2:0     3:1     4:1     5:0     6:0     7:18    8:0     9:1     10:1    11:0    12:0    13:0    14:0    15:0    16:0    # さ
あ、あの人を罰して下さい。
-1      1:18    2:0     3:1     4:1     5:0     6:0     7:27    8:0     9:0     10:1    11:0    12:0    13:0    14:0    15:0    16:0    # ど
うとも勝手に、罰して下さい。
```

判定するクラス
$y^{(i)}$

素性番号:値
$x^{(i)}_j$

訓練結果（モデル）

```
SVM-light Version V6.02
0 # kernel type
3 # kernel parameter -d
1 # kernel parameter -g
1 # kernel parameter -s
1 # kernel parameter -r
empty# kernel parameter -u
16 # highest feature index
392 # number of training documents
247 # number of support vectors plus 1
1.0005788 # threshold b, each following line is a SV (starting with alpha*y)
-0.00039746743794739382554109807799136 1:6 2:0 3:0 4:1 5:0 6:0 7:37 8:0 9:1 10:1 11:0 12:1
13:1 14:1 15:0 16:0 # 火と水と。
0.00039746743794739382554109807799136 1:24 2:0 3:1 4:1 5:0 6:0 7:91 8:0 9:1 10:1 11:0 12:1
13:0 14:1 15:1 16:1 # その人は、ずいぶん不仕合せな男なのです。
-0.00039746743794739382554109807799136 1:13 2:0 3:0 4:1 5:0 6:0 7:172 8:0 9:1 10:1 11:0 12:1
13:0 14:1 15:0 16:1 # 六日まえのことでした。
0.00039746743794739382554109807799136 1:9 2:0 3:0 4:1 5:0 6:0 7:6 8:0 9:0 10:1 11:0 12:0 13:1
14:0 15:0 16:0 # 生かして置けねえ。
-0.00039746743794739382554109807799136 1:155 2:0 3:1 4:1 5:0 6:1 7:22 8:0 9:0 10:1 11:0 12:0
13:0 14:1 15:1 16:0 # 大群集、老いも若きも、あの人のあとにつき従い、やがて、エルサレムの宮が間近になったころ、あの人は、一匹の老いぼれた驢馬の子に乗りて来り給う」と予言されてある通りの形なのだと、弟子たちに晴れがましい顔をして教えましたが、私ひとりは、なんだか浮かぬ気持でありました。
0.00039746743794739382554109807799136 1:45 2:0 3:1 4:1 5:0 6:1 7:223 8:0 9:1 10:1 11:0 12:1
13:1 14:0 15:0 16:2 # あの人ひとりに心を捧げ、これ迄どんな女にも心を動かしたことは無いのだ。
0.00039746743794739382554109807799136 1:223 2:0 3:1 4:1 5:0 6:1 7:15 8:0 9:0 10:1 11:0 12:0
13:0 14:1 15:0 16:1 # マルタの妹のマリヤは、姉のマルタが骨組頑丈で牛のように大きく、気象も荒く、どたばた立ち働くのだけが取柄で、なんの見どころも無い百姓女でありますが、あれは違って骨も細く、皮膚は透きとおる程の青白さで、手足もふっくらして小さく、湖水のように深く澄んだ大きい眼が、いつも夢みるように、うっとり遠くを眺めていて、あの村では皆、不思議がっているほどの気高い娘でありました。
0.00039746743794739382554109807799136 1:8 2:0 3:1 4:1 5:0 6:0 7:88 8:0 9:1 10:1 11:0 12:1 13:1
14:1 15:1 16:1 # それにきまった。
0.00039746743794739382554109807799136 1:5 2:0 3:0 4:1 5:0 6:0 7:16 8:0 9:0 10:1 11:0 12:0 13:0
14:0 15:1 16:0 # 主です。
```

まとめ

- 概要
  - 特徴ベクトル
  - 訓練データ
  - 概要
- 評価
  - 再現率、適合率
  - オープンテスト、クローズテスト
  - 交差検定
- いくつか紹介
  - 決定木学習
  - 回帰分析
  - SVM
- 多クラス分類
- 実例：SVM^{light}

課題

下の表を見て、天候W、温度T、湿度H、強風Sから「開催」、「中止」を決定する式を求めよ。
条件判定、論理演算（＜、＞、＝等）が少ないものがより良い解とする。

天候	温度（° F）	湿度（%）	強風	クラス
晴れ	75	70	真	開催
晴れ	80	90	真	中止
晴れ	85	85	偽	中止
晴れ	72	95	偽	中止
晴れ	69	70	偽	開催
曇り	72	90	真	開催
曇り	83	78	偽	開催
曇り	64	65	真	開催
曇り	81	75	偽	開催
雨	71	80	真	中止
雨	65	70	真	中止
雨	75	80	偽	開催
雨	68	80	偽	開催
雨	70	96	偽	開催