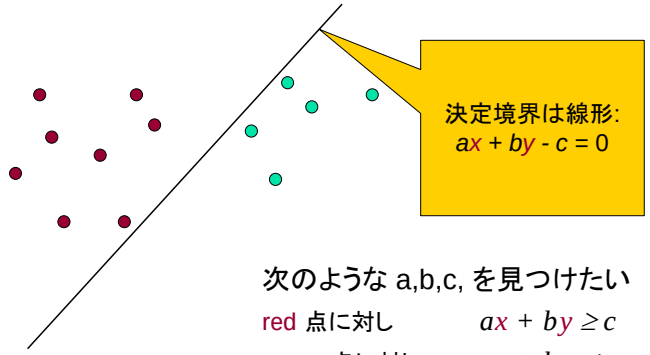
情報意味論(7)
サポートベクターマシン

理工学部管理工学科
櫻井彰人

基礎的復習: 線形判別関数

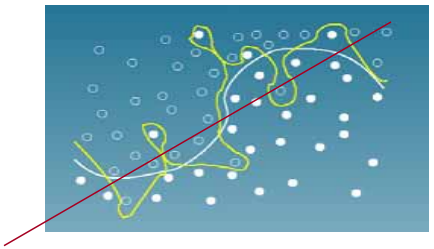


次のような a,b,c, を見つけたい

red 点に対し $ax + by \geq c$

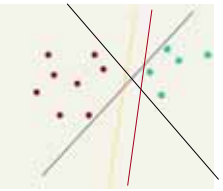
green 点に対し $ax + by \leq c.$

これも復習: 複雑な境界は？



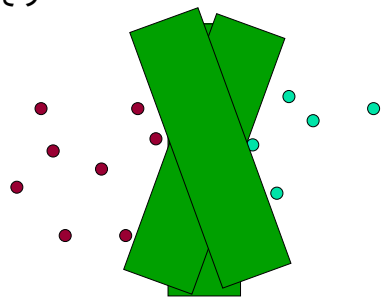
どの超平面を選ぶべきか?

- *a,b,c* にはいくつもの可能性あり
- 見つけたどれもが最良なわけではない
  [何か「よさ」の基準を設ける必要はある]
  - 例, パーセプトロン学習アルゴリズム
- サポートベクターマシンは最良のものをみつける.
  - 超平面とそれに近い「困難点」との距離を最大化する
  - 直感的解釈: 決定境界に近いところに(別のクラスの)点がなければ、決定の不確実さは少なかろう



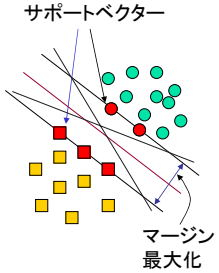
直感的解釈をもう一つ

- 分離境界を幅のある帯に置き換えてみよう。選択範囲がせばまり、汎化誤差の減少につながりそう



サポートベクターマシン (SVM)

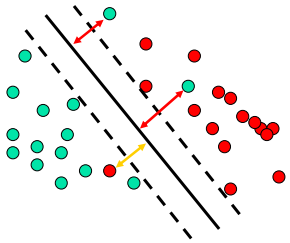
- SVM は、分離超平面周囲のマージンを最大化する.
  - ラージマージン分類器ともいう
- 決定関数はサポートベクターと呼ばれる訓練データによって完全に定まる.
- 2次計画問題である
- 広範囲の問題に対してうまくいく方法であると考えられている



ラージマージン分類器

線形分離可能でないならば

- 誤りを許す
  - コストを払って、本来あるべき場所に動かす
- ただ, 超平面はどちらのクラスからも遠ざける

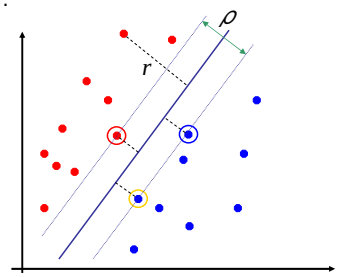


最大マージン: 定式化

- w: 決定超平面への垂線ベクトル
- x_i: i 番目のデータ点
- y_i: 属するクラス (+1 or -1)　注: 1/0 ではない
- 分類器: $\mathrm{sign}(w^T x_i + b)$
- そのとき x_i の関数マージン: $y_i (w^T x_i + b)$
  - 勿論 w を大きくすればマージンは増大する、そこで、、、
- 訓練データ全体の関数マージンは、上記の値の最大

幾何的マージン

- データ点から分離超平面までの距離 $r = \frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$
- 分離超平面に最も近い点がサポートベクター.
- 分離超平面のマージン ρ は別クラスのサポートベクターがどの程度分離しているかを示す.



線形 SVM を数学的に

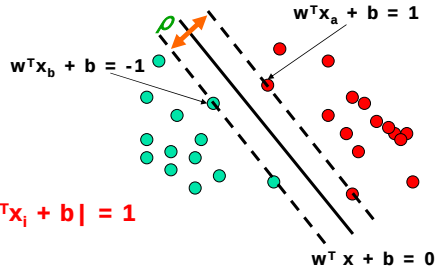
- 全ての点が超平面から距離 1 離れていると仮定しよう. そうであれば次の2つの制約が訓練データ集合 $\{(\mathbf{x}_i, y_i)\}$ から得られる

$$\mathbf{w}^T\mathbf{x}_i + b \geq 1 \quad \text{if } y_i = 1$$
$$\mathbf{w}^T\mathbf{x}_i + b \leq -1 \quad \text{if } y_i = -1$$

- サポートベクターに対しては、上記不等式は等式となる; そうなると, 各データの超平面からの距離は $r = \frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$ であるから、マージンは次の値となる: $\rho = \frac{2}{\|\mathbf{w}\|}$

線形サポートベクターマシン

- **超平面**
  $\mathbf{w}^T\mathbf{x} + b = 0$
- **制約**:
  $\min_{i=1,\ldots,n} |\mathbf{w}^T\mathbf{x}_i + b| = 1$
- 書換えると:
  $\mathbf{w}^T(\mathbf{x}_a - \mathbf{x}_b) = 2$
  $\rho = \|\mathbf{x}_a - \mathbf{x}_b\|_2 = 2/\|\mathbf{w}\|_2$



線形サポートベクターマシン

- 次の2次計画問題が得られる:

次のような $\mathbf{w}$ と b を見出せ:
$\rho = \frac{2}{\|\mathbf{w}\|}$ は最大であり, 全ての $\{(\mathbf{x}_i, y_i)\}$ につき
$\mathbf{w}^T\mathbf{x}_i + b \geq 1$ if $y_i=1$; $\mathbf{w}^T\mathbf{x}_i + b \leq -1$ if $y_i = -1$

- よりよい定式化 ($\min \|\mathbf{w}\| = \max 1/\|\mathbf{w}\|$):

次のような $\mathbf{w}$ と b を見出せ:
$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$ は最小であり, すべての $\{(\mathbf{x}_i, y_i)\}$ につき
$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

最適化問題の解法

次のような $\mathbf{w}$ と b とを見出せ
最小化: $\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$;
全ての $\{(\mathbf{x_i}, y_i)\}$ につき: $y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$

- 線形制約のもとでの2次関数の最適化
- 2次計画問題は、よく知られた数理計画問題の一つ. 多くの解法が知られている
- 解法にあたっては、ラグランジュ乗数 α_i を主問題の各制約に割付けた双対問題を構成する:

次のような $\alpha_1 \ldots \alpha_N$ を見出せ
最大化: $\mathbf{Q}(\boldsymbol{\alpha}) = \sum \alpha_i - \frac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}$;
(1) $\sum \alpha_i y_i = 0$
(2) $\alpha_i \geq 0$, 任意の α_i

最適化問題の解法

- 解の形は:

$\mathbf{w} = \sum \alpha_i y_i \mathbf{x_i}$, かつ $\alpha_k \neq 0$ なるすべての $\mathbf{x_k}$ につき $b = y_k - \mathbf{w}^T\mathbf{x_k}$

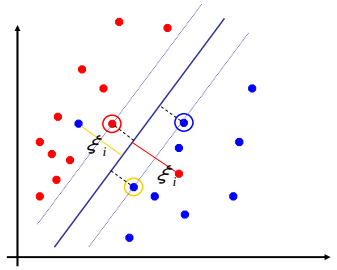
- 各非零の α_i は、対応する $\mathbf{x_i}$ がサポートベクターであることを示す.
- 識別関数は次のようになる:

$$f(\mathbf{x}) = \sum \alpha_i y_i \mathbf{x_i}^T\mathbf{x} + b$$

- 当該式は新規点とサポートベクトル $\mathbf{x_i}$ の**内積**であることに注意.
- また、当該最適化問題を解くには、訓練データのすべての組合せに関する内積 $\mathbf{x_i}^T\mathbf{x_j}$ の計算が含まれていることを注意しておく.

ソフトマージン分類器

- もし訓練データが線形分離可能でなければ, スラック変数 ξ_i を用いて分類が難しい点やノイズがのった点の誤分類を許すようにする.



ソフトマージン分類器の数学

- 以前の定式化:

次のような $\mathbf{w}$ と b とを見出せ
最小化: $\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$;
すべての $\{(\mathbf{x_i}, y_i)\}$ について: $y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$

- スラック変数を含む、新しい定式化:

次のような $\mathbf{w}$ と b とを見出せ
最小化: $\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum \xi_i$;
すべての $\{(\mathbf{x_i}, y_i)\}$ について: $y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i$, かつ
すべての i について: $\xi_i \geq 0$

- パラメータ C は過学習を制御する方法と見ることができる.

ソフトマージン分類器 – 解

- ソフトマージン分類器の双対問題:

次のような $\alpha_1 \ldots \alpha_N$ を見出せ:
最大化: $\mathbf{Q}(\boldsymbol{\alpha}) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$; ただし
(1) $\sum \alpha_i y_i = 0$
(2) すべての α_i につき $0 \leq \alpha_i \leq C$

- スラック変数 ξ_i もラグランジュ乗数も、双対問題には表れていない！
- 再び, 非零の α_i に対応する $\mathbf{x}_i$ はサポートベクターである.
- 当該双対問題への解は:

$\mathbf{w} = \sum \alpha_i y_i \mathbf{x}_i$
$b = y_k(1 - \xi_k) - \mathbf{w}^T \mathbf{x}_k$ where $k = \underset{k}{\operatorname{argmax}}\ \alpha_k$

明示的には w がなくても分類できる!

$f(\mathbf{x}) = \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$

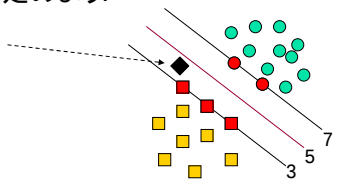
SVMを用いた分類

- 所与の新点 (*x1,x2*) に対し, その超平面への垂直射影を計る(scoreとしよう):
  - 2次元の場合: score = $w_1 x_1 + w_2 x_2 + b$.
  - すなわち: score = $wx + b = \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$
  - 信頼限度 t を定めよう.

score > t : yes

score < −t : no

それ以外: 判定放棄



線形 SVM: まとめ

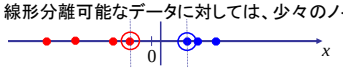
- 分類器は、分離超平面 *separating hyperplane.*
- 最も重要な訓練データ点がサポートベクターとなる; それが当該超平面を決める.
- 2次計画問題を解けば、どの点 $\mathbf{x}_i$ がサポートベクターで非零のラグランジュ乗数 α_i に対応するかが分かる.
- 当該問題の双対問題においても解法においても、訓練データ点は、内積の中にしか現れない:

次のような $\alpha_1 \ldots \alpha_N$ を見出せ:
最大化: $\mathbf{Q}(\boldsymbol{\alpha}) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$,
但し
(1) $\sum \alpha_i y_i = 0$
(2) すべての α_i につき: $0 \leq \alpha_i \leq C$

$f(\mathbf{x}) = \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$

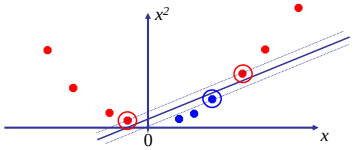
非線形 SVM

- 線形分離可能なデータに対しては、少々のノイズがあっても、うまくいく:



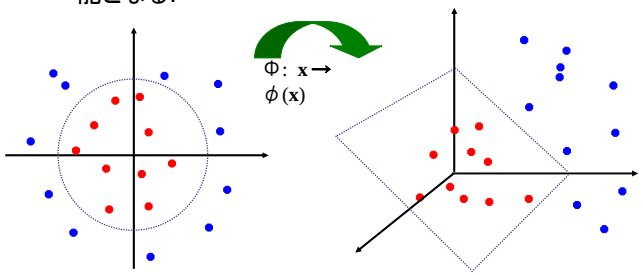
- しかし、データ集合が線形分離可能でなかったらどうしよう?

- 例えば… データをより高次元の空間に写像したらどうだろうか:



非線形 SVM: 特徴空間

- 一般的なアイデア: もともとの特徴空間は、いつでも、ある高次元特徴空間に写像すれば、線形分離可能となる:



カーネルトリック “Kernel Trick”

- 線形分類器が依拠していたのは、ベクター間の内積 $K(\mathbf{x}_i,\mathbf{x}_j)=\mathbf{x}_i^T\mathbf{x}_j$
- もし各点を高次元空間に、変換 $\Phi: \mathbf{x} \rightarrow \phi(\mathbf{x})$ を用いて写像すると, 内積は:

$$K(\mathbf{x}_i,\mathbf{x}_j)= \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$$

- **カーネル関数**は、変換後の内積の値が、変換前の内積の関数となるようなもの.
- 例:

2-次元ベクトル $\mathbf{x}=[x_1\ x_2]$ に対し $K(\mathbf{x}_i,\mathbf{x}_j)=(1+\mathbf{x}_i^T\mathbf{x}_j)^2$ とおく
このとき、次の式が成立する $K(\mathbf{x}_i,\mathbf{x}_j)= \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$:

$$K(\mathbf{x}_i,\mathbf{x}_j)=(1+\mathbf{x}_i^T\mathbf{x}_j)^2 = 1+ x_{i1}^2x_{j1}^2 + 2\,x_{i1}x_{j1}\,x_{i2}x_{j2} + x_{i2}^2x_{j2}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2} =$$
$$= [1\ \ x_{i1}^2\ \sqrt{2}\,x_{i1}x_{i2}\ \ x_{i2}^2\ \sqrt{2}x_{i1}\ \sqrt{2}x_{i2}]^T\,[1\ \ x_{j1}^2\ \sqrt{2}\,x_{j1}x_{j2}\ \ x_{j2}^2\ \sqrt{2}x_{j1}\ \sqrt{2}x_{j2}]$$
$$= \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \qquad \text{ただし } \phi(\mathbf{x}) = [1\ \ x_1^2\ \sqrt{2}\,x_1x_2\ \ x_2^2\ \ \sqrt{2}x_1\ \sqrt{2}x_2]$$

カーネル関数

- なぜカーネルを用いるか?
  - 分離可能でないものを分離可能にする.
  - データをより適切な表現空間に写像する
- よく使われるカーネル
  - 線形
  - 多項式 $K(x,z) = (1+x^Tz)^d$
  - RBF Radial basis function

$$K(\mathbf{x}_i,\mathbf{x}_j) = e^{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2/2\sigma^2}$$

SVM: 汎化能力の推定

- 汎化能力最大(新規データに対して最も正確)の分類器がほしい.
- 良い汎化性能を得るための糸口は?
  - 訓練データを大きくする
  - 訓練データに対する誤りを小さくする
  - 容量/分散 (モデル記述パラメータ数, モデルの表現能力) をおおきくする
- SVM では、これらの量に基づいて、新規データに対する誤差限界を明示的に示すことができる.

容量/分散: VC 次元

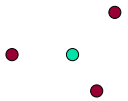
- 理論的なリスク限界:

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}}$$

- Risk = 平均誤り率
- α – 当該モデル (パラメータで決まる)
- R_{emp} – 経験リスク, l – 観測数, h – VC 次元, 当該式は確率 $(1-\eta)$ で正しい
  - VC (Vapnik-Chervonenkis) 次元/容量: shatterできる点の最大数
  - ある点集合がshatterできるとは、その任意のラベル付けを当該分類器が行えること.
- 重要な理論的性質; 実際にはあまり使われない

演習

- d 次元空間に n 個の点があり、それらは、red か green とラベルが付けられていると仮定する. n を (d の関数として) どれだけ大きくとれば, red 点と green 点が線形分離でなくなる例が作れるか?
- 例, d =2 に対しては $n \ge 4$.



スケッチ: マージン最大化の理論的な正当化

- Vapnik は次のことを証明した:
  最適な線形判別器クラスの VC 次元 h は、次の上界をもつ

$$h \le \min\left\{\left\lceil \frac{D^2}{\rho^2} \right\rceil, m_0\right\} + 1$$

  ただし ρ はマージン, D は訓練事例をすべて囲い込む最小の超球の直径, そして m_0 は(事例の表現空間の)次元である.

- 直感的に, これは空間の次元 m_0 にかかわりなく, マージン ρ を最大化することにより、VC 次元を最小化することができる.
- こうして, 分類器の複雑度は、次元数に関わりなく小さく保つことができる.

SVM の性能

- SVM は、最良の性能を持つと考える人は多い.
- 多くの場合、統計的な有意性は明確ではない.
- SVM と同程度の性能をもつ手法は他にもある.
- 例: regularized logistic regression (Zhang & Oles)
  - Tong Zhang, Frank J. Oles: Text Categorization Based on Regularized Linear Classification Methods. Information Retrieval 4(1): 5-31 (2001)
- 比較研究の例: Yang & Liu
  - Yiming Yang, Xin Liu: A re-examination of text categorization methods, 22nd Annual International SIGIR (1999).

評価例: 古典的な Reuters データ

- 非常によく使われたデータセット
- 21578 documents
- 9603 training, 3299 test articles (ModApte split)
- 118 categories
  - 一つの article は複数の category に属しうる
  - 118 個の2値分類
- 1 document 当たりの category 数
  - 1.24
- 10 categories のみ大きい（全 118 categories）

大きめの categories (#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Reuters Text Categorization data set (**Reuters-21578)** document 例

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

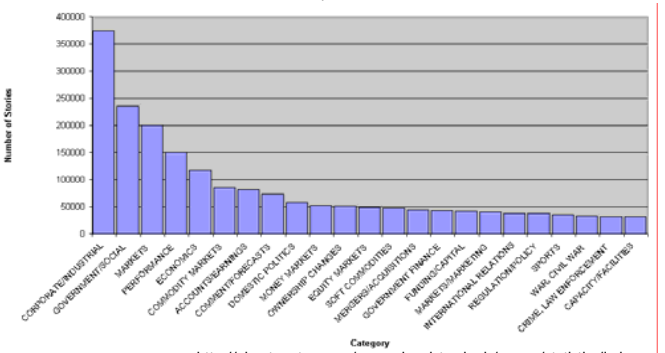
Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

New Reuters: RCV1: 810,000 文書

- Reuters RCV1 の頻出トピック



http://about.reuters.com/researchandstandards/corpus/statistics/index.asp

（クラス当たりの）評価尺度

- Recall: クラス *i* のdocument中、正しく *i* に分類されたものの割合:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

- Precision:クラス *i* に分類されたdocument中、本当にクラス *i* に属するものの割合:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

- “Correct rate”: (1- error rate) 正しく分類されたdocumentの割合:

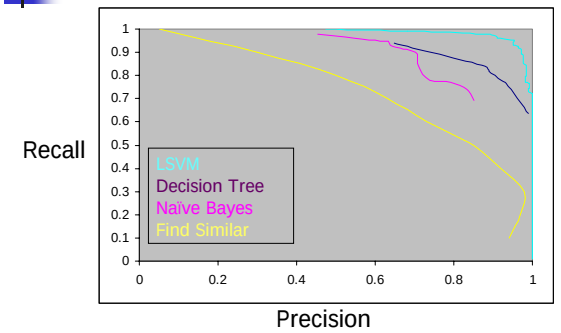
$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Dumais et al. 1998: Reuters – Break-Even Performance

	Rocchio	NBayes	Trees	LinearSVM	
earn	92.9%	95.9%	97.8%	98.2%	
acq	64.7%	87.8%	89.7%	92.8%	
money-fx	46.7%	56.6%	66.2%	74.0%	
grain	67.5%	78.8%	85.0%	92.4%	
crude	70.1%	79.5%	85.0%	88.3%	
trade	65.1%	63.9%	72.5%	73.5%	
interest	63.4%	64.9%	67.1%	76.3%	
ship	49.2%	85.4%	74.2%	78.0%	
wheat	68.9%	69.7%	92.5%	89.7%	
corn	48.2%	65.3%	91.8%	91.1%	
Avg Top 10	64.6%	81.5%	88.4%	91.4%	
Avg All Cat	61.7%	75.2%	na	86.4%	

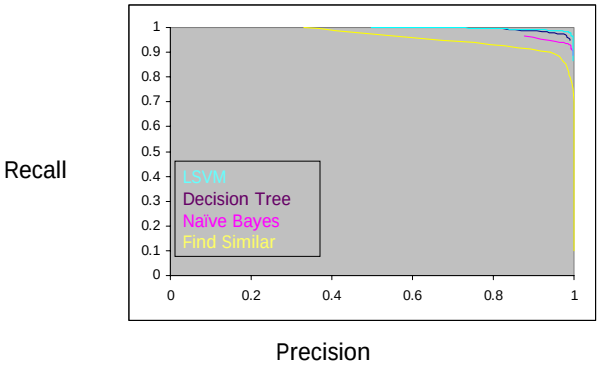
Break Even: (Recall + Precision) / 2

Precision vs. Recall - Category “Grain”

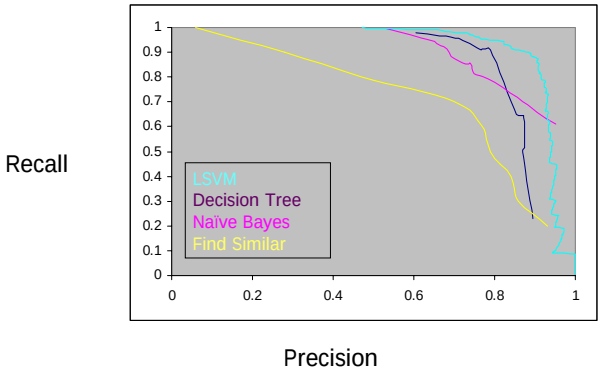


Recall: = TP/(TP+TN); % 当該カテゴリ中そのカテゴリに属すると判定したもの
Precision: = TP/(TP+FP); % そのカテゴリに属するとした中で本当にそのカテゴリに属するもの

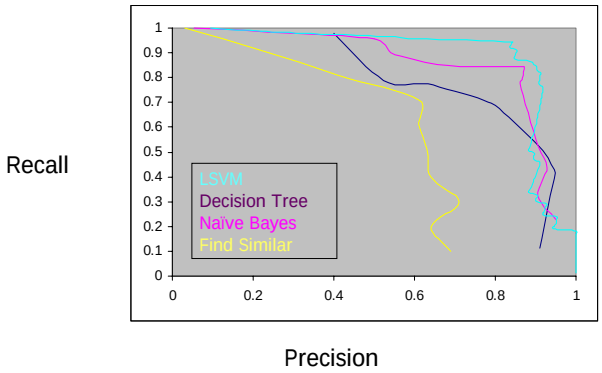
Precision vs. Recall - Category “Earn”



Precision vs. Recall - Category “Crude”



Precision vs. Recall - Category “Ship”



カーネルによる違い (Joachims)

	Bayes	Rocchio	C4.5	k-NN	SVM (poly) degree d = 1	2	3	4	5	SVM (rbf) width γ = 0.6	0.8	1.0	1.2
earn	95.9	96.1	96.1	97.3	98.2	98.4	**98.5**	98.4	98.3	**98.5**	98.5	98.4	98.3
acq	91.5	92.1	85.3	92.0	92.6	94.6	**95.2**	95.2	95.3	95.0	95.3	95.3	**95.4**
money-fx	62.9	67.6	69.4	78.2	66.9	72.5	75.4	74.9	**76.2**	74.0	75.4	**76.3**	75.9
grain	72.5	79.5	89.1	82.2	91.3	93.1	**92.4**	91.3	89.9	**93.1**	91.9	91.9	90.6
crude	81.0	81.5	75.5	85.7	86.0	87.3	88.6	**88.9**	87.8	**88.9**	89.0	88.9	88.2
trade	50.0	77.4	59.2	77.4	69.2	75.5	76.6	77.3	**77.1**	76.9	78.0	**77.8**	76.8
interest	58.0	72.5	49.1	74.0	69.8	63.3	67.9	73.1	**76.2**	74.4	75.0	**76.2**	76.1
ship	78.7	83.1	80.9	79.2	82.0	85.4	86.0	**86.5**	86.0	**85.4**	86.5	87.6	87.1
wheat	60.6	79.4	85.5	76.6	83.1	84.5	85.2	**85.9**	83.8	**85.2**	85.9	85.9	85.9
corn	47.3	62.2	87.7	77.9	86.0	86.5	85.3	**85.7**	83.9	**85.1**	85.7	85.7	84.5
microavg.	**72.0**	**79.9**	**79.4**	**82.3**	84.2	85.1	85.9	86.2	85.9	86.1	86.5	86.3	86.2
					combined: **86.0**					combined: **86.4**			

T. Joachims, Learning to Classify Text using Support Vector Machines. Kluwer, 2002.

Yang&Liu: SVM vs 他の手法

Table 1: Performance summary of classifiers

method	miR	miP	miF1	maF1	error
SVM	.8120	.9137	.8599	.5251	.00365
KNN	.8339	.8807	.8567	.5242	.00385
LSF	.8507	.8489	.8498	.5008	.00414
NNet	.7842	.8785	.8287	.3765	.00447
NB	.7688	.8245	.7956	.3886	.00544

miR = micro-avg recall; miP = micro-avg prec.;
miF1 = micro-avg F1; maF1 = macro-avg F1.

まとめ

- サポートベクターマシン (SVM) は
  - サポートベクターに基づいて超平面を決める
    - Support vector = 判定境界付近のクリティカルな点
  - 線形 SVM は線形分類器.
  - カーネル: 高次元へ写像するが、その内積は低次元の内積で簡単に計算できる
  - リスクの上界 (リスク = テストデータでの期待誤り)
  - (邪魔な属性が多いときの)分類器としてベスト?
  - ポピュラー: SVMlight がきっかけ?
    - 速くて無料 (研究目的には)
  - 他にもいくつか: TinySVM, libsvm, ….

参考

- A Tutorial on Support Vector Machines for Pattern Recognition (1998) Christopher J. C. Burges
- S. T. Dumais, Using SVMs for text categorization, IEEE Intelligent Systems, 13(4):21-23, Jul/Aug 1998
- S. T. Dumais, J. Platt, D. Heckerman and M. Sahami. 1998. Inductive learning algorithms and representations for text categorization. *Proceedings of CIKM '98*, pp. 148-155.
- A re-examination of text categorization methods (1999) Yiming Yang, Xin Liu 22nd Annual International SIGIR
- Tong Zhang, Frank J. Oles: Text Categorization Based on Regularized Linear Classification Methods. Information Retrieval 4(1): 5-31 (2001)
- Trevor Hastie, Robert Tibshirani and Jerome Friedman, "Elements of Statistical Learning: Data Mining, Inference and Prediction" Springer-Verlag, New York.
- 'Classic' Reuters data set: http://www.daviddlewis.com /resources /testcollections/reuters21578/
- T. Joachims, *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.