言語処理のための機械学習入門

4. 分類
4.1 準備
4.2 ナイーブベイズ分類器
4.3 サポートベクトルマシン

河野和平

クラスタリングと分類

- クラスタリング
  - 何らかの意味で似ているものをグループ化
  - どの様なクラスタができるかは分からない

- 分類
  - あらかじめ決まったグループに分ける

    例

    電子メールを「仕事関係」「プライベート」「スパム」に分類
  - グループをクラス(カテゴリ)と呼ぶ

ラベル付きデータ

- データ集合

$$D = \left\{\left(d^{(1)}, c^{(1)}\right), \left(d^{(2)}, c^{(2)}\right), \cdots, \left(d^{(|D|)}, c^{(|D|)}\right)\right\}$$

$d^{(n)}$: 事例　　$c^{(n)}$: 事例の属するクラス（ラベル）

  - ラベル付きデータ（⇔ラベル無しデータ）
  - 教師付き学習（⇔教師無し学習）

ナイーブベイズ分類器

- 事例 d に対して、$P(c|d)$ が最大となるクラス c を求める。

$$c_{max} = \arg\max_c \frac{P(c)P(d|c)}{P(d)}$$

$$= \arg\max_c P(c)P(d|c)$$

起こり得るdは膨大で
すべてのdについて$P(d|c)$ を求めるのは困難

多変数ベルヌーイモデル

- 語彙 V に含まれる各単語 w とクラス c

- ベルヌーイ分布に従う確率変数 $X_{w,c}$
  - w が事例内に出現するとき1,そうでないなら0
  - $p_{w,c}$: $X_{w,c}$ が1になる確率

- 文書 d の生起確率

$$P(d|c) = \prod_{w \in V} {p_{w,c}}^{\delta_{w,c}} \left(1 - p_{w,c}\right)^{1-\delta_{w,c}}$$

多変数ベルヌーイモデル

- ナイーブベイズ分類器

$$P(c)P(d|c) = p_c \prod_{w \in V} {p_{w,c}}^{\delta_{w,c}} \left(1 - p_{w,c}\right)^{1-\delta_{w,c}}$$

を最大化する c を求める。

- ◆ $P(d|c)$ は単語の組合せ
- ◆ $p_{w,c}$ は一つの単語にのみ注目

多変数ベルヌーイモデル

- パラメータの最尤推定

$$\log P(D) = \sum_{(d,c)\in D} \log\left(p_c \prod_{w\in V} {p_{w,c}}^{\delta_{w,c}} \left(1-p_{w,c}\right)^{1-\delta_{w,c}} \right)$$

$$= \sum_{(d,c)\in D} \left(\log p_c + \sum_{w\in V} \delta_{w,c} \log p_{w,c} + \left(1-\delta_{w,c}\right) \log\left(1-p_{w,c}\right) \right)$$

$$= \sum_{c} N_c \log p_c + \sum_{c} \sum_{w\in V} N_{w,c} \log p_{w,c} + \sum_{c} \sum_{w\in V} \left(N_c - N_{w,c}\right) \log\left(1-p_{w,c}\right)$$

N_c : クラス c である訓練文書数

$N_{w,c}$: クラス c かつ w を含む訓練文書数

多変数ベルヌーイモデル

- 制約付き最適化問題

$$max. \log P(D)$$
$$s.t. \sum_c p_c = 1$$

- パラメータ

$$p_{w,c} = \frac{N_{w,c}}{N_c} = \frac{\text{クラス}c\text{に属する訓練文書のうち}w\text{を含む文書数}}{\text{クラス}c\text{に属する訓練文書数}}$$

$$p_c = \frac{N_c}{\sum_c N_c} = \frac{\text{クラス}c\text{に属する訓練文書数}}{\text{訓練文書数}}$$

多変数ベルヌーイモデル

- 分類器の構築

  例

  P氏の文書

  $$d^{(1)} = "good\ bad\ good\ good"$$
  $$d^{(2)} = "exciting\ exciting"$$
  $$d^{(3)} = "good\ good\ exciting\ boring"$$

  N氏の文書

  $$d^{(4)} = "bad\ boring\ boring\ boring"$$
  $$d^{(5)} = "bad\ good\ bad"$$
  $$d^{(6)} = "bad\ bad\ boring\ exciting"$$

  語彙 $V = \{bad, boring, exciting, good\}$

多変数ベルヌーイモデル

- 分類器の構築

$$p_{\mathrm{P}} = \frac{N_{\mathrm{P}}}{N_{\mathrm{P}} + N_{\mathrm{N}}} = \frac{3}{3+3} = 0.50, \quad p_{\mathrm{N}} = \frac{N_{\mathrm{N}}}{N_{\mathrm{P}} + N_{\mathrm{N}}} = \frac{3}{3+3} = 0.50,$$

$$p_{\mathrm{bad,P}} = \frac{N_{\mathrm{bad,P}}}{N_{\mathrm{P}}} = \frac{1}{3} = 0.33, \quad p_{\mathrm{bad,N}} = \frac{N_{\mathrm{bad,N}}}{N_{\mathrm{N}}} = \frac{3}{3} = 1.00,$$

$$p_{\mathrm{boring,P}} = \frac{N_{\mathrm{boring,P}}}{N_{\mathrm{P}}} = \frac{1}{3} = 0.33, \quad p_{\mathrm{boring,N}} = \frac{N_{\mathrm{boring,N}}}{N_{\mathrm{N}}} = \frac{2}{3} = 0.67,$$

$$p_{\mathrm{exciting,P}} = \frac{N_{\mathrm{exciting,P}}}{N_{\mathrm{P}}} = \frac{2}{3} = 0.67,$$

$$p_{\mathrm{exciting,N}} = \frac{N_{\mathrm{exciting,N}}}{N_{\mathrm{N}}} = \frac{1}{3} = 0.33,$$

$$p_{\mathrm{good,P}} = \frac{N_{\mathrm{good,P}}}{N_{\mathrm{P}}} = \frac{2}{3} = 0.67, \quad p_{\mathrm{good,N}} = \frac{N_{\mathrm{good,N}}}{N_{\mathrm{N}}} = \frac{1}{3} = 0.33.$$

多変数ベルヌーイモデル

- 分類の例1

$$d = \text{"good good bad boring"}$$

$$P(P)P(d|P) = p_p \times p_{bad,P} \times p_{boring,P} \times \left(1 - p_{exciting,P}\right) \times p_{good,P}$$
$$= 0.012$$

$$P(N)P(d|N) = p_N \times p_{bad,N} \times p_{boring,N} \times \left(1 - p_{exciting,N}\right) \times p_{good,N}$$
$$= 0.074$$

⇒ N氏によって書かれたと予測

多変数ベルヌーイモデル

- 分類器の構築

  例

  P氏の文書

  $d^{(1)} = "good\ bad\ good\ good"$ ⇒ $"good\ bad\ good\ good\ fine"$

  $d^{(2)} = "exciting\ exciting"$

  $d^{(3)} = "good\ good\ exciting\ boring"$

  N氏の文書

  $d^{(4)} = "bad\ boring\ boring\ boring"$

  $d^{(5)} = "bad\ good\ bad"$

  $d^{(6)} = "bad\ bad\ boring\ exciting"$

  語彙 $V = \{bad, boring, exciting, good, fine\}$

多変数ベルヌーイモデル

- 分類器の構築

$$p_{\text{fine,P}} = \frac{N_{\text{fine,P}}}{N_{\text{P}}} = \frac{1}{3} = 0.33, \quad p_{\text{fine,N}} = \frac{N_{\text{fine,N}}}{N_{\text{N}}} = \frac{0}{3} = 0.00.$$

- 分類の例2

$d = "bad\ bad\ boring\ boring\ fine"$ ⇐ fine以外はN氏の文章と似ている

$$P(P)P(d|P) = p_p \times p_{bad,P} \times p_{boring,P} \times \left(1 - p_{exciting,P}\right) \times p_{fine,P} \times \left(1 - p_{good,P}\right)$$
$$= 0.002$$
$$P(N)P(d|N) = p_N \times p_{bad,N} \times p_{boring,N} \times \left(1 - p_{exciting,N}\right) \times p_{fine,N} \times \left(1 - p_{good,N}\right)$$
$$= 0.00$$

⇒ P氏によって書かれたと予測

多変数ベルヌーイモデル

- MAP推定
  - すべてのパラメータが0にならないよう推定する。
  - 0に近い値を取る確率が小さい事前分布をパラメータに与える(ディリクレ分布)

目的関数

$$\log P(\theta) + \log P(D)$$
$$= \log\left(\prod_c p_c^{\alpha-1}\right) \times \left(\prod_c \left(p_{w,c}^{\alpha-1}(1-p_{w,c})^{\alpha-1}\right)\right) + \log P(D) + \text{定数}$$

多変数ベルヌーイモデル

- MAP推定

  目的関数

$$\log P(\theta) + \log P(D)$$

$$= \log\left(\prod_c p_c^{\alpha-1}\right) \times \left(\prod_c \left(p_{w,c}^{\alpha-1}(1-p_{w,c})^{\alpha-1}\right)\right) + \log P(D) + \text{定数}$$

$\alpha = 2$ のとき

$$p_{w,c} = \frac{N_{w,c}+1}{N_c+2} \qquad p_c = \frac{N_c+1}{\sum_c N_c + |C|}$$

多項モデル

- 多変数ベルヌーイモデル
  - 各単語が起こるか起こらないかをモデル化

- 多項モデル
  - 文書の各位置でどんな単語が起こるかをモデル化
  - 語彙 V の中から1単語を選ぶ操作を $|d|$ 回繰り返す

多項モデル

- 文書d 内で単語w が $n_{w,d}$ 回起こる確率

$$\frac{\left(\sum_w n_{w,d}\right)!}{\prod_{w\in V} n_{w,d}!}\prod_{w\in V} q_{w,c}^{n_{w,d}}$$

$q_{w,c}$ ：
クラスcとしたとき
単語wが選ばれる確率

$$P(d|c) = P\left(K=\sum_w n_{w,d}\right)\frac{\left(\sum_w n_{w,d}\right)!}{\prod_{w\in V} n_{w,d}!}\prod_{w\in V} q_{w,c}^{n_{w,d}}$$

- 最大化問題

$$\arg\max_c P(c)P(d|c) = \arg\max_c p_c P\left(K=\sum_w n_{w,d}\right)\frac{\left(\sum_w n_{w,d}\right)!}{\prod_{w\in V} n_{w,d}!}\prod_{w\in V} q_{w,c}^{n_{w,d}}$$

$$= \arg\max_c p_c \prod_{w\in V} q_{w,c}^{n_{w,d}}$$

多項モデル

- 制約付き最適化問題

$$max. \log P(D)$$

$$s.t. \sum_c p_c = 1$$

$$\sum_{w \in V} q_{w,c} = 1$$

- パラメータ

$$q_{w,c} = \frac{n_{w,c}}{\sum_w n_{w,c}} = \frac{\text{クラス}c\text{に属する訓練文書全体での}w\text{の出現回数}}{\text{クラス}c\text{に属する訓練文書全体での全単語の出現数}}$$

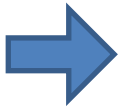
⇧ 文書内での単語の生起回数が考慮される

多項モデル

- MAP推定

  多変数ベルヌーイモデルと同様,

  一部のパラメータが0になり分類に悪影響を及ぼす。



ディリクレ分布を事前分布として用いることで

多変数ベルヌーイモデルと同様の手順で解決可能。

サポートベクトルマシン

- Support Vector Machine (SVM)
- 線形二値分類器
- 非常に高い分類性能を持つ

- 分類するクラスを正クラス，負クラスと呼ぶ。
- 正クラスに属する事例：正例
- 負クラスに属する事例：負例

サポートベクトルマシン

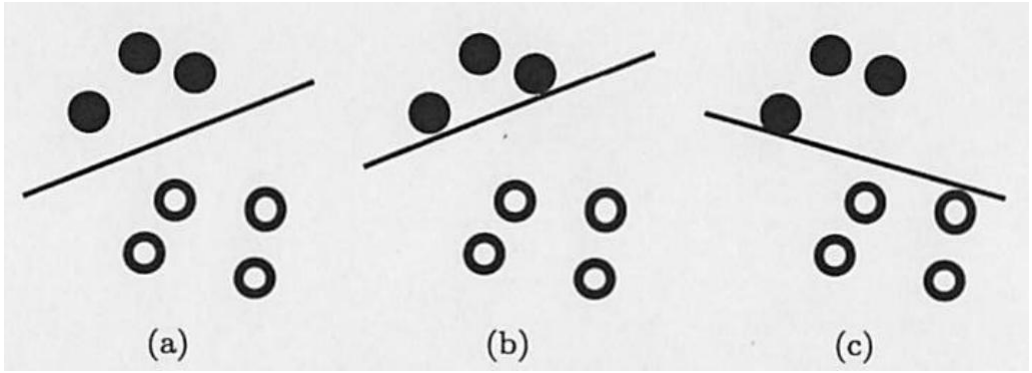
- 訓練データ

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

- 関数 $f(x) = \boldsymbol{w} \cdot \boldsymbol{x} - b$
  - 分類平面の方向ベクトル $\boldsymbol{w}$
  - 分離平面の切片 b
  - $f(x) \geq 0$なら正クラス,$f(x) < 0$ なら負クラスに分類

サポートベクトルマシン

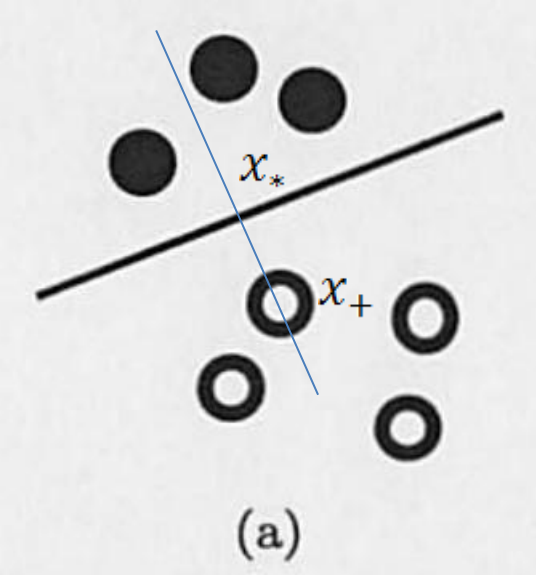
- マージン最大化



どちらのクラスからもなるべく遠い位置で分離したい
(マージン最大化)

サポートベクトルマシン

- マージン最大化



マージン: $|x_+ - x_*|$

分離平面: $w \cdot x = b$

分離平面の方向ベクトル: w

$$w \cdot x_* = b$$

$$w \cdot x_+ - b = 1$$

$$w \cdot (x_+ - x_*) = |w||x_+ - x_*|$$

$$|w||x_+ - x_*| = 1$$

サポートベクトルマシン

- マージン最大化

$$|w||x_+ - x_*| = 1$$

$$|x_+ - x_*| = \frac{1}{|w|}$$

$|w|$ の最少化で表わされる。
(絶対値は扱いにくいのでw^2の最少化を行う。)

厳密制約下のSVMモデル

- $y^{(i)} = +1$ (正クラス)であるような訓練事例

$$w \cdot x^{(i)} - b \geq 1$$

- $y^{(i)} = -1$(負クラス)であるような訓練事例

$$w \cdot x^{(i)} - b \leq -1$$

$$\Rightarrow \quad y^{(i)}(w \cdot x^{(i)} - b) \geq 1$$

- 最適化問題

$$\min. \ \frac{1}{2}w^2$$

$$s.t. \ y^{(i)}\left(w \cdot x^{(i)} - b\right) - 1 \geq 0$$

緩和制約下のSVMモデル

- 厳密制約下のSVMは実際のデータに対してうまく動かない。

- 訓練データの中には例外的な事例が存在する。
  - 分離平面に大きな影響を与える。

- 制約を緩めることで解決する。

緩和制約下のSVMモデル

- 緩和した制約

$$y^{(i)}\left(w \cdot x^{(i)} - b\right) - 1 \geq -\xi_i$$

ξ_i ：i 番目の訓練事例がうまく分けられない度合い

- 最適化問題

$$\min.\ \frac{1}{2}w^2 + C\sum_i \xi_i$$

$$s.t.\ y^{(i)}\left(w \cdot x^{(i)} - b\right) - 1 \geq -\xi_i$$

$$\xi_i \geq 0$$

関数距離

- SVMの分類

  事例 x

  $f(x) \geq 0$ なら正クラス

  $f(x) < 0$ なら負クラス

  関数距離

  $f(x) = 0.00001$
  $f(x) = 1000$ } どちらも正クラスに分類される

  $f(x)$ の値が大きい方が分類結果に誤りが少ない

多値分類器

- one-versus-rest法
  - 各クラスについて1つの分離平面を作る。
    (そのクラスに属するか否かを判別)
  - SVMによりn個の分離平面を求める。
  - 1つの事例で複数のクラスで正例となる場合
    関数距離が最も大きいクラスに分類する。

多値分類器

- ペアワイズ法
  - クラス対ごとに分離平面を作る
    (C_1,C_2,C_3ならC_1とC_2,C_2とC_3,C_1とC_3で3つの分離平面を作る)
    比較的多くの平面を作る
  - 事例を各分離平面で分離
    最も多く分類されたクラスに決定する。
  - ある対の分離平面を作るとき
    ２つのクラスに属さない訓練事例は利用しない

⇒ 比較的、訓練時間は少なくなることが多い。