4章 分類

08t4023y　金田昌侑

４，１準備

分類とはあらかじめ決まったグループに分けることであり、そのグループをクラス、またはカテゴリと呼ぶ。そしてその分類をする手段として分類器をつくる。

分類器の作り方のひとつに人間が分類規則を書く方法がある。規則ベース手法と呼ばれる。

$$D = \{(d^{(1)}, c^{(1)}), (d^{(2)}, c^{(2)}), \ldots, (d^{(|D|)}, c^{(|D|)})\} \qquad (4.1)$$

分類器を構築するには上記のようなデータ集合が与えられている必要がある。dは事例を表し、cは事例の属するクラスであるラベルを表す。集合Dのことをラベル付きデータと呼ぶ。それに対し、クラスタリングに用いたようなラベルのないデータはラベル無しデータと呼ぶ。

教師付き学習

ラベル付きデータを用いて行う学習。

教師無し学習

クラスタリングのように、ラベル無しデータを用いる学習

4.2 ナイーブベイズ分類器

事例dに対し、P(*c*|d)が最大となるc∈Cを出力する。ベイズの定理の次の式を利用する。

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \quad (4.2)$$

以下の右辺が最大となるクラスcを出力する。

$$c_{\max} = \arg \max_{c} \frac{P(c)P(d \mid c)}{P(d)} \quad (4.3)$$

$$= \arg \max_{c} P(c)P(d \mid c) \quad (4.4)$$

続き

式(4.4)の右辺を求めればよいが、P(d|c)は簡単に計算できないので、文書dを簡単化したモデルを仮定してP(d|c)の値を求める。そのモデルとして、多変数ベルヌーイモデルと多項モデルの2種類が使われる。

4.2.1 多変数ベルヌーイモデル

[1]モデルの導入

語彙Vに含まれる各単語wとクラスcにベルヌーイ分布に従う確率変数$X_{w,c}$を考える。

wが事例内で存在するとき1、しないときを0とする。$X_{w,c}$が1となる確率は$p_{w,c}$で表す。つまり$p_{w,c}=P(X_{w,c}=1)$となる。同時に$p_c=P(c)$とする。

クラスcが与えられている時のwの生起するかどうかの確率は

$$p_{w,c}^{\delta_{w,c}} (1 - p_{w,c})^{1-\delta_{w,d}}$$

で表される。

文書dの生起確率を数式にすると

$$p(d \mid c) = \prod p_{w,c}^{\delta_{w,c}} (1 - p_{w,c})^{1-\delta_{w,c}} \quad (4.5)$$

よって、多変数ベルヌーイモデルは、

$$P(c)p(d|c) = p_c \prod_{w \in V} (p_{w,c}^{\delta_{w,c}} (1 - p_{w,c})^{1-\delta_{w,c}}) \quad (4,6)$$

を最大化するようなcを出力する。

[2]多変数ベルヌーイモデルのパラメータの最尤推定

求めるパラメータは$p_{w,c}$とp_cである。最尤推定なので、

$$\log p(D) = \sum_{(d,c)\in D} \log P(d,c)$$

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

を最大化することになる。

この最大化問題は次の制約付き最適化問題で書くことができる。

$$\max . \quad \log \ p(D)$$

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

ラグランジュの未定乗数法により解く。

未定乗数λを導入し次のように定義する。

$$L(\theta,\lambda) = \log P(D) + \lambda(\sum p_c - 1) \qquad (4.7)$$

各パラメータに関する偏微分を計算する。

$$\frac{\partial L(\theta,\lambda)}{\partial p_{w,c}} = \frac{N_{w,c}}{p_{w,c}} - \frac{N_c - N_{w,c}}{1 - p_{w,c}} \quad \frac{\partial L(\theta,\lambda)}{\partial p_c} = \frac{N_c}{p_c} + \lambda$$

となり、これらをそれぞれ0とし、Σ p_c＝1とあわせると

$$p_{w,c} = \frac{N_{w,c}}{N_c} \qquad p_c = \frac{N_c}{\sum_c N_c} \qquad (4.8)$$

結果として、

$$p_{w,c} = \frac{\left(\text{クラス}c\text{に属する訓練文書のうち}w\text{を含む文書の数}\right)}{\left(\text{クラス}c\text{に属する訓練文書数}\right)}$$

$$p_c = \frac{\left(\text{クラス}c\text{に属する訓練文書数}\right)}{\left(\text{訓練文書数}\right)}$$

を推定していることになる

[3]多変数ベルヌーイモデルのパラメータのMAP推定

値が0.00にならない様なパラメータ推定をマップ推定を用いて考える。

0.00に近いような値をとる確率が小さいディリクレ分布を事前分布に使う。

MAP推定の目的関数は

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

$$= (\alpha - 1)\sum_{c} \log p_c + (\alpha - 1)\sum_{w,c} (\log p_{w,c} + \log(1 - p_{w,c}))$$

$$+ \sum_{(d,c)\in D} \log(p_c \prod_{w\in V} (p_{w,c}^{\delta_{w,c}} (1 - p_{w,c})^{1-\delta_{w,d}})) + (\text{定数})$$

であり、これを Σ p(θ)=1の制約の元で最大化する。

ラグランジュ関数は、

$$L(\theta,\lambda)=\log P(\theta)+\log P(D)+\lambda(\sum_{c} p_c-1)$$

となり、偏微分を計算すると、

$$\frac{\partial L(\theta,\lambda)}{\partial p_{w,c}}=\frac{(\alpha-1)}{p_{w,c}}-\frac{(\alpha-1)}{1-p_{w,c}}+\frac{N_{w,c}}{p_{w,c}}-\frac{N_c-N_{w,c}}{1-p_{w,c}}$$

$$\frac{\partial L(\theta,\lambda)}{\partial p_c}=\frac{(\alpha-1)}{p_c}+\frac{N_c}{p_c}+\lambda$$

となる。

それぞれ0とおき、Σ p_c=1とあわせると

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

α=2とすると、

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

となり全ての生起回数に1を足してパラメータを計算していることになる。

4.2.2 多項モデル

[1]モデルの導入

単語wが選ばれる確率を、$q_{w,c}$で表す。

文書dの生起確率を数式にすると

$$p(d \mid c) = p(k = \sum_{w} n_{w.d}) \frac{(\sum_{w} n_{w,d})!}{\prod_{w \in V} n_{w,d}!} \prod_{w \in V} q_{w,c}^{n_{w,d}}$$

多項モデルのナイーブベイズ分類器は

$$P(c)P(d\mid c)=p_c P(\sum_w n_{w,d})\frac{(\sum_w n_{w,d})!}{\prod_{w\in V} n_{w,d}!}\prod_{w\in V} q_{w,c}^{n_{w,d}} \qquad (4.13)$$

を最大化するようなcを出力する。よって

$$\arg\max_c P(c)P(d\mid c)=\arg\max_c p_c\prod_{w\in V} q_{w,c}^{n_{w,d}} \qquad (4.14)$$

多項モデルはベルヌーイモデルと違い、単語wが文書d内で生起した回数が分類に影響を与える。

|d|= Σ $n_{w,d}$と表し、最尤推定であるので、

$$\log P(D) = \sum_{(d,c)\in D} \log P(d,c)$$

$$= \sum_{(d,c)\in D} \log \frac{P(|d|)|d|!}{\prod_{w\in c} n_{w,d}!} + \sum_{c} N_c \log p_c + \sum_{c}\sum_{w\in c} n_{w,c} \log q_{w,c}$$

を最大化することになる。

多項モデルおいて次のような制約付き最適化問題で書く。

$$\max \log P(d) \quad s.t \ \sum_{c\in C} p_c = 1$$

$$\sum_{w\in V} q_{w,c} = 1; \forall c \in C \qquad (4.14)$$

ラグランジュ未定乗数法を用い次のように定義する。

$$L(\theta,\beta,\gamma)=\log P(d)+\sum_{c\in C}\beta_c(\sum_{w\in V}q_{w,c}-1)+\gamma(\sum_{c\in C}p_c-1)$$

$q_{w,c}$に関する偏微分が0になればよい。

偏微分を計算し、これらを0とし Σ qw,c=1とあわせると次の式が得られる。

$$q_{w,c}=\frac{n_{w,c}}{\sum_{w}n_{w,c}} \qquad (4.15)$$

式(4.15)を言葉で表すと

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

である。p_cについては多変数ベルヌーイモデルの時と同じ。

[3]多項モデルのパラメータのMAP推定

ディリクレ分布を事前分布として用いる。

目的関数は

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

$$= (\alpha - 1)(\sum_{c} \log p_c + \sum_{w,c} \log q_{w,c})$$

$$+ \sum_{(d,c) \in D} \log(\frac{p(|d|)\,|d|!}{\prod_{w \in V} n_{w,d}!} p_c \prod_{w \in V} q_{w,c}^{n_{w,d}}) + (\text{定数})$$

これを式(4.14)のもとで最大化する。

ラグランジュ関数は

$$L(\theta,\beta,\gamma)=\log P(\theta)+\log P(D)$$
$$+\sum_{c\in C}\beta_c(\sum_{w\in V}p_{w,c}-1)+\gamma(\sum_{c\in C}p_c-1)$$

となる。

偏微分を計算し、これを0とおき、$\Sigma\, q_{w,c}=1$とあわせると次の式が得られる。

$$q_{w,c}=\frac{n_{w,c}+(\alpha-1)}{\sum_{w}n_{w,c}+|W|(\alpha-1)} \qquad (4.16)$$

|W|は単語数を表す。α=2とすると

$$q_{w,c} = \frac{n_{w,c} + 1}{\sum_{w} n_{w,c} + |W|} \qquad p_c = \frac{N_c + 1}{\sum_{c} N_c + |C|}$$

となり、wとcの対について生起回数に1を足して、パラメータの値を計算していることになる。

4.3 サポートベクトルマシン

サポートベクトルマシン(SVM)は1990年代の終わりころから使われ始めた線形二分類器であり、非常に高い分類性能を持つ。

正クラス、負クラスの2つのクラスがあり、正クラスに属する事例は正例、負クラスに属する事例は負例という。

訓練データDとして

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(|D|)}, y^{(|D|)})\} \quad (4.17)$$

が与えられているとする。xは事例ベクトル、yは事例のクラスラベルである。線形分類であるので方向ベクトルwと切片bをパラメータとして

$$f(x) = w \cdot x - b \quad (4.18)$$

f(x)≥0なら正クラス、f(x)<0なら負クラスである。

4.3.1 マージン最大化

図4.1のように訓練データが分布しているとする

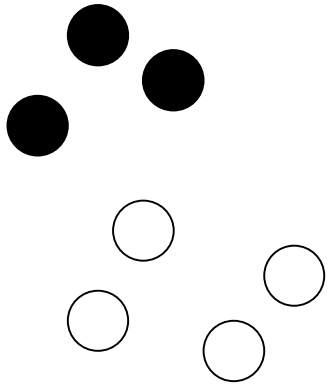

図4.1

これを分類する平面を分離平面とよぶ。

マージン最大化は図4.2のようにどちらのクラスからもなるべく遠い位置で分けるという方法である。

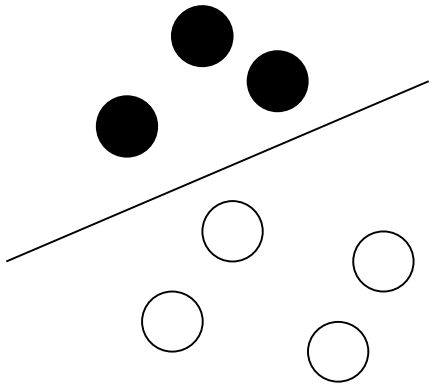

図4.2

分離平面のマージンとは、もっとも近い訓練事例への距離として定義される

分離平面の最も近くにある正例をx_+、分離平面を結ぶ垂線の足をx^*で表す。wとx_+、x^*は同じ方向を向いているので

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

が成り立つ。w・x=bは式全体を定数倍しても変わらず、x^*は分離平面状にあるので、
w・x^*=bである。よって、

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

$$= (b+1) - b = 1$$

この2つの式を合わせると

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

が導ける。よってこの分離平面のマージンは1/|w|となる。絶対値は扱いにくいので2乗して、1/w^2とする。この分離平面のマージンを最大化するので、w^2を最小化することになる。

4.3.2 厳密制約下のSVMモデル

y^(i)=+1のときw･x^(i)－b≧1、y^(i)=-1のとき
w･x^(i)－b≦1である。この二つの条件は

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

とまとめて表せる。

よってこれを制約した次の最適化問題を解く

$$\min \quad \frac{1}{2}w^2 \qquad s.t \;\; y^{(i)}(w\cdot x^{(i)}-b)-1\geq 0;\forall i$$

(4.19) (4.20)

ラグランジュ法用い、ラグランジュ乗数 α_i を導入すると

$$L(w,b,\alpha)=\frac{1}{2}w^2-\sum_i \alpha_i(y^{(i)}(w\cdot x^{(i)}-b)-1)$$

となり、これを偏微分し、それぞれ0と置くことで

$$w^*=\sum_i \alpha_i y^{(i)} x^{(i)} \quad (4.22) \qquad \sum_i \alpha_i y^{(i)}=0 \quad (4.23)$$

が得られる。

式(4.22)を分離平面の式に代入すると、

$$f(x) = \sum_i \alpha_i y^{(i)} x^{(i)} \cdot x - b \tag{4.24}$$

となり、後はα_iとbを求めれば分離平面を得られる。そこで式(4.22)を式(4.21)のラグランジュ関数に代入すると、

$$L(w^*, b, \alpha) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(j)} \cdot x^{(i)} + \sum_i \alpha_i \tag{4.25}$$

となる。

式(4.25)でwとbが消えたので、あとはこの関数を最大化するα_iを求めれば、wが求まり、切片bも解くことができる。この最大化問題は2次計画問題としてよく知られた問題である。

4.3.3 緩和制約下のSVMモデル

4.2.2で導出したSVMは実際のデータだとうまく動かない。極端な場合では、訓練データが線形関数で分離できない場合、制約を満たす解が存在しないことになる。

ここでは式(4.20)の制約を少し緩め、

$$y^{(i)}(w\cdot x^{(i)}-b)-1\geq-\xi_i \quad (4.26)$$

とする。変数ξi（≥0）はi番目の訓練事例がうまく分けられない度合いを示す。

最適化問題はつぎのようになる

$$\min \ \frac{1}{2}w^2+C\sum_i\xi_i \quad (4.27) \qquad s.t \quad y^{(i)}(w\cdot x^{(i)}-b)\geq 1-\xi;\forall i$$

$$\xi\geq 0\forall i; \quad (4.28)$$

ラグランジュ法を用いてと関数は

$$L(w,b,\xi,\alpha,\beta)=\frac{1}{2}w^2+C\sum_i\xi_i$$
$$-\sum_i\alpha_i(y^{(i)}(w\cdot x^{(i)}-b)-1+\xi_i)-\sum_i\beta_i\xi_i$$

表せる。これを偏微分しそれぞれを0とすると、

w、bは厳密制約下と同じになり、ξiに関して

$$C=\alpha_i+\beta_i$$

が得られる。

よって、0≤αi≤Cの条件のもとで

$$L(w^*, b, \alpha, \beta) = -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)} \cdot x^{(j)} + \sum_i \alpha_i$$

を最大化するαを求める。

4.3.4 関数距離

- 事例xを、f(x)>=0なら正クラス、f(x)<0なら負クラスとする。このf(x)を関数距離という。
- 関数距離が大きい事例だけを集めると高い確率で正例となる事例が集められ、0に近い事例だけを集めると高い確率で分類結果が誤りとなる事例を集められる。

4.3.5 多値分類器への拡張

[1] one-versus-rest法

各クラスについての分離平面を作り、そのクラスに属するか否かを判別する。

クラス数がnならば、SVMによりn個の分離平面を求める。

4.3.5の続き

[2]ペアワイズ法

クラス対ごとに、その二つのクラスの内どちらであるか分ける平面を作る。

クラス数がnの場合、n(n-1)/2個の分離平面を作る。

one-versus-rest法と比べると多くの平面が作られるが、全体としては訓練時間が減ることが多い。