機械学習の続きと
前回の課題について

グループI4
2008MI233 鈴木健太
2008MI214 沢田天馬

目次

- 分類知識の学習の実装
- ベイズの定理
- ベイジアンフィルタ
- 前回の課題の続き
- 今後の課題
- 参考文献

分類知識の学習の実装(1/2)

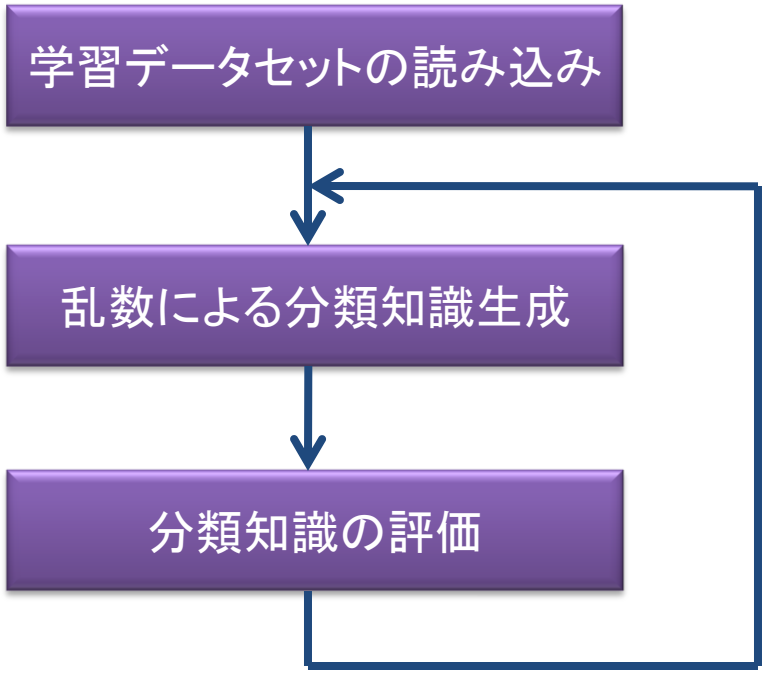
表1　formula[a] [b] 配列の要素と命題の対応関係

a \ b	0	1	2	3	4	5	6	7
0	p1	¬p1	p2	¬p2	p3	¬p3	p4	¬p4
1	p1	¬p1	p2	¬p2	p3	¬p3	p4	¬p4
2	p1	¬p1	p2	¬p2	p3	¬p3	p4	¬p4

表2　(p1 ∨　p2) ∧　p3　∧　¬p4　の、2次元配列による表現

a \ b	0	1	2	3	4	5	6	7
0	1	0	1	0	0	0	0	0
1	0	0	0	0	1	0	0	0
2	0	0	0	0	0	0	0	1

分類知識の学習の実装(2/2)



ex.txt

```
1 1 1 0 1   +1
1 0 1 0 1
0 1 1 0 1
1 1 1 0 1
0 0 1 0 0   +0
1 1 0 0 0
1 1 1 1 0   +1
1 1 0 0 0
1 1 0 1 0
0 0 1 1 0
```

p1 p2 p3 p4　カテゴリ

(¬p3 ∨ ¬p4) ∧ (¬p1 ∨ ¬p2 ∨ p3)

```
ファイル(F)  編集(E)  表示(V)  端末(T)  タ
[08mi214@08mi214 08mi214]$ cd kikai
[08mi214@08mi214 kikai]$ ./eml < ex.txt

スコア:0.800000
00100100
01001000
10111111

スコア:0.900000
00000101
01101111
01011000

スコア:0.900000
00000001
10101110
01011010
```

p3∨¬p3 となり、分類に
何も貢献しないので無視する

ベイズの定理

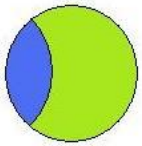
ベイズの定理とは、条件付き確率を変形したものにすぎない

条件付き確率：ある事象Bが起こる条件下で、別の事象Aが起こる確率

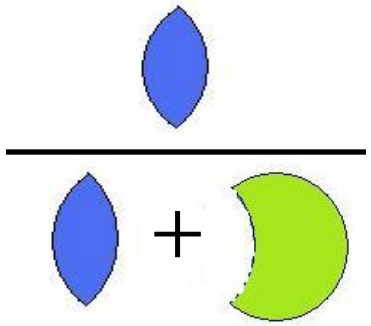
記号ではP(A | B)と表す

例　あなたが人間だとして、アメリカ人である確率は？　→　P(アメリカ人 | 人間)

$$P(A \mid B) = P(A \cap B) \,/\, P(B)$$



$$= P(A \cap B) \,/\, P(A \cap B) + P(\bar{A} \cap B)$$



ベイジアンフィルタ

ベイズの定理を実装した教師あり学習

ベイジアンフィルタで重要になる点

予測をするために、過去の情報を利用する

予測された答えも過去の情報となり利用される

学習量が増えると
フィルタの分類精度の上昇につながる

現状では

・アキネーター

・Mozilla Thunderbirdのスパム分類フィルタ

ベイジアンフィルタの実装(1/2)

利用した言語

Python(コンパイルを必要としないスクリプト言語に属する)

利用したもの

- Cygwin　(Windows上で開発環境を整えるためのツール)
- Yahoo!デベロッパーのテキスト解析API(日本語形態素解析)
- BeautifulSoup (XMLをパージングするPythonライブラリ)

今回は、簡単な理解のためにサンプルプログラムの実行のみを行った

ベイジアンフィルタの実装(2/2)

プログラムの流れ

用意された日本語の文章に対して日本語形態素解析APIを利用

今回は、Wikiの
Python,Ruby,機械学習
の項目の一文を利用

日本語は英文のようにスペースで区切る
「わかち書き」がされていないため、形態素解析
で文脈の解析や単語の分解をする必要がある

BeautifulSoupを利用して、XMLの解析を行う（タグの抽出やソート）

カテゴリは
「機械学習」,「Ruby」,「Python」

ベイズの定理のアルゴリズムを用いて、未知の文章に対してのカテゴリを推定する
カテゴリの推定には尤度（もっともらしさ）が用いられる

与えられた文章がどのカテゴリに属するのが
もっともらしいかを単語の出現確率から求めている

実行結果

与えてあるWikiの一文

Python（パイソン）は、オランダ人のグイド・ヴァンロッサムが作った・・・

Ruby（ルビー）は、まつもとゆきひろ（通称Matz）により開発されたオブジェクト指向スクリプト言語であり・・・・

豊富な機械学習（きかいがくしゅう、Machine learning）とは、人工知能における研究課題の一つで・・・・

これらの文をもとに未知の文をカテゴリに分類

未知の文章に対するカテゴリの分類

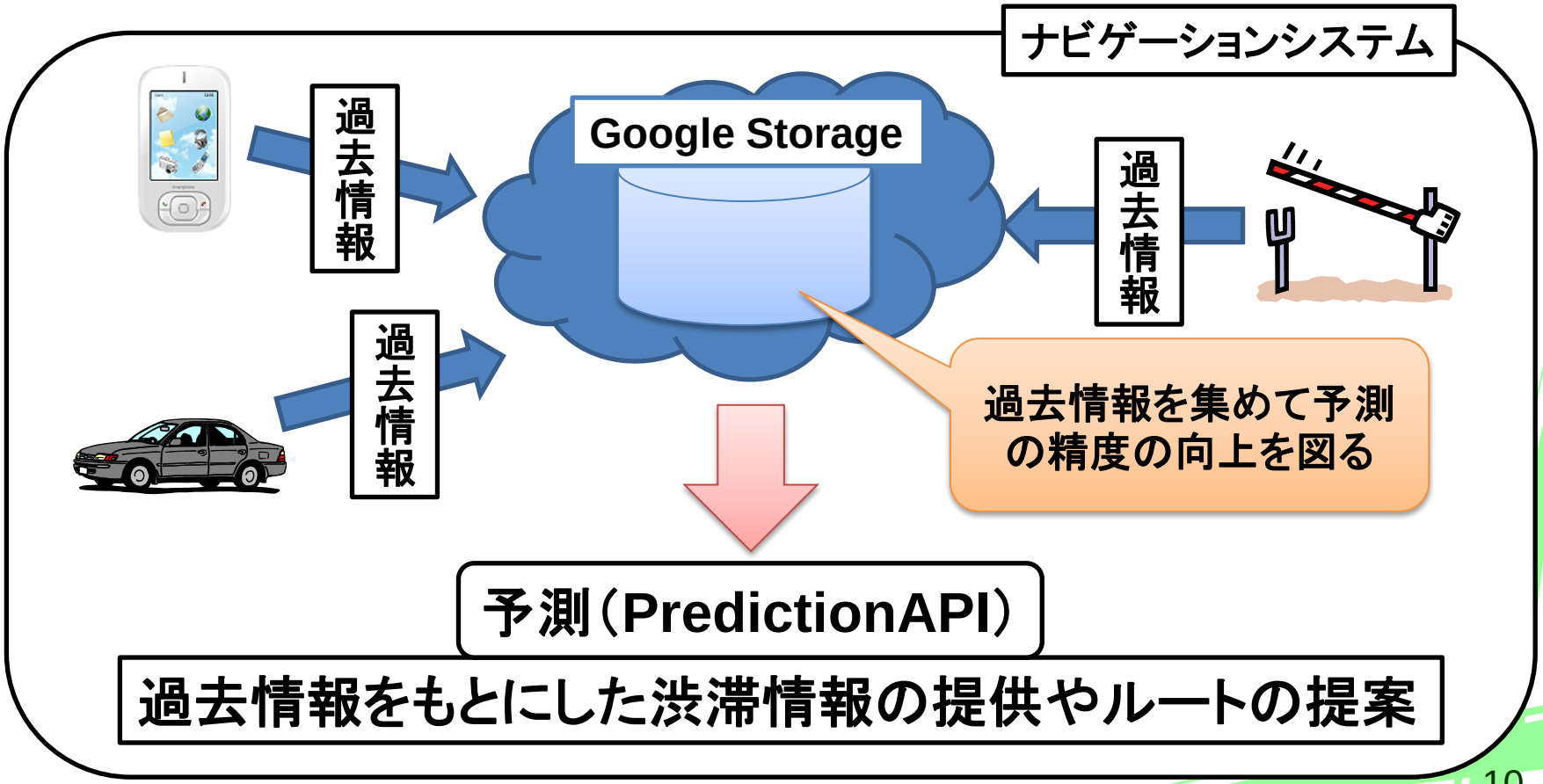
```
$ python naivebayes.py
ヴァンロッサム氏によって開発されました。 => 推定カテゴリ: Python
豊富なドキュメントや豊富なライブラリがあります。 => 推定カテゴリ: Python
純粋なオブジェクト指向言語です。 => 推定カテゴリ: Ruby
Rubyはまつもとゆきひろ氏(通称Matz)により開発されました。 => 推定カテゴリ: Ruby
「機械学習 はじめよう」が始まりました。 => 推定カテゴリ: 機械学習
検索エンジンや画像認識に利用されています。 => 推定カテゴリ: 機械学習
```

前回の課題の続き

与えられた情報を機械学習に利用するための仕組みはどうすればいいのか

これに対する一案

過去の情報をもとに、未来を予測するというベイズの定理の考えを利用



今後の課題

- コンテキストアウェアなサービス提供技術によって、何が解決できるのか、問題は何なのかをはっきりさせていくこと

現時点では、
「サービスはユーザー側から要求があって提供できるもの」
と考えている

機械学習を用いて、そのような状況を変えていくことができないか

機械学習を用いるとずっと言ってきたが、改めて、機械学習を利用するメリット・デメリットの理解の必要性

参考文献

- はじめての機械学習　小高知宏著
- 確率統計　馬場敬之　久池井茂著
- 技術評論社　http://gihyo.jp/
- e-Words　http://e-words.jp/
- Yahoo!デベロッパーネットワーク
  http://developer.yahoo.co.jp/