集合知プログラミング
第10章　特徴を発見する

06T4007A　伊藤輝将

はじめに

- これまでの章
  →基本的に教師ありの分類器について

- 本章
  →データセットに内在している重要な特徴たち
  を抽出する方法について

特徴の抽出(1)

- クラスタリング

  -データセット中のすべての行は階層中のグループもしくは点に割り当てられていた

  ↓このアイデアをさらに一般化

  つまり、特徴の抽出とは、、、

  組み合わせることでもとのデータセットの行を再構築できるような新しい行たちを探し出そうと試みるもの

特徴の抽出(2)

・データの重要な特徴を抽出するための技術

→非負値行列因子分解

（NFM：non-negative matrix factorization)

-これは本書でカバーする技術の中では最も洗練されたものの中のひとつ

-詳しくは次節で

行列に変換する

・特徴抽出のアルゴリズムは非常に大きな数字の行列を利用する

行：アイテム　　列：プロパティ

例）

articles = ['A','B','C',・・・

words = ['hurricane','democrats','world',・・・

matrix = [[3,0,1,・・・]
[1,2,0,・・・]
[0,0,2,・・・]
・・・]

実行結果

```
C:¥Documents and Settings¥kato¥デスクトップ¥chapter10>python
Python 2.5.1 (r251:54863, Apr 18 2007, 08:51:08) [MSC v.1310 32 bit
    (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import newsfeatures1
>>> allw,artw,artt=newsfeatures1.getarticlewords()
>>> wordmatrix,wordvec=newsfeatures1.makematrix(allw,artw)
>>> wordvec[0:10]
['protest', 'sanford', 'chinese', 'believed', 'leaders', 'street', 'china',
    'military', 'sarah', 'would']
>>> artt[1]
u'U.S. and Russia Take Step to Cut Nuclear Arms'
>>> wordmatrix[1][0:10]
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

これまでのアプローチ

- ベイジアン分類器
  -教師あり学習の手法の一つ
  -すべてのテーマについてテーマごとにいくつかの例でトレーニングする必要があるため、カテゴリが少なく、それぞれのカテゴリ中の例が多いものに向いている

- クラスタリング
  -教師なし学習の手法の一つ
  -人物についてのニュース記事は分類できないことが時々ある

実行結果

```
>>> def wordmatrixfeatures(x):
...     return [wordvec[w] for w in range(len(x)) if x[w]>0]
...
>>> wordmatrixfeatures(wordmatrix[0])
['china', 'after', 'down', 'deadly']
>>> import docclass
>>> calssfier=docclass.naivebayes(wordmatrixfeatures)
>>> classfier.setdb('newstest.db')
>>> artt[0]
u'China Locks Down Restive Region After Deadly Clashes'
>>> #イラクに関する記事としてトレーニング
>>> classfier.train(wordmatrix[0],'iraq')
>>> artt[1]
u'U.S. and Russia Take Step to Cut Nuclear Arms'
>>> #インドに関する記事としてトレーニング
>>> classfier.train(wordmatrix[1],'india')
>>> artt[2]
u'Health Co-op Offers Model for Overhaul'
>>> #この記事はどのように分類されるだろうか？
>>> classfier.classify(wordmatrix[1])
u'india'
```