

データマイニング

- テキストマイニング
- 自然言語処理
- 機械学習
- 情報検索

ナイーブ・ベイズ

ナイーブ・ベイズ学習とは...

■ データをクラスに分類するための一手法

- 大量にあるデータをクラスに分類する
- 文書分類：大量にある新聞記事を、記事の内容に対応したクラスに分類する（例：「経済」「スポーツ」「娯楽」等）
- メールフィルタリング：受信したEメールを「迷惑メール」と「普通メール」に分類する
- データにクラスに対応するラベルを付与することと同義
- 今回は、1つのデータに対して、1つのクラスに対応するラベルを付与

クラス分類の例

■ Yahooのトップページ

The screenshot shows the Yahoo! Japan homepage with the 'Economy' (経済) category selected. The page features a search bar, navigation links, and a list of news items. A blue arrow points from the 'Economy' category tab to a text box explaining the classification process.

YAHOO! JAPAN
ゴールデンウィーク特集

ウェブ | 画像 | 動画 | 辞書 | 知恵袋 | 地図 | リアルタイム | 一覧

検索

の節句、簡単レシピほかお楽しみ情報 フレッツ光で最大50,000円バック 明日公開映画「宇宙兄弟」を“スマホ

トピックス **経済** エンタメ スポーツ その他

7時36分更新

ECB、政策金利を据え置き
三菱UFJ銀 大企業100社支援へ
中韓FTA交渉へ 日本に逆風
日中韓 相互の国債購入で合意
米自動車ビッグ3が黒字確保
日産ルノー 露社と合併設立
関電管内、計画停電立案へ
親を入社式へ招く企業 理由は
昨日の話題(5件) 一覧

日経平均株価
9,380.25
+29.30

NYダウ
13,206.59
-61.98

米国ドル
80.19

YAHOO! JAPAN mobage

記事が経済、エンタメ、スポーツ、その他に分類されている



各記事に「経済」、「エンタメ」、「スポーツ」、「その他」のラベルを付与

自動化したい

ナイーブ・ベイズ学習の例

■ 学習用データ

天気	温度	湿度	風	ゴルフプレイ
晴	暑	高	無	×
晴	暑	高	有	×
曇	暑	高	無	○
雨	暖	高	無	○
雨	涼	普通	無	○
雨	涼	普通	有	×
曇	涼	普通	有	○
晴	暖	高	無	×
晴	涼	普通	無	○
雨	暖	普通	無	○
晴	暖	普通	有	○
曇	暖	高	有	○
曇	暑	普通	無	○
雨	暖	高	有	×

■ テスト用データ(新規事例)

天気	温度	湿度	風	ゴルフプレイ
晴	涼	高	有	?



新規事例データ → 晴 涼 高 有
○ 尤度:0.00529100529100527
× 尤度:0.0205714285714286
判定:×

素性 「天気」「温度」「湿度」「風」

素性値 「晴」「暑」「高」「無」

ナイーブ・ベイズ

- クラス（ゴルフプレイ＝○，ゴルフプレイ＝×）に分類するにおいて、素性（天気、温度、湿度、風）が互いに独立であると仮定して学習を行う手法



素性が互いに独立と仮定（ナイーブ）：
現実世界ではあまり成立しないが、それなりにうまくいく

素性	「天気」「温度」「湿度」「風」
----	-----------------

素性値	「晴」「暑」「高」「無」
-----	--------------

ナイーブ・ベイズ学習式

■ ナイーブ・ベイズ学習式

$$\hat{c} = \arg \max_{c=\{\bigcirc, \times\}} P(c | E)$$

c: クラス(ゴルフプレイ=○、×)

E: 新規事例データの素性値列(例: 天気=晴, 温度=暑, 湿度=高, 風=無)

素性

「天気」「温度」「湿度」「風」

素性値


「晴」「暑」「高」「無」

ナイーブ・ベイズ学習式の変形

- 学習データから計算できるようにナイーブ・ベイズ学習式を変形

※ベイズの定理に変形(証明は黒板に板書)

$$\begin{aligned}\hat{c} &= \arg \max_{c=\{\circ, \times\}} P(c \mid E) \\ &= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E \mid c)}{P(E)} = P(c)P(E \mid c)\end{aligned}$$

 $P(E)$ は c に依らないので考えなくてよい

ナイーブ・ベイズ学習式の変形

クラスcの出現確率→
学習用データから推定可能

$$\hat{c} = \arg \max_{c=\{\circ, \times\}} P(c | E)$$

$$= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E | c)}{P(E)} = P(c) \underbrace{P(E | c)}$$

- $P(\circ) = 9/14$
- $P(\times) = 5/14$

クラスcが与えられたときの素性値列Eの出現確率は、
様々な事例データが存在するため、推定することは困難

c: クラス(ゴルフプレイ=○、×)

E: 事例データの素性値列(例: 天気=晴, 温度=暑, 湿度=高, 風=無)

ナイーブ・ベイズ学習式の変形

■ $P(E | c)$ を計算

$$P(E | c) = P(\{e_1, e_2, e_3, e_4\} | c)$$

ここで、 E に含まれる各素性値 e_1, e_2, e_3, e_4 が互いに独立に生起すると仮定すると...

$$\begin{aligned} P(E | c) &= P(\{e_1, e_2, e_3, e_4\} | c) \\ &\approx P(e_1 | c)P(e_2 | c)P(e_3 | c)P(e_4 | c) = \prod_{i=1}^4 P(e_i | c) \end{aligned}$$

ナイーブ・ベイズ学習式のパラメータ推定

■ $P(e_1 | c), P(e_2 | c), P(e_3 | c), P(e_4 | c)$ を学習用データから推定

$$P(E | c) \approx P(e_1 | c)P(e_2 | c)P(e_3 | c)P(e_4 | c) = \prod_{i=1}^4 P(e_i | c)$$

天気	温度	湿度	風	ゴルフプレイ
曇	暑	高	無	○
雨	暖	高	無	○
雨	涼	普通	無	○
曇	涼	普通	有	○
晴	涼	普通	無	○
雨	暖	普通	無	○
晴	暖	普通	有	○
曇	暖	高	有	○
曇	暑	普通	無	○

$P(E | \text{ゴルフプレイ} = \text{○})$

$= P(\text{天気} = \text{晴れ} | \text{○}) \times P(\text{温度} = \text{涼} | \text{○})$

$\times P(\text{湿度} = \text{高} | \text{○}) \times P(\text{風} = \text{有} | \text{○})$

$= (2/9) \times (3/9) \times (3/9) \times (3/9)$

ナイーブ・ベイズ学習式(最終)

- 学習データからパラメータを推定できるようにナイーブ・ベイズ学習式を変形

$$\begin{aligned}\hat{c} &= \arg \max_{c=\{\circ, \times\}} P(c \mid E) \\ &= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E \mid c)}{P(E)} \approx \frac{P(c)}{\prod_{i=1}^4 P(e_i \mid c)}\end{aligned}$$

学習用データから推定可能

課題

- サンプルデータ(各自、ダウンロードせよ)の学習用データを訓練データとして、テスト用データをナイーブベイズ手法で分類するプログラムを実装せよ。素性として名詞を用いよ。

・訓練データの1記事(全体で227記事)

```
<id> 2012-01-24_1 </id>
<date> 2012/01/24 </date>
<company> エレコム </company>
<class> IT 周辺機器 </class>
<title> エレコム、名刺用紙「なっとく。名刺」シリーズからスーパーファイン用紙など2タイプを発売 </title>
  しっかりとした厚みで上品に仕上がる「特厚」タイプ！
  切り口がすっきりきれいなクリアカットタイプの名刺用紙「なっとく。名刺」の新シリーズを発売
  エレコム株式会社(本社:大阪府中央区、取締役社長:葉田順治)は、手持ちのプリンタなどでオリジナル
  の名刺やメッセージカードが簡単に作成できる名刺用紙「なっとく。名刺」シリーズについて、新たに2タイプ4
  種類を2月上旬より新発売いたします。 ご家庭のプリンタを使って、簡単・手軽にオリジナルの名刺が作成
  できる「なっとく。名刺」は、人気の名刺用紙シリーズです。
```


課題

- サンプルデータ(各自、ダウンロードせよ)の学習用データを訓練データとして、テスト用データをナイーブベイズ手法で分類するプログラムを実装せよ。素性として名詞を用いよ。

・テスト用データの1記事(全体で12記事)

推定せよ

```
<id> 2012-04-25_0 </id>  
<date> 2012/04/25 </date>  
<company> グーグル </company>
```

```
<class> </class>
```

```
<title> グーグル、全てのファイルを安心して保存・共有できるGoogleドライブを発表 </title>  
Googleドライブを使って、何でも保存、共有しよう。
```

```
Posted by デービッド ウォルツ／Googleドライブ・プロダクトマネージャー
```

本日、全てのファイルを安心して保存、共有できるGoogleドライブを発表しました。ドライブでは写真、動画、Googleドキュメント、PDFなどさまざまなファイルを一つの場所に保管できるようになります。例えば、顧客とのビジネスプランの相談や、来年度の予算計画を同僚と立てる場合も、共同作業が効率よくでき、プロジェクトを迅速に進められるでしょう。Googleドキュメントが進化したGoogleドライブは、クラウドでの働き方を大きく変えます。

ナイーブ・ベイズ学習式

- 学習データからパラメータを推定できるようにナイーブ・ベイズ学習式を変形

$$\hat{c} = \arg \max_{c=\{\circ, \times\}} P(c | E)$$

$$= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E | c)}{P(E)} \approx \underbrace{P(c)}_{\text{学習用データから推定可能}} \prod_{i=1}^n \underbrace{P(e_i | c)}_{\text{学習用データから推定可能}}$$

$$P(c) = \frac{fa(c, S)}{|S|}$$

$$P(e_i | c) = \frac{f(e_i, c)}{f(c)}$$

- $fa(c, S)$: テストデータ記事集合 S 中において、クラス c が出現する記事の頻度
- $f(e, c)$: クラス c の記事集合において、素性の名詞 e の出現数
- $f(c)$: クラス c の記事集合において、素性の名詞の総出現数

ナイーブ・ベイズ学習式の実装

- 確率の積和は、ただでさえ小さい値の積なので、実装においては問題が生じる→0に極めて近い値となり、最終的に0になってしまう

$$\hat{c} = \arg \max_{c=\{\circ, \times\}} P(c | E)$$

$$= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E | c)}{P(E)} \approx \frac{P(c)}{\prod_{i=1}^n P(e_i | c)}$$

学習用データから推定可能

- 実装上の解決方法として、確率値のlogをとって、積和を和に変換する → 小さい値も負の整数になって好都合

ナイーブ・ベイズ学習式(最終)

- 実装上の解決方法として、**確率値のlogをとって**、積和を和に変換する → 小さい値も負の整数になって好都合

$$\hat{c} = \arg \max_{c=\{\circ, \times\}} P(c | E)$$

$$= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E | c)}{P(E)} \approx \underbrace{P(c)}_{\text{学習用データから推定可能}} \prod_{i=1}^n \underbrace{P(e_i | c)}_{\text{学習用データから推定可能}}$$



$$\hat{c} = \arg \max_{c=\{\circ, \times\}} P(c | E)$$

$$= \arg \max_{c=\{\circ, \times\}} \frac{P(c)P(E | c)}{P(E)} \approx \log(P(c)) + \sum_{i=1}^n \log(P(e_i | c))$$