知的情報処理
4. ナイーブなベイズ法

櫻井彰人
慶應義塾大学理工学部

1

今日の目標

- ベイズ推定ということを少し考える
- ナイーブ・ベイズ法の原理と実装を知る
  - 実務でのベイズ推定方法の一つとして

2

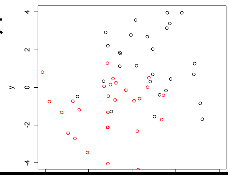
目次

- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

3

今日扱う問題

- 被説明変数値は2値。但し、確率的に決まる場合
  - （正確には）
    決定的に決まっていてもよい
    ただ、未観測の変数に依存している
- この事情は



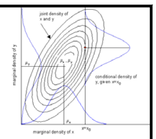
4

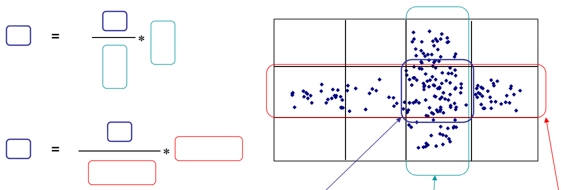
目次

- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

5

復習: 条件付確率

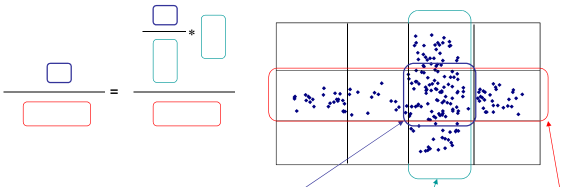




$$p(m \mid x)p(x) = p(x, m) = p(x \mid m)p(m)$$

6

Bayesの定理



$$p(m\,|\,x) = \frac{p(x,m)}{p(x)} = \frac{p(x\,|\,m)\,p(m)}{p(x)}$$

7

目次

8

では、ベイズ推論とは

- ある証拠に基づいて、その原因となった事象を推定するための確率論的方法である。
- Bayesian inference is a method of statistical inference in which some kind of evidence or observations are used to calculate the probability that a hypothesis may be true, or else to update its previously-calculated probability.

$$p(m\,|\,x) = \frac{p(x\,|\,m)\,p(m)}{p(x)}$$

Wikipedia より

9

脱線: 一般記事にも

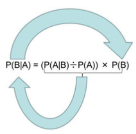
- **「今、話題の自動運転車は、なぜ自動で運転できるのか? その基本メカニズムを「ベイズ理論」にまで遡って解説」**
  - 小林 雅一. **現代ビジネス、ITトレンド・セレクト.**

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

P(B|A) = (P(A|B)÷P(A)) × P(B)

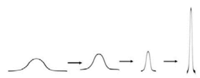
事後確率 = (実験・測定・観測などの結果) × 事前確率

これをさらに噛み砕いて説明すると、次のようになる。つまり、「まず最初は『いい加減』というか、かなり適当に決めた不正確な確率(事前確率)から出発し、これを何らかの実験や測定、観測などによって、もっと正確な確率(事後確率)へと改良していこう」という考え方だ。これがベイズ定理の真意なのである。



【図1】 ベイズ定理は繰り返し、循環的に適用される

http://gendai.ismedia.jp/articles/-/37143



【図3】 カルマン・フィルターの原理: センサーによる位置測定とベイズ定理の適用を繰り返すことで、誤差を徐々に収束させて、移動体の位置を正確に把握する

カルマンフィルター (Kalman filter) は、誤差のある観測値を用いて、ある動的システムの状態を推定あるいは制御するための、無限インパルス応答フィルターの一種である

10

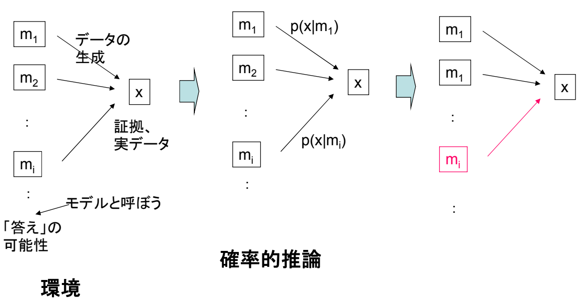
ベイズ推論補足

- 証拠を x で、原因を m で表す
  - 原因の候補を m_i で表す。
- ベイズ推論は、ある方法で、$p(m_i|x)$ を計算し、(m_iの中から)原因 m を推定すること

$$p(m\,|\,x) = \frac{p(x\,|\,m)\,p(m)}{p(x)}$$

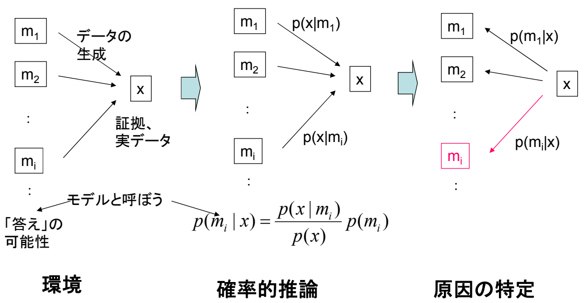
11

(確率的)推論の枠組み



12

（ベイズ）推論の枠組み



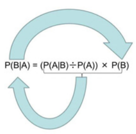
$p(m)$ と $p(x|m)$ の推定

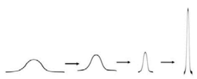
$$\underbrace{p(m \mid x)}_{\text{事後確率}} = \frac{p(x,m)}{p(x)} = \frac{\overbrace{p(x \mid m)}^{\text{条件付き確率}}}{p(x)}\overbrace{p(m)}^{\text{事前確率}}$$

- p(m) はクラスmの発生度数を用いて推定すればよい
- では、p(x|m) はどうしたら推定できるだろうか？
  - p(x|m) はモデルmからデータ x が生成される確率を表す。個別のxに対する p(x|m) を知るには、任意のxに対するp(x|m)をしっていればよい（あたりまえ）。通常はモデルの記述そのものである。
- いろいろありうるよなぁ。正規分布か、多項分布か、、、
- そうしたモデル記述の一つの方法が naïve Bayes である

ベイズ推論とナイーブベイズ

- ベイズ推論





- ナイーブベイズ
  – ベイズ推論を簡便に行う方法

目次

Naïve Bayes なモデル記述

- まず、証拠 x は、複数の属性で記述されているとする
  – これは普通（例外はあるが）。
  – 属性とは、人間であれば、性別、年齢、住所、体重、趣味、、、、; 販売であれば、商品名、個数、単価、販売日時、顧客の性別、、、、
- 各属性は統計的に独立とする
  身長と体重は独立？
  年齢と趣味とは独立？
  ありえない！
  –「そんなバカなことするなよ。ありえない！」と言うのが真っ当な人の台詞。しかし、それをあえて仮定するのが、naïve たるゆえん。

複数の属性で記述されている

というのは

- 「証拠」x の属性が $<a_1,\ldots,a_n>$であるとすれば、x と書いても $<a_1,\ldots,a_n>$ と書いても同じということ
  – 例
  –「太郎」君は、身長、体重、学科、性別が <172, 63, 管理、男性> である
  –「伝票123」は、<2013年10月15日18時30分、日吉駅前店、男性、20代、ジュース、おにぎり> である

属性が統計的に独立というのは

- 「証拠」x の属性が $<a_1,\ldots,a_n>$であるとき、

$$p(X=x)=p(A_1=a_1,\ldots,A_n=a_n)$$
$$=\prod_{i=1}^{n} p(A_i=a_i)$$

- あとで「条件付独立」というのが出てきます。

$$p(X=x\,|\,C=c)=p(A_1=a_1,\ldots,A_n=a_n\,|\,C=c)$$
$$=\prod_{i=1}^{n} p(A_i=a_i\,|\,C=c)$$

Naïve Bayes なモデル記述

とは、

- 「証拠」x は、その属性で $<a_1,\ldots,a_n>$ と書け
- 次が成り立つこと

$$p(X=x)=p(A_1=a_1,\ldots,A_n=a_n)$$
$$=\prod_{i=1}^{n} p(A_i=a_i)$$
$$p(X=x\,|\,C=c)=p(A_1=a_1,\ldots,A_n=a_n\,|\,C=c)$$
$$=\prod_{i=1}^{n} p(A_i=a_i\,|\,C=c)$$

話を戻して

- 欲しいのは、$p(m|x)$ であった。

$$p(m\,|\,x)=\frac{p(x,m)}{p(x)}=\frac{p(x\,|\,m)}{p(x)}p(m)=\frac{p(a_1,\ldots,a_n\,|\,m)}{p(x)}p(m)$$

であるから

$$p(m\,|\,x)=\frac{\prod_{i=1}^{n} p(a_i\,|\,m)}{p(x)}p(m)$$

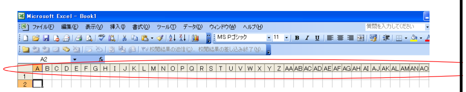
とするのが、naïve Bayes

で、それは何がいいの？

- それは、属性の数が問題の種だから。

「どんな問題」の種？

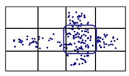
- 「属性数が多いと、(確率分布の）パラメータ推定に必要なデータ数が非常に大きくなる」という問題



目次

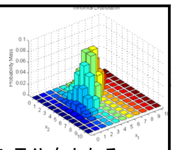
- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

属性数について（続）



- 属性を表す確率変数は離散値をとるものとしよう。以下、例で考えよう。
- $<A_1,A_2,A_3,A_4>$ は<身長、体重、胸囲、座高>であり、どの変数も、高、中、低の3値（0,1,2と略記する）をとるものとする。
- 特に分布は仮定しない（前提知識なし）。となると、実は、3^4=81個の, $<A_1,A_2,A_3,A_4>$ の可能な値一組につき一個の確率 $p_{<A_1,A_2,A_3,A_4>}$ が決まれば分布が決まったことになる。総和が1という制約があるので、80個の値が決まればよい。
- この値をデータから決める（推定する）には、データは何個ぐらい必要なのであろうか、考えてみよう（かなりいい加減に）。

多項分布



- 各データ(証拠)は互いに独立であるとする。
- <A_1,A_2,A_3,A_4> の発生回数の確率分布は、多項分布となる。
- 多項分布: 事象 e_i が発生する確率を p_i とする(p_iの総和は1)。総計n回繰り返した時に事象 e_i がn_i回発生する確率は

$$p(n_1,\ldots,n_k;n,p_1,\ldots,p_k)=\frac{n!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k}$$

- なお、期待値、分散、共分散は

$$E(N_i)=np_i,\ \mathrm{var}(N_i)=np_i(1-p_i),\ \mathrm{cov}(N_i,N_j)=-np_ip_j$$

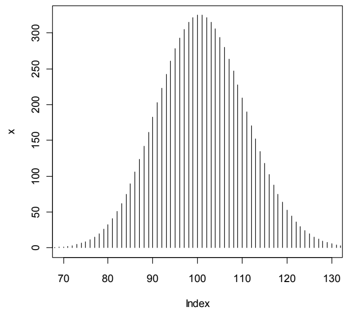
25

属性数について(続々)

- $p_{<A_1,A_2,A_3,A_4>}$ は81個あるので、仮に $p_{<0,0,0,0>}$ =1/81 とし、これだけを推定するものとしよう。
- <0,0,0,0>の分布は、2項分布であり、例えば、n=8100 とすれば、平均 $np_{<0,0,0,0>}$ =100, 分散$np_{<0,0,0,0>}$ $(1-p_{<0,0,0,0>})$≈98.8, 標準偏差≈9.94となる。
- ということは、$p_{<0,0,0,0>}$ を推定するのに、n=8100 としても、<0,0,0,0> の個数が、100±10以内 (誤差10%以内)となる確率は概算約68%(ほぼ1σだから)。
  - 悪い :-(
- ところが、各属性が独立だとすると、 $p_{<0,0,0,0>}=\prod p_{A_i=0}$ 故、各 $p_{Ai=0}$ を推定すればよく、それぞれに全データ(今の例では n=8100 個)が使える。
- ということは: $p_{Ai=0}$ =1/3とすると、n=8100 に対し、平均2700, 分散 1800, 標準偏差≈42.4 となる。2700±270以内 (誤差10%以内)となる確率は概算1 – 2/10億(6σ) 以上
  - n=300とすれば、平均100, 分散≈66.7, 標準偏差≈8.16故、100±10以内 (誤差10%以内)となる確率は68%よりは大きいが、まぁ、同じくらい(1σより大)

26

```
> x<-dbinom(0:200,8100,1/81)*8100
> plot(x,type="h",xlim=c(70,130))
>
```

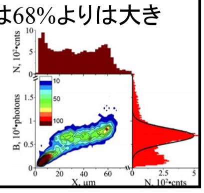


27

属性数について(続々々)

ざっと纏めれば

- $p_{<A_1,A_2,A_3,A_4>}$ を、仮定なしに、推定しようとすると、n=8100 で誤差10%以内となる確率は概算約68% (ほぼ1σ)
- 一方、naïve Bayes的に属性の独立を仮定すると、n=300 で誤差10%以内となる確率は68%よりは大きいが、まぁ、同じくらい
  - n=8100もあれば、誤差10%以内となる確率は 1 – 2/10億(6σ) 以上



28

独立ならよいことばかりか？

- 真に独立なら、よいことばかりである。
- しかし、真に独立なわけはない
  - 風邪か否かの診断を考えよう。咳が酷ければ、喉が炎症を起こし、熱が出る。
  - だから、<咳, 喉の炎症, 熱> という3属性は互いに独立ではない
- 独立でないのに、独立を仮定すると何が起こるか？
- めちゃくちゃになる(何をしているか分からなくなる)はず。
  - 実際、naïve Bayesによって推定した確率値はまったく合っていないといわれている。
- しかし、実際には、naïve Bayes がうまく機能することが多い。これは、
  - 誤った独立性仮定による誤りの増加より、独立性仮定によってパラメータ数を減らしてパラメータの推定精度を向上させたことによる誤りの減少が勝っている
    - 分布を推定しているわけではなく、クラス・分類を推定しているのである。
  - 実際には、独立でなくとも独立として十分近似できることが多い

からではないかと考えられる。

29

従って、naïve Bayes

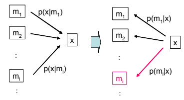
- まぁ使ってみよう(と昔の人は考えた)。
- 実際、結構うまくいく。
  - 確率値の推定はだめです。
  - うまくいくのは、分類に使う場合
- では、「分類」に使う方法を以下に。

30

目次

- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

Naïve Bayes 分類器



- 前のスライドに戻って、

$$p(m \mid x) = \frac{p(x \mid m)}{p(x)} p(m)$$

において、m_1 として クラス1, m_2 としてクラス2 を考える

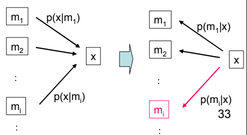
- 証拠 x は観測（データ1個）の集合で、各データは、属性 $<A_1,\ldots,A_n>$ で記述される。
  - 各属性は離散値をとる
- 各属性は統計的に独立である
- クラスは、各属性の分布で特徴付けられる
  - 各クラスごと、A_i のとる値 $a_{i1},\ldots,a_{ik}$ に関する確率 $p_{i1},\ldots,p_{ik}$ が決まっている（これを決めるのが「学習」）

Naïve Bayes 分類器（続）

- 以上の仮定のもと

$$\begin{aligned} p(m_j \mid x) &= \frac{p(x \mid m_j)}{p(x)} p(m_j) \\ &\approx p(x \mid m_j) p(m_j) \\ &= p(a_1,\ldots,a_n \mid m_j) p(m_j) \\ &= p(m_j) \prod_{i=1}^{n} p(a_i \mid m_j) \end{aligned}$$

$$m_{\text{MAP}} = \arg\max_j p(m_j \mid x)$$



Naïve Bayes 分類器（続々）

- モデル m を記述するパラメータ（この場合は確率 $p_{i1},\ldots,p_{ik}$です）の推定は次のように行う。
- モデル m から生成されたデータを $<y_{j1},\ldots,y_{jn}>$ $(j=1,\ldots,N)$ としよう
- 属性 A_i $(i=1,\ldots,n)$ について、$y_{1i},\ldots,y_{Ni}$ のヒストグラムを作る。例えば、1,2,3 の3個の値をとるなら、1,2,3 の度数を数える。
- これを元に、p_{i1}, p_{i2}, p_{i3} を推定する。例えば、p_{i1}=1の度数/N, p_{i2}=2の度数/N, p_{i3}=3の度数/N というようにする。

目次

- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

簡単な例で: 天気とテニス

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

（テニスを行う）Play=Yes と（テニスを行わない）Play=No の2つのクラスがある

このとき、下記の（未知、つまり学習データにない）条件では、Play=Yesであった（あろう）かPlay=Noであった（あろう）かを推定する。

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Tom Mitchell の Machine Learning という書籍から. よく使われます

（回りくどいが）データをクラスに分割

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Rainy	Cool	Normal	True	No
Sunny	Mild	High	False	No
Rainy	Mild	High	True	No

37

データの数を数えて、推定

	A1=Outlook		A2=Temperature		A3=Humidity		A4=Windy	
度数	Sunny	2	Hot	2	High	3	False	6
	Overcast	4	Mild	4	Normal	6	True	3
	Rainy	3	Cool	3				
	合計	9	合計	9	合計	9	合計	9
確率の推定	Sunny	2/9	Hot	2/9	High	3/9	False	6/9
	Overcast	4/9	Mild	4/9	Normal	6/9	True	3/9
	Rainy	3/9	Cool	3/9				

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

	A1=Outlook		A2=Temperature		A3=Humidity		A4=Windy	
度数	Sunny	3	Hot	2	High	4	False	2
	Overcast	0	Mild	2	Normal	1	True	3
	Rainy	2	Cool	1				
	合計	5	合計	5	合計	5	合計	5
確率の推定	Sunny	3/5	Hot	2/5	High	4/5	False	2/5
	Overcast	0/5	Mild	2/5	Normal	1/5	True	3/5
	Rainy	2/5	Cool	1/5				

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Rainy	Cool	Normal	True	No
Sunny	Mild	High	False	No
Rainy	Mild	High	True	No

38

一つの表に纏めておこう

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Rainy	Cool	Normal	True	No
Sunny	Mild	High	False	No
Rainy	Mild	High	True	No

p(m) に関する説明を省きましたが（忘れた、が正しいのだが）、それは、これ

A1=Outlook			A2=Temperature			A3=Humidity			A4=Windy			m=Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

39

推論をしよう

$$p(m_j \mid x) = p(x \mid m_j)p(m_j) \Big/ p(x)$$
$$= p(a_1, \ldots, a_n \mid m_j)p(m_j) \Big/ p(x)$$
$$= \left(\prod_{i=1}^{n} p(a_i \mid m_j)\right) p(m_j) \Big/ p(x)$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

未知の x

p(Play=yes | x)
= p(Outlook=Sunny | Play=yes)
 * p(Temp=Cool | Play=yes)
 * p(Humidity=High | Play=yes)
 * p(Windy=True | Play=yes)
 * p(Play=yes) / p(x)
= (2/9) * (3/9) * (3/9) * (3/9)
 *(9/14) / p(x)
= 0.0053 / p(x)

p(Play=no | x)
= p(Outlook=Sunny | Play=no)
 * p(Temp=Cool | Play=no)
 * p(Humidity=High | Play=no)
 * p(Windy=True | Play=no)
 * p(Play=no) / p(x)
= (3/5) * (1/5) * (4/5) * (3/5)
 *(5/14) / p(x)
= 0.0206 / p(x)

言い換えれば、p(Play=yes | x) < p(Play=no | x)
すなわち、「テニスは行わなかった（行わないだろう）」

注: 1/p(x) は気にしなくてよいことが分る; 比較すべき相手すべてに共通だから.

40

目次

- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

41

Rでは？

```
# package e1071 をインストールした後、
> library(e1071)
> setwd("D:/R/Sample")
> xy<-read.csv("04PlayTennis.csv", header=TRUE)
> xyt<-read.csv("04PlayTennisTest01.csv",header=TRUE,as.is=TRUE)
> tt<-data.frame(factor(xyt[,1],levels=levels(xy[,1])))
> for (i in 2:5) {
>   tt<-data.frame(tt,factor(xyt[,i],levels=levels(xy[,i])))
> }
> names(tt)<-names(xy)
> tt
  Outlook Temp. Humidity Windy Play
1   Sunny  Cool     High  True <NA>
> m <- naiveBayes(xy[,-5], xy[,5])
> predict(m, tt)
[1] No
Levels: No Yes
>
```

xyt を直接（as.is=FALSEで）使うことができないのは、テストデータをfactorに変換するとき（read.csv内）に、xy のlevelsを参照するような指定ができないためである。上記のように手で変換せざるをえない。

42

補足1

Forループではなく、applyを使いたいのだが、levels が結合されてしまい、うまくいかない。

```
# package e107 をインストールした後、
> library(e1071)
> setwd("D:/R/Sample")
> xy<-read.csv("04PlayTennis.csv", header=TRUE)
> xyt<-read.csv("04PlayTennisTest01.csv", header=TRUE, as.is=TRUE)
> tt<-apply(as.data.frame(1:5),1,
        function(i) factor(xyt[,i],levels=levels(xy[,i])))
> tt
[1] Sunny Cool  High  True  <NA>
Levels: Overcast Rainy Sunny Cool Hot Mild High Normal False True No Yes
>
```


補足2

予測の確率を出力することもできる。type="raw" を加えればよい。
正規化（総和が1）した値が出力される。

```
> predict(m, tt, type="raw")
            No       Yes
[1,] 0.7954173 0.2045827
>
```


目次

- 今日扱う問題の特徴
- 復習
  - 条件付確率とベイズの定理
  - ベイズ推論
- ナイーブベイズ
  - ナイーブな記述ということ
  - 属性数について
  - 分類器
  - 簡単な例
  - Rでは
  - 学習誤差

パラメータと学習誤差（訓練誤差）

```
> m

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = xy[, -5], y = xy[, 5])

A-priori probabilities:
xy[, 5]
       No       Yes
0.3571429 0.6428571

Conditional probabilities:
       Outlook
xy[, 5]  Overcast     Rainy     Sunny
    No  0.0000000 0.4000000 0.6000000
    Yes 0.4444444 0.3333333 0.2222222

       Temp.
xy[, 5]      Cool       Hot      Mild
    No  0.2000000 0.4000000 0.4000000
    Yes 0.3333333 0.2222222 0.4444444

       Humidity
xy[, 5]      High    Normal
    No  0.8000000 0.2000000
    Yes 0.3333333 0.6666667

       Windy
xy[, 5]     False      True
    No  0.4000000 0.6000000
    Yes 0.6666667 0.3333333
```

confusion matrix:
（Weka と行・列が逆）

```
> table(predict(m, xy[,-5]), xy[,5])

      No Yes
  No   4   0
  Yes  1   9
>
```

Yesと予測した

正解はNo

A1=Outlook			A2=Temperature			A3=Humidity			A4=Windy			m=Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

今日の課題

- Naïve Bayes 法を用いて、下図左の訓練データが与えられたとき、下図右のテストデータの属性「スキー」の値を推定せよ。
- Rを使ってください。データはファイルに用意してあります。「確率」も出してください。

雪	天気	シーズン	体調	スキー
ベタ	霧	ロー	回復	no
新雪	晴	ロー	回復	yes
新雪	霧	ロー	回復	yes
ざらめ	霧	ロー	怪我	no
新雪	晴	ロー	怪我	no
ベタ	晴	ロー	回復	yes
新雪	霧	ロー	回復	yes
ベタ	晴	半ば	回復	yes
新雪	晴	ハイ	回復	yes
新雪	風	ロー	回復	yes
ざらめ	霧	半ば	回復	no
新雪	風	ロー	回復	yes
新雪	晴	半ば	回復	yes
ざらめ	風	ハイ	疲労	no

雪	天気	シーズン	体調	スキー
ベタ	風	半ば	疲労	?

今でも研究が

Francesco Sambo, Alberto Malovini, Niina Sandholm, Monica Stavarachi et al. (31 authors)
Novel genetic susceptibility loci for diabetic end-stage renal disease identified through robust naive Bayes classification
Diabetologia, 57(8), 1611-1622, 2014.

Jessica Minnier, Ming Yuan, Jun S. Liu & Tianxi Cai
Risk Classification with an Adaptive Naive Bayes Kernel Machine Model
Journal of the American Statistical Association, online: 08 Apr 2014

Nayyar A. Zaidi, Jesús Cerquides, Mark J. Carman, Geoffrey I. Webb
Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting
Journal of Machine Learning Research, 14(Jul):1947−1988, 2013.

Liangxiao Jiang, Zhihua Cai, Harry Zhang & Dianhong Wang
Naive Bayes text classifiers: a locally weighted learning approach
Journal of Experimental & Theoretical Artificial Intelligence, 25(2), 273-286, 2013

Sotiris Kotsiantis
Increasing the accuracy of incremental naive bayes classifier using instance based learning
International Journal of Control, Automation and Systems, 11(1), 159-166, 2013

まとめ

- ベイズ推論
  - 証拠を発生させた原因(モデル)の確率(条件付確率)を求め、それに基づき、原因(モデル)を推定する
- 困難点
  - (簡単な形のモデルが仮定できない場合)パラメータを決めるのに必要な学習データ数が膨大にある
- ナイーブベイズ
  - その困難点の解決方法の一つ
  - (データを表現する)属性が統計的に独立だとする
    - 実際には成立し得ない仮定であるが、結構うまく働く
- 今でも改善の研究が!

49