テキストマイニング入門
（ベイズ統計とその応用）

前田利之（まえだ　としゆき）
maechan@hannan-u.ac.jp

阪南大学　経営情報学部

(2024.12.11)

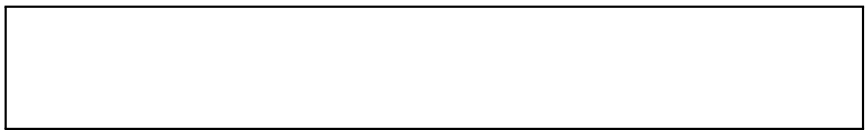
実践（復習）

```
rslt2 <- read.table("I:/rdata/hindo.txt")
c1 <- t(rslt2) # 転置（縦横入替）
c2 <- dist(c1) # ベクトルの距離計算
c3 <- cmdscale(c2,eig=T) # MDS
c4 <- c3$points # ポイントデータ
c5 <- kmeans(c2,2)
c6 <- c5$cluster
plot(c4,xlim=c(-30,30),t="n")
text(c4,rownames(c1),font=c6)
```

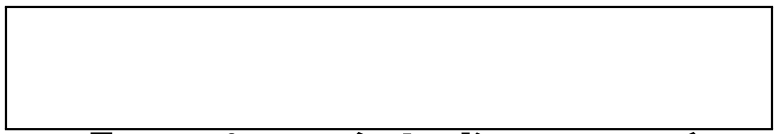
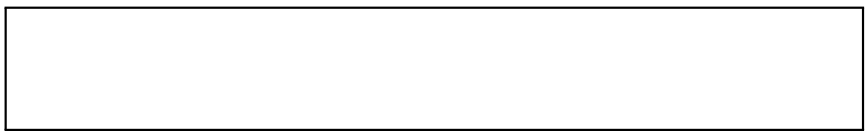
テキスト分類

- 与えられた文書（Web ページ、メール、等）を、あらかじめ与えられたいくつかの

  に自動で分類すること
  - Web ページを「政治・経済」「科学・学問」「コンピュータ・IT」「ゲーム・アニメ」などのカテゴリへ自動分類する 
  - 電子メールを「スパム（迷惑メール）」と「それ以外」というカテゴリへ自動分類して、スパムはゴミ箱へ捨てる 

テキスト分類

- 与えられた文書（Web ページ、メール、等）を、あらかじめ与えられたいくつかの **カテゴリ（クラス）** に自動で分類すること
  - Web ページを「政治・経済」「科学・学問」「コンピュータ・IT」「ゲーム・アニメ」などのカテゴリへ自動分類する 
  - 電子メールを「スパム（迷惑メール）」と「それ以外」というカテゴリへ自動分類して、スパムはゴミ箱へ捨てる 

テキスト分類

- 与えられた文書（Web ページ、メール、等）を、あらかじめ与えられたいくつかの **カテゴリ（クラス）** に自動で分類すること
  - Web ページを「政治・経済」「科学・学問」「コンピュータ・IT」「ゲーム・アニメ」などのカテゴリへ自動分類する **【まとめサイト】**
  - 電子メールを「スパム（迷惑メール）」と「それ以外」というカテゴリへ自動分類して、スパムはゴミ箱へ捨てる

テキスト分類

- 与えられた文書（Web ページ、メール、等）を、あらかじめ与えられたいくつかの **カテゴリ（クラス）** に自動で分類すること
  - Web ページを「政治・経済」「科学・学問」「コンピュータ・IT」「ゲーム・アニメ」などのカテゴリへ自動分類する **【まとめサイト】**
  - 電子メールを「スパム（迷惑メール）」と「それ以外」というカテゴリへ自動分類して、スパムはゴミ箱へ捨てる **【スパムフィルタ】**

テキスト分類（続き）

- [ ]：すでに分類された正解のデータでプログラムを訓練する
  - データは [ ] の場合が多い
- [ ]：非常によく使われていて、実装も簡単、処理速度は高速。精度評価の基準となることが多い

テキスト分類（続き）

- **教師あり学習**：すでに分類された正解のデータでプログラムを訓練する
  - データは [　　] の場合が多い
- [　　]：非常によく使われていて、実装も簡単、処理速度は高速。精度評価の基準となることが多い

テキスト分類（続き）

- **教師あり学習**：すでに分類された正解のデータでプログラムを訓練する
  - データは**語彙の頻度表** (Bag of Words) の場合が多い
- [　　]：非常によく使われていて、実装も簡単、処理速度は高速。精度評価の基準となることが多い

テキスト分類（続き）

- **教師あり学習**：すでに分類された正解のデータでプログラムを訓練する
  - データは**語彙の頻度表** (Bag of Words) の場合が多い
- **ナイーブベイズ** (Naïve Bayes)：非常によく使われていて、実装も簡単、処理速度は高速。精度評価の基準となることが多い

ベイズ統計とは

- ベイズ統計とは [　　] を元に、得られたデータから新たな [　　] する統計学
- [　　] 統計とは大きく異なった考え方
- [　　] 統計＝ [　　] ・ [　　]

ベイズ統計とは

- ベイズ統計とは **事前確率** を元に、得られたデータから新たな [　　] する統計学
- [　　] 統計とは大きく異なった考え方
- [　　] 統計＝ [　　] ・ [　　]

ベイズ統計とは

- ベイズ統計とは **事前確率** を元に、得られたデータから新たな **確率を導出** する統計学
- [　　] 統計とは大きく異なった考え方
- [　　] 統計＝ [　　] ・ [　　]

ベイズ統計とは

- ベイズ統計とは **事前確率** を元に、得られたデータから新たな **確率を導出** する統計学
- **頻度論的** 統計とは大きく異なった考え方
- [ ] 統計＝ [ ] ・ [ ]

ベイズ統計とは

- ベイズ統計とは **事前確率** を元に、得られたデータから新たな **確率を導出** する統計学
- **頻度論的** 統計とは大きく異なった考え方
- **頻度論的** 統計＝ [　　] ・ [　　]

ベイズ統計とは

- ベイズ統計とは **事前確率** を元に、得られたデータから新たな **確率を導出** する統計学
- **頻度論的** 統計とは大きく異なった考え方
- **頻度論的** 統計＝ **記述統計学** ・

ベイズ統計とは

- ベイズ統計とは **事前確率** を元に、得られたデータから新たな **確率を導出** する統計学
- **頻度論的** 統計とは大きく異なった考え方
- **頻度論的** 統計＝ **記述統計学** ・ **推計統計学**

統計学の分類

- 記述統計学：標本に見られる特徴をわかりやすく表す
- 推計統計学：標本を分析して、母集団について推測する
- ベイズ統計学：完全な標本を  しない＝不十分なデータでも何とかして確率を導く
  - ベイズ統計だけ、標本を必要としない
  - ベイズ統計学を支持する人は特別に“ベイジアン”と呼ばれる

統計学の分類

- 記述統計学：標本に見られる特徴をわかりやすく表す
- 推計統計学：標本を分析して、母集団について推測する
- ベイズ統計学：完全な標本を必ずしも必要としない＝不十分なデータでも何とかして確率を導く
  - ベイズ統計だけ、標本を必要としない
  - ベイズ統計学を支持する人は特別に“ベイジアン”と呼ばれる

頻度論とベイズ論の違い

- 頻度論：得られたデータが [　　] どれくらいの頻度（確率）で発生するのか、という基本的な考え方
- ＝パラメータが定数、データが変数（確率変数）
- ベイズ統計の考え方は逆＝パラメータが [　　]、データが [　　]
- 言い換えれば、既存のデータがどのようなパラメータに基づく [　　] から得られたのかを推測する、ということ

頻度論とベイズ論の違い

- 頻度論：得られたデータが **母集団から** どれくらいの頻度（確率）で発生するのか、という基本的な考え方
- ＝パラメータが定数、データが変数（確率変数）
- ベイズ統計の考え方は逆＝パラメータが [　　]、データが [　　]
- 言い換えれば、既存のデータがどのようなパラメータに基づく [　　] から得られたのかを推測する、ということ

頻度論とベイズ論の違い

- 頻度論：得られたデータが **母集団から** どれくらいの頻度（確率）で発生するのか、という基本的な考え方
- ＝パラメータが定数、データが変数（確率変数）
- ベイズ統計の考え方は逆＝パラメータが **変数（確率変数）**、データが
- 言い換えれば、既存のデータがどのようなパラメータに基づく から得られたのかを推測する、ということ

頻度論とベイズ論の違い

- 頻度論：得られたデータが **母集団から** どれくらいの頻度（確率）で発生するのか、という基本的な考え方
- ＝パラメータが定数、データが変数（確率変数）
- ベイズ統計の考え方は逆＝パラメータが **変数（確率変数）**、データが **定数**
- 言い換えれば、既存のデータがどのようなパラメータに基づく [　　] から得られたのかを推測する、ということ

頻度論とベイズ論の違い

- 頻度論：得られたデータが **母集団から** どれくらいの頻度（確率）で発生するのか、という基本的な考え方
- ＝パラメータが定数、データが変数（確率変数）
- ベイズ統計の考え方は逆＝パラメータが **変数（確率変数）**、データが **定数**
- 言い換えれば、既存のデータがどのようなパラメータに基づく **母集団** から得られたのかを推測する、ということ

ベイズ統計と学習能力

- [ ]：適当に（？）あてた確率＝事前確率（の初期値）
  - 例：「明日が雨になる確率は？」「２０％（５日に１回雨？）」
  - ここで、空がだんだん暗くなってきたので、５０％に修正する
  - 明朝になると、雨は降らなかったので１５％に修正する
  - 夜中になっても雨は降らなかったので０％に修正して確定する

ベイズ統計と学習能力

- **主観確率**：適当に（？）あてた確率＝事前確率（の初期値）
  - 例：「明日が雨になる確率は？」「２０％（５日に１回雨？）」
  - ここで、空がだんだん暗くなってきたので、５０％に修正する
  - 明朝になると、雨は降らなかったので１５％に修正する
  - 夜中になっても雨は降らなかったので０％に修正して確定する

ベイズ統計と学習能力

- 事前確率を設定した段階から、何か新たな情報をゲットして、事後確率を [ ] （例：Viagra という単語がメールに含まれていたら、それはスパムである確率が高い、等）
- また新たな情報を手に入れたら、前回は事後確率だったものが [ ] となって、さらに確率の更新が行われる
- これを繰り返すことで、真の値に近付くことが期待できる

ベイズ統計と学習能力

- 事前確率を設定した段階から、何か新たな情報をゲットして、事後確率を **更新**（例：Viagra という単語がメールに含まれていたら、それはスパムである確率が高い、等）
- また新たな情報を手に入れたら、前回は事後確率だったものが　　となって、さらに確率の更新が行われる
- これを繰り返すことで、真の値に近付くことが期待できる

ベイズ統計と学習能力

- 事前確率を設定した段階から、何か新たな情報をゲットして、事後確率を **更新**（例：Viagra という単語がメールに含まれていたら、それはスパムである確率が高い、等）
- また新たな情報を手に入れたら、前回は事後確率だったものが **事前確率** となって、さらに確率の更新が行われる
- これを繰り返すことで、真の値に近付くことが期待できる

ベイズの定理

$$P(cat|doc) = \frac{P(cat)P(doc|cat)}{P(doc)}$$

$$\propto P(cat)P(doc|cat)$$

- $\propto =$ [ ] … $P(doc)$ は固定（事象に共通）の確率なので、大小を比較するだけなら無視できるということ

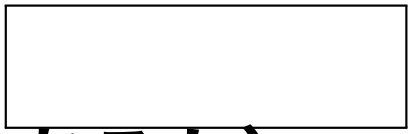
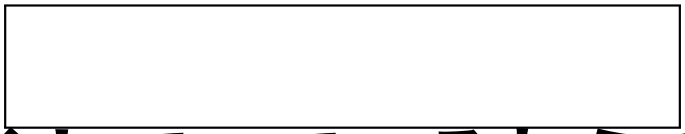
ベイズの定理

$$P(cat|doc) = \frac{P(cat)P(doc|cat)}{P(doc)}$$

$$\propto P(cat)P(doc|cat)$$

- $\propto$= **比例する** … $P(doc)$ **は固定（事象に共通）の確率なので、大小を比較するだけなら無視できるということ**

ベイズの定理（続き）

- $P(cat)$:  （もともとそのカテゴリがどれくらいあるか）
- $P(doc|cat)$:  ：観測データはすべて出尽くしていて、それらのデータに対してあるパラメータの確率分布を当てはめた時、どれだけ尤もらしいかを意味している（確率とは式の形は一緒だが意味するところはちょっと違う）
- $P(doc)$: そのデータ doc が得られる確率

ベイズの定理（続き）

- $P(cat)$: **事前確率**（もともとそのカテゴリがどれくらいあるか）
- $P(doc|cat)$: ：観測データはすべて出尽くしていて、それらのデータに対してあるパラメータの確率分布を当てはめた時、どれだけ尤もらしいかを意味している（確率とは式の形は一緒だが意味するところはちょっと違う）
- $P(doc)$: そのデータ doc が得られる確率

ベイズの定理（続き）

- $P(cat)$: **事前確率**（もともとそのカテゴリがどれくらいあるか）
- $P(doc|cat)$: **尤度（ゆうど）**：観測データはすべて出尽くしていて、それらのデータに対してあるパラメータの確率分布を当てはめた時、どれだけ尤もらしいかを意味している（確率とは式の形は一緒だが意味するところはちょっと違う）
- $P(doc)$: そのデータ doc が得られる確率

ベイズの定理（続き）

- $P(cat|doc)$: [ ] ＝文書 doc が与えられたときカテゴリ cat である確率
- つまり、事後確率は [ ] によって与えられる
- [ ]：カテゴリを予測したい未知の文書は、事後確率がもっとも高いカテゴリへ分類する

ベイズの定理（続き）

- $P(cat|doc)$: (cat に関する）事後確率＝文書 doc が与えられたときカテゴリ cat である確率
- つまり、事後確率は [ ] によって与えられる
- [ ]：カテゴリを予測したい未知の文書は、事後確率がもっとも高いカテゴリへ分類する

ベイズの定理（続き）

- $P(cat|doc)$: (cat に関する）事後確率＝文書 doc が与えられたときカテゴリ cat である確率
- つまり、事後確率は 尤度と事前確率 によって与えられる
- [ ]: カテゴリを予測したい未知の文書は、事後確率がもっとも高いカテゴリへ分類する

ベイズの定理（続き）

- $P(cat|doc)$: (cat に関する）事後確率＝文書 doc が与えられたときカテゴリ cat である確率
- つまり、事後確率は 尤度と事前確率 によって与えられる
- MAP 推定：カテゴリを予測したい未知の文書は、事後確率がもっとも高いカテゴリへ分類する

ベイズの定理（続き）

MAP 推定とは

- というものを導入する（唯一の厳密な解を探索することをあきらめる？）
- 有限の確率的に起こるデータからパラメータ自体も、集めたデータに依存する であると考える

ベイズの定理（続き）

MAP (Maximum a posteriori, **最大事後確率**) **推定とは**

- **[　　]というものを導入する（唯一の厳密な解を探索することをあきらめる？）**
- **有限の確率的に起こるデータからパラメータ自体も、集めたデータに依存する[　　]であると考える**

ベイズの定理（続き）

MAP (Maximum a posteriori, **最大事後確率**) 推定とは

- **統計パラメータ**というものを導入する（唯一の厳密な解を探索することをあきらめる？）
- 有限の確率的に起こるデータからパラメータ自体も、集めたデータに依存する　　であると考える

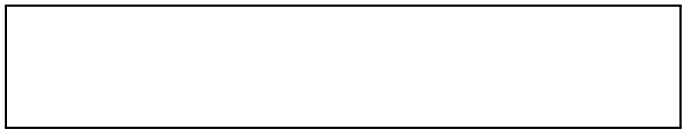
ベイズの定理（続き）

MAP (Maximum a posteriori, **最大事後確率**) **推定とは**

- **統計パラメータ**というものを導入する（唯一の厳密な解を探索することをあきらめる？）
- 有限の確率的に起こるデータからパラメータ自体も、集めたデータに依存する **確率的な変数** であると考える

ベイズの定理（続き）

MAP 推定とは（続き）

- 推定しようというパラメータ自身がデータに依存する確率変数であることを認める
- そうすれば、データが新たに手に入った時、その新たなデータを反映することで簡単にパラメータを更新  することが出来る

ベイズの定理（続き）

MAP 推定とは（続き）

- 推定しようというパラメータ自身がデータに依存する確率変数であることを認める
- そうすれば、データが新たに手に入った時、その新たなデータを反映することで簡単にパラメータを更新 **（ベイズ更新）** することが出来る

ベイズの定理（モンティホール問題）

「プレイヤーの前に 3 つのドアがあって、1 つのドアの後ろには景品の新車が、2 つのドアの後ろにはヤギ（はずれ）がいる。プレイヤーは新車のドアを当てると新車がもらえる。プレイヤーが 1 つのドアを選択した後、番組の司会者であるモンティが残りのドアのうちヤギがいるドアを開けてヤギを見せる。ここでプレイヤーは最初に選んだドアを、残っている開けられていないドアに変更してもよいと言われる。プレイヤーはドアを変更すべきだろうか？」

(wikipedia **より…この問題は答えを説明されても納得しない人が多い問題として有名です。多くの人は、動かしても動かなくても同じと考えるようです。**)

参考：午前０時の森(2023)

ベイズの定理（モンティホール問題）

確率論的に考える前に、まずシミュレーション（?!）

- http://puffin.hannan-u.ac.jp/lect/dats/monty.r **を** I:¥rdata **にダウンロードする**
- R **を立ち上げ** source('I:/rdata/monty.r') **として実行する**
- source('I:/rdata/monty.r') **の最後のコメント部分を実行する**

```
print(sprintf('Rate if not change: %1.4f',play(num,F)/num))
print(sprintf('Rate if change: %1.4f',play(num,T)/num))
```

- **...**

ベイズの定理（モンティホール問題）

確率論的に考える前に、まずシミュレーション（?!）

- http://puffin.hannan-u.ac.jp/lect/dats/monty.r **を** I:¥rdata **にダウンロードする**
- R **を立ち上げ** source('I:/rdata/monty.r') **として実行する**
- source('I:/rdata/monty.r') **の最後のコメント部分を実行する**

```
print(sprintf('Rate if not change: %1.4f',play(num,F)/num))
print(sprintf('Rate if change: %1.4f',play(num,T)/num))
```

- ... **実は、移動させるほうが有利（！）**

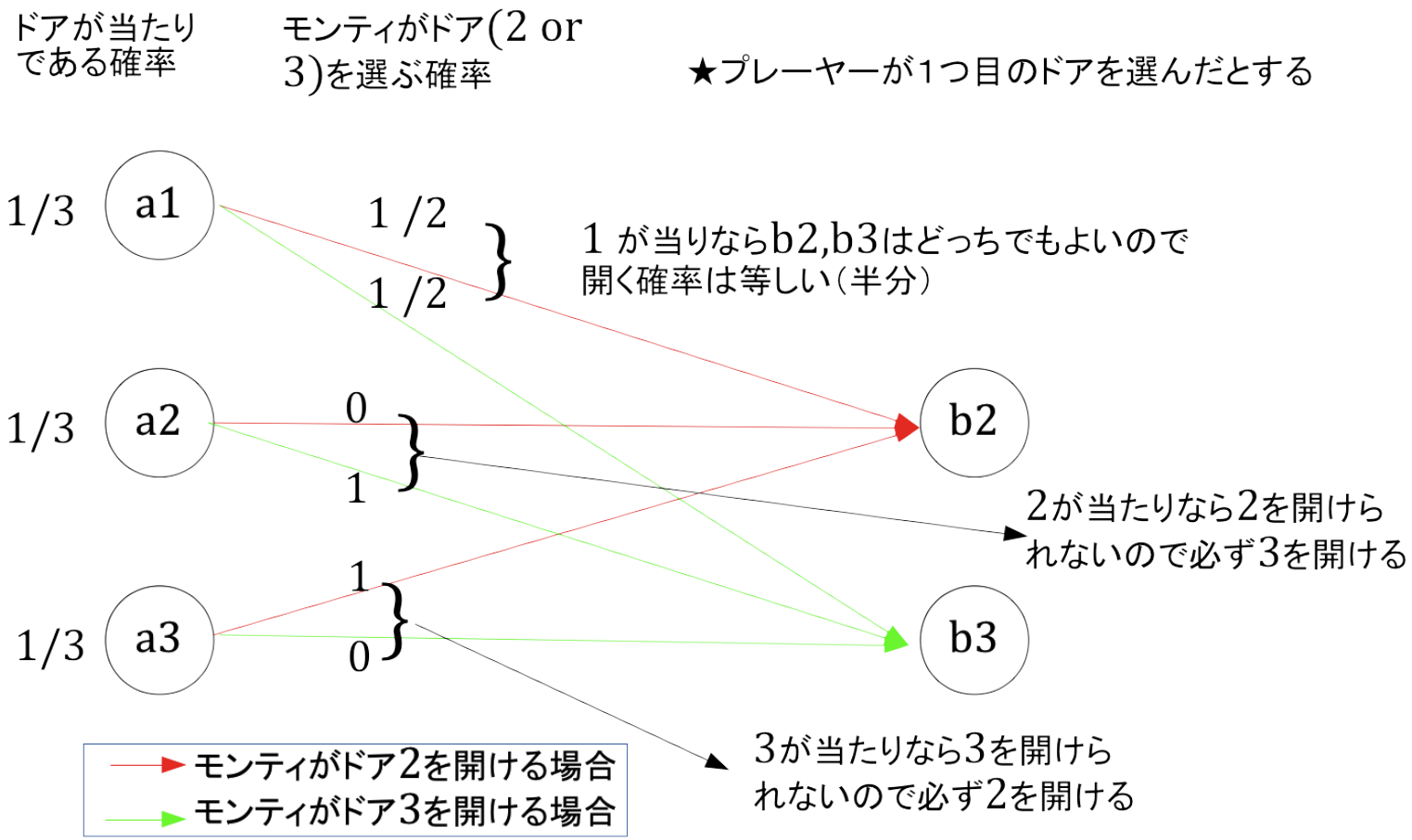
ベイズの定理（モンテカルロ法）

- **モンテカルロ法**（ランダム法とも呼ばれる）によるシミュレーション
- シミュレーションや数値計算を乱数を用いて行う手法の総称
- 数値解析の分野においてはモンテカルロ法はよく確率を近似的に求める手法として使われる

ベイズの定理（モンテカルロ法）

- （近似的）確率の求めかた
  - n 回シミュレーションを行いある事象が m 回起これば、その事象の起こる確率は m/n で近似される
  - 試行回数が少なければ近似は荒く、試行回数が多ければよい近似となる。

ベイズの定理（モンティホール問題）



ベイズの定理（モンティホール問題）

前提：プレーヤが１を選んでいてモンティがドア２を開けた場合、図の赤線 のどれかが起こっているので；

- 変更しないで当たる確率：

$$p(a_1|b_2) = \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \square$$

- 変更して当たる確率：

$$p(a_3|b_2) = \frac{\frac{1}{3}\cdot 1}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \square$$

よって変更したほうが □ （他の場合も同様）

ベイズの定理（モンティホール問題）

前提：プレーヤが１を選んでいてモンティがドア２を開けた場合、図の赤線 のどれかが起こっているので；

- 変更しないで当たる確率：

$$p(a_1|b_2) = \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \boxed{\frac{1}{3}}$$

- 変更して当たる確率：

$$p(a_3|b_2) = \frac{\frac{1}{3}\cdot 1}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \boxed{}$$

よって変更したほうが ☐ （他の場合も同様）

ベイズの定理（モンティホール問題）

前提：プレーヤが１を選んでいてモンティがドア２を開けた場合、図の赤線 のどれかが起こっているので；

- 変更しないで当たる確率：

$$p(a_1|b_2) = \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \boxed{\frac{1}{3}}$$

- 変更して当たる確率：

$$p(a_3|b_2) = \frac{\frac{1}{3}\cdot 1}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \boxed{\frac{2}{3}}$$

よって変更したほうが [　　] （他の場合も同様）

ベイズの定理（モンティホール問題）

前提：プレーヤが１を選んでいてモンティがドア２を開けた場合、図の赤線 のどれかが起こっているので；

- 変更しないで当たる確率：

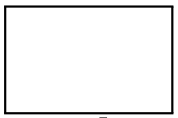
$$p(a_1|b_2) = \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \boxed{\frac{1}{3}}$$

- 変更して当たる確率：

$$p(a_3|b_2) = \frac{\frac{1}{3}\cdot 1}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \boxed{\frac{2}{3}}$$

よって変更したほうが **確率が高い**（他の場合も同様）

ベイズの定理（モンティホール問題）

- ※もっと直感的に言うと、２回目に必ず替える方針だと：
  - １回目で当りだった場合（１つ）：替えるので必ず「外れる」→確率は 
  - １回目で外れだった場合（２つ）：替える対象は必ず当りなので「当る」→確率は 
- よって、替えると  で当りを選べることになる

ベイズの定理（モンティホール問題）

- ※もっと直感的に言うと、２回目に必ず替える方針だと：
  - １回目で当りだった場合（１つ）：替えるので必ず「外れる」→確率は 1/3
  - １回目で外れだった場合（２つ）：替える対象は必ず当りなので「当る」→確率は □
- よって、替えると □ で当りを選べることになる

ベイズの定理（モンティホール問題）

- ※もっと直感的に言うと、２回目に必ず替える方針だと：
  - １回目で当りだった場合（１つ）：替えるので必ず「外れる」→確率は 1/3
  - １回目で外れだった場合（２つ）：替える対象は必ず当りなので「当る」→確率は 2/3
- よって、替えると □ で当りを選べることになる

ベイズの定理（モンティホール問題）

- ※もっと直感的に言うと、２回目に必ず替える方針だと：
  - １回目で当りだった場合（１つ）：替えるので必ず「外れる」→確率は 1/3
  - １回目で外れだった場合（２つ）：替える対象は必ず当りなので「当る」→確率は 2/3
- よって、替えると 2/3 で当りを選べることになる

ナイーブベイズ

- [ ]に基づいた単純な識別器
- [ ]（項目間の相関が無い）を仮定している
- 学習が非常に[ ]で、かつ[ ]が出る
- スパムフィルタなどに利用されている

ナイーブベイズ

- **ベイズの定理**に基づいた単純な識別器
- [　　]（項目間の相関が無い）を仮定している
- 学習が非常に[　　]で、かつ[　　]が出る
- スパムフィルタなどに利用されている

ナイーブベイズ

- **ベイズの定理**に基づいた単純な識別器
- **条件付き独立**（項目間の相関が無い）を仮定している
- 学習が非常に [　　] で、かつ [　　] が出る
- スパムフィルタなどに利用されている

ナイーブベイズ

- **ベイズの定理**に基づいた単純な識別器
- **条件付き独立**（項目間の相関が無い）を仮定している
- 学習が非常に**高速**で、かつ　　が出る
- スパムフィルタなどに利用されている

ナイーブベイズ

- **ベイズの定理**に基づいた単純な識別器
- **条件付き独立**（項目間の相関が無い）を仮定している
- 学習が非常に**高速**で、かつ**実用的な精度**が出る
- スパムフィルタなどに利用されている

ナイーブベイズ（実践）

1. R で用意されている、形態素解析・集計済みのデータセットを用いるため、２つのパッケージ (e1071 ライブラリ, spam データセット) をインストールし、ライブラリを利用可能とする
2. R を立ち上げ、以下を実行する

```
install.packages("e1071") # naiveBayes
install.packages("kernlab")  # spam データは ke

library("e1071")
library("kernlab")
data(spam)
```

ナイーブベイズ（実践）

3. 引き続き、以下を実行する

```
n <- 10*(1:(nrow(spam)/10))
# 10のうち9を学習用とするための数列を作っている
# 半分にする場合は n <- seq(1,nrow(spam),by=2)
train <- spam[-n,] # 学習（訓練）用データ
test <- spam[n,]   # 評価（テスト）用データ
num <- 58 # (no)spam のタイプが書いてある列番号
```

ナイーブベイズ（実践）

4 引き続き、以下を実行する

```
# 学習モデルを作る
mdl <- naiveBayes(type~.,data=train)
# 上で作ったモデルで判定をしてみる
pred <- predict(mdl,test[,-num])
t <- table(test[,num],pred)
a <- (t[1,1]+t[2,2]) /          # 正解率を求める
  (t[1,1]+t[1,2]+t[2,1]+t[2,2]) #本当は１行
rslt <- round(a,3) # 3桁にまとめる
cat(rslt) # 結果の出力 ((0.72 になるはず))
```

おしまい

質問があれば、お気軽に！