知的情報処理
4. ナイーブなベイズ法 補足
(Laplace correction)

慶應義塾大学理工学部
櫻井 彰人

前置き

- Laplace correction については、講義資料「4. ナイーブなベイズ法」の p.5 右上のスライドに記し、また、講義中でも説明した。
- しかし、聞き落とした人がいるようなので、以下、説明を追加しようと思う。

- しかし、ただ、追加するんじゃ、解答を示すようなもので、ちょっと癪だし、聞いていた人には不公平だろう。
- というわけで、ちょっと意地悪なんだが、説明を少し詳しくし、余分な話題を入れて、目くらましとしてみました。
- どうだろうか？

Laplace's Law of Succession

- ラプラスの「連の法則」とでも訳しましょうか。
- 同じ事象がn回連続して起こった場合に、その事象がn+1回目も起こる確率は、$(n+1)/(n+2)$ である。

- この定理の不思議さ:
  - 普通に、丁半勝負を考えよう。
  - 丁ばかり10回続いた。丁が出る確率は最尤推定すると 10/10 = 1 である。従って、次に丁となる確率は1である。
    - まぁ、10回も続けば、いかさまサイコロだと思うよね。
  - しかし、ラプラスによれば、11/12

最尤推定 vs Laplace

- 10回などといわずに、2回にしてみよう。すなわち、丁が2回続いたとしよう。
- 最尤推定なら、丁が出る確率は 2/2=1ゆえ、次に丁の出る確率は1。
- Laplaceによれば、3/4。
- どっちが正しい？

- では、丁が1回続いた（続いてない？）
- 最尤推定なら、丁の出る確率は1/1=1ゆえ、次に丁の出る確率は1。
- Laplaceによれば、2/3。

- さて、どちらが正しそうか？

証明

- ここでは、丁半勝負、ではなかった、ベルヌーイ試行 (coin-tossing) を考える。
- ただしコインはN+1種類あり（表の出る確率は異なる）、どのコインであるかは、出題者は知っているが回答者は知らないとする。
- i番目のコインの表の出る確率は i/N (i=0,..,N) とする。
- 出題者は、N+1個のコインを等確率で選び、一回選べば、あとはそれを使い続けるとする。
- n回連続して表が出たときに、n+1回目も表が出る確率を求めたい。

- コインiを選んだという条件下で、n回連続して表が出る確率は $(i/N)^n$ である。従って、n回連続して表が出る確率は $(1/(N+1))\sum_{i=0}^{N}(i/N)^n$ これを $F_{N,n}$ と表す。
- n+1回連続して表が出る確率は $F_{N,n+1} =(1/(N+1))\sum_{i=0}^{N}(i/N)^{n+1}$ である
- 従って、求める確率は $F_{N,n+1} / F_{N,n}$ である（式は複雑）。そこで次頁。

証明（続）

- $N\to\infty$ の極限を考えよう。
  - これは、コインの表が出る確率が、[0,1] 上の一様分布に従うことを意味する。
- $F_{N,n} =(1/(N+1))\sum_{i=0}^{N}(i/N)^n \to \int_0^1 x^n\, dx =1/(n+1)$
- 従って、求める確率は、$N\to\infty$ のとき、$F_{N,n+1}/F_{N,n} \to F_{\infty,n+1}/F_{\infty,n} = (n+1)/(n+2)$

- つまり、より正確に述べれば、Laplace's law of succession は、（coin tossingを例にすれば）出題者は、表が出る確率を一様分布に従ってランダムに決め、それを用いてcoin tossをする、一方回答者は、表が出る確率を知らないで（しかし、一様分布に従っていることは知っている）、n回連続して表が出たときに、n+1回目も表が出る確率を求めるという状況において、その確率は、$(n+1)/(n+2)$ であると主張している。

補足: 誤った証明(?)

- コインiが選ばれたとき、n回連続して表が出る確率は $(i/N)^n$ であり、$n+1$回連続して表が出る確率は $(i/N)^{n+1}$ である。
- 従って、コインiが選ばれたとき、n回連続して表が出た後、$n+1$回目も表が出る確率は、$(i/N)^{n+1}$ / $(i/N)^n$ = i/N である。
- コインiが選ばれる確率は $1/(N+1)$ であるから、n回連続して表が出た後、$n+1$回目も表が出る確率は $(1/(N+1))\sum_{i=0}^{N}(i/N) = ½$ である。

補足: 式で書いて比較しよう

- $P(n+1$回表 $\mid n$回表$)$
  $= P(n+1$回表$, n$回表$) / P(n$回表$)$
  $= P(n+1$回表$) / P(n$回表$)$
- $P(n$回表$)$
  $= \sum_{i=0}^{N} P(n$回表 $\mid$ コイン $i) P($コイン$i)$
  $= \sum_{i=0}^{N}(i/N)^n (1/(N+1)) \to 1/(n+1)$
- $P(n+1$回表 $\mid n$回表$) \to (n+1)/(n+2)$

どこに間違いがあるか？

- $P(n+1$回表 $\mid n$回表 $)$
  $= \sum_{i=0}^{N} P(n+1$回表 $\mid n$回表, コイン $i) P($ コイン$i)$
  $= \sum_{i=0}^{N} P($コイン$i) P(n+1$回表$, n$回表, コイン $i) / P(n$回表, コイン $i)$
  $= \sum_{i=0}^{N} P($コイン$i) P(n+1$回表, コイン $i) / P(n$回表, コイン $i)$
  $= (1/(N+1))\sum_{i=0}^{N}(i/N)$
  $= ½$

Laplace correction

Laplace smoothing とも言います

- 先ほどの問題で、最尤法を用いてより直感に合う結果を得る方法でもある。
- 再び最尤推定で考えよう。
- 最初からn回続けて表が出れば、表が出る確率の最尤推定量は $n/n = 1$.
- しかし、これはおかしい。裏が出る確率が0というのはありえない。仮に、回答者が見る前に、表・裏一回ずつ出たものとしてみよう（これが Laplace correction）。
- そうすると、全部で $n+2$回試行し、$n+1$回表が出たことになるので、表が出る確率の最尤推定量は、$(n+1)/(n+2)$

- Laplace's law of succession の結果とぴったり合う！

パラメータの事前分布

- 最尤推定ではなく、事後確率分布によるパラメータの期待値を推定値としてみよう。すなわち、表が出る確率 p を
  $\bar{p} = \int_0^1 p\, P(p \mid n$回表 $)\, dp$
  で推定したい。

  –ではなく ∧ にしたいのですが、できませんでした。

- この推定を行うには、パラメータ p の事前分布を知る必要がある。仮に、[0,1]上の一様分布だと考えよう。そうすると、
  $\bar{p} = \int_0^1 p\, P(n$回表 $\mid p)\, P(p)/P(n$回表$)\, dp$
  $= \int_0^1 p^{n+1} / (\int_0^1 p^n\, dp)\, dp$
  $= (n+1)/(n+2)$

  $P(n$回表$)$
  $= \int_0^1 P(n$回表 $\mid p)P(p)dp$
  $= \int_0^1 p^n\, dp$

- これはLaplace's law of successionの結果に一致する！
- すなわち、Laplace's law of succession の結果は（Laplace correctionの結果は）、パラメータの事前分布を一様分布と仮定し、事後分布によるパラメータの期待値を推定値とすることに相当する！

パラメータの事前分布（続）

- 一様分布では気に食わない？
  - 誰でもそう思う。
- では、ベータ分布 $f(x;\alpha,\beta)=x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ を考えよう。なお$f(x;1,1)$が一様分布である。
  - なぜベータ分布か？って？ 二項分布の双対分布だから。
  - なぜ双対分布か？って？ そうでないと計算ができないから、、、、
- このとき、事後分布によるパラメータの期待値を推定値とすると、その値は、$(n+\alpha)/(n+\alpha+\beta)$ となる

- これをLaplace correction の拡張として用いることができる。$(n+1)/(n+2)$ の代わりに $(n+1/2)/(n+1)$ や$(n+1/3)/(n+2/3)$ などを用いる方法となる。
- もっとも、そうなると、どれを使ったらよいのだろうか？という迷いが生まれる
  - 選択肢が多いと困ることもある
- 実は、（理論的な）答えは、ない