離散数理工学　第 10 回
離散確率論：乱択データ構造とアルゴリズム (発展)

岡本 吉央
okamotoy@uec.ac.jp

電気通信大学

2019 年 1 月 17 日

最終更新：2019 年 1 月 15 日　14:35

スケジュール 前半

スケジュール 後半 (予定)

注意：予定の変更もありうる

今日の目標

今日の目標

典型的な乱択アルゴリズムの設計と解析ができるようになる
- 確率の推定 (中央値トリック)

目次

不公平な硬貨

設定
- 硬貨が 1 つある
- 投げたとき，表が出る確率はいつも変わらない
- その確率が分からない
- $\boxed{\text{目標}}$：表が出る確率を知りたい
- 可能な操作：硬貨を投げる (ことのみ)

不公平な硬貨

設定
- 考えている硬貨について

$$\Pr(\text{表}) = p$$

ただし，$0 \le p \le 1$
- $\boxed{\text{目標}}$：p を知りたい

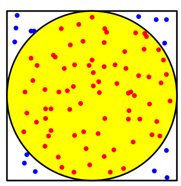
応用例：モンテカルロ法

モンテカルロ法：(実際には用いられない) 例

円周率の計算のために次を行う
1. $[-1,1]^2$ 内の点 (x,y) を一様分布に従って発生させる
2. $x^2+y^2 \le 1$ ならば，1 を出力，そうでなければ 0 を出力

このとき，この方法が 1 を出力する確率 $= \pi/4$

つまり，$p = \pi/4$ とした不公平な硬貨を考えていることになる



モンテカルロ法は，次の「単純なアルゴリズム」を実行する

単純なアルゴリズム

単純なアルゴリズム

硬貨を何度も投げてみる

- n 回投げるとする (独立な試行)
- 確率変数 X_i を次で定義 ($i \in \{1,\ldots,n\}$)

$$X_i = \begin{cases} 0 & (i \text{ 回目に投げたとき裏が出る}) \\ 1 & (i \text{ 回目に投げたとき表が出る}) \end{cases}$$

- 次の量を出力

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

期待値の解析

- 出力の期待値は

$$\begin{aligned} \mathrm{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} (0\cdot(1-p) + 1\cdot p) \\ &= p \end{aligned}$$

期待値は正しい「推測」になっている

問題点

必ず「p」を出力するわけではない $\rightsquigarrow$ 誤差が出る

n を大きくすれば，誤差は小さくなりそう

誤差の解析 (1)

以後，$X = X_1 + X_2 + \cdots + X_n$ とする

- 真の値 p から出力 $\frac{X}{n}$ がどれだけずれるか？
- そのずれが ε 未満である確率を知りたい
- その確率は次のように書ける

$$\Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right)$$

- 計算

$$\begin{aligned} \Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) &= 1 - \Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right), \\ \Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right) &\leq \frac{\mathrm{E}\left[\left|\frac{X}{n} - p\right|\right]}{\varepsilon} \quad (\text{マルコフの不等式}) \end{aligned}$$

しかし，$\mathrm{E}\left[\left|\frac{X}{n} - p\right|\right]$ はどう計算したらいいか分からない

誤差の解析 (2)

$$\Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right) = \Pr\left(\left|\frac{X}{n} - p\right|^2 \geq \varepsilon^2\right) \leq \frac{\mathrm{E}\left[\left|\frac{X}{n} - p\right|^2\right]}{\varepsilon^2}$$

$\mathrm{E}\left[\left|\frac{X}{n} - p\right|^2\right]$ を計算してみる

誤差の解析 (3)

$$\mathrm{E}\left[\left|\frac{X}{n} - p\right|^2\right] = \mathrm{E}\left[\left(\frac{X}{n}\right)^2 - 2p\frac{X}{n} + p^2\right] = \frac{1}{n^2}\mathrm{E}[X^2] - \frac{2p}{n}\mathrm{E}[X] + p^2$$

ここで，

$$\begin{aligned} \mathrm{E}[X] &= \mathrm{E}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} \mathrm{E}[X_i] = pn \\ \mathrm{E}[X^2] &= \mathrm{E}[(X_1 + \cdots + X_n)^2] = \sum_{i=1}^{n} \mathrm{E}[X_i^2] + \sum_{i=1}^{n} \sum_{j=1,(i\neq j)}^{n} \mathrm{E}[X_i X_j] \end{aligned}$$

誤差の解析 (4)

任意の $i \in \{1,\ldots,n\}$ に対して

$$\mathrm{E}[X_i^2] = (1-p)\cdot 0^2 + p \cdot 1^2 = p$$

したがって，

$$\sum_{i=1}^{n} \mathrm{E}[X_i^2] = pn$$

任意の異なる $i, j \in \{1,\ldots,n\}$ に対して，X_i と X_j は独立なので，

$$\mathrm{E}[X_i X_j] = \mathrm{E}[X_i]\mathrm{E}[X_j] = p \cdot p = p^2$$

したがって，

$$\sum_{i=1}^{n} \sum_{j=1,(i\neq j)}^{n} \mathrm{E}[X_i X_j] = p^2 n(n-1)$$

誤差の解析 (5)

ここまで，まとめると

$$\begin{aligned} \mathrm{E}\left[\left|\frac{X}{n} - p\right|^2\right] &= \frac{1}{n^2}\mathrm{E}[X^2] - \frac{2p}{n}\mathrm{E}[X] + p^2 \\ &= \frac{1}{n^2}(pn + p^2 n(n-1)) - \frac{2p}{n}pn + p^2 \\ &= \frac{p}{n} + \frac{p^2(n-1)}{n} - p^2 \\ &= \frac{p - p^2}{n} \\ &= \frac{p(1-p)}{n} \end{aligned}$$

誤差の解析 (6)

すなわち，

$$\Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right) \leq \frac{\mathrm{E}\left[\left|\frac{X}{n} - p\right|^2\right]}{\varepsilon^2} = \frac{1}{\varepsilon^2}\frac{p(1-p)}{n}$$

- この不等式はチェビシェフの不等式と呼ばれる (ものの特殊な場合)
- この右辺を δ 以下にするには，$n \geq \frac{1}{\varepsilon^2}\frac{p(1-p)}{\delta}$ とすればよい

結論

誤差が ε 以上になる確率を δ 以下とするためには，

$$n \geq \frac{1}{\varepsilon^2}\frac{p(1-p)}{\delta}$$

とすればよい

疑問

単純なアルゴリズム

硬貨を何度も投げてみる

- n 回投げるとする (独立な試行)
- 確率変数 X_i を次で定義 ($i \in \{1, \dots, n\}$)　　(標示確率変数)

$$X_i = \begin{cases} 1 & (i \text{ 回目に投げたとき表が出る}) \\ 0 & (i \text{ 回目に投げたとき裏が出る}) \end{cases}$$

- 次の量を出力

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

疑問

この「単純なアルゴリズム」よりもよいアルゴリズムは無いのか？

目次

中央値トリック (median trick)

中央値トリック

- n 回投げるとする (独立な試行)
- $n = (2k-1)t$ とする (k, t は自然数)
- 確率変数 X_i を次で定義 ($i \in \{1, \dots, n\}$)

$$X_i = \begin{cases} 0 & (i \text{ 回目に投げたとき裏が出る}) \\ 1 & (i \text{ 回目に投げたとき表が出る}) \end{cases}$$

- 確率変数 Y_j を次で定義 ($j \in \{1, \dots, 2k-1\}$)

$$Y_j = \frac{X_{(j-1)t+1} + \cdots + X_{(j-1)t+t}}{t}$$

- 次の量を出力

$$Y = \mathrm{med}\{Y_1, \dots, Y_{2k-1}\}$$

med は中央値：$\mathrm{med}\{5, 1, 6, 2, 4\} = 4$

中央値トリック：誤差の解析 (1)

- 次が成り立つために n が満たす条件を見つけたい

$$\Pr(|Y - p| \geq \varepsilon) \leq \delta$$

- 今までの議論から，任意の $j \in \{1, \dots, 2k-1\}$ に対して

$$\Pr(|Y_j - p| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\frac{p(1-p)}{t}$$

- $t \geq \dfrac{8p(1-p)}{\varepsilon^2}$ とすると，$\dfrac{1}{t} \leq \dfrac{1}{8}\dfrac{\varepsilon^2}{p(1-p)}$ なので，

$$\Pr(|Y_j - p| \geq \varepsilon) \leq \frac{1}{8}$$

中央値トリック：誤差の解析 (2)

- このとき，合併上界から

$$\Pr(k \text{ 個の } j \text{ に対して}, |Y_j - p| \geq \varepsilon) \leq \binom{2k-1}{k}\left(\frac{1}{8}\right)^k$$

- 二項係数に対する上界を用いて，右辺を整理すると

$$\binom{2k-1}{k}\left(\frac{1}{8}\right)^k \leq \left(\frac{\mathrm{e}(2k-1)}{k}\right)^k\left(\frac{1}{8}\right)^k < \left(\frac{2\,\mathrm{e}}{8}\right)^k < \left(\frac{3}{4}\right)^k$$

二項係数：簡単な評価　(第 1 回講義より)

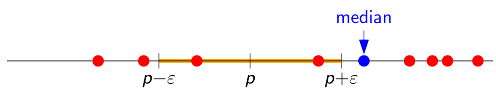
任意の自然数 $a \geq 1$ と任意の自然数 $b \geq 1$ に対して，$a \geq b$ であるとき，

$$\left(\frac{a}{b}\right)^b \leq \binom{a}{b} \leq \left(\frac{\mathrm{e}\,a}{b}\right)^b$$

中央値トリック：誤差の解析 (3)

- したがって (演習問題 11.3 参照)，

$$\begin{aligned}\Pr(|Y - p| \geq \varepsilon) &\leq \Pr(k \text{ 個の } j \text{ に対して}, |Y_j - p| \geq \varepsilon) \\ &< \left(\frac{3}{4}\right)^k\end{aligned}$$



- $k \geq \log_{3/4}\delta$ とすると

$$\Pr(|Y - p| \geq \varepsilon) < \left(\frac{3}{4}\right)^k \leq \delta$$

中央値トリック：誤差の解析 (まとめ)

中央値トリック：誤差の解析 (まとめ)

- $t \geq \dfrac{8p(1-p)}{\varepsilon^2}$，$k \geq \log_{3/4}\delta$ とすると
  誤差が ε 以上になる確率を δ 以下にできる
- このとき，硬貨を投げる回数 n は

$$n = (2k-1)t \geq \Omega\left(\frac{p(1-p)}{\varepsilon^2}\log\frac{1}{\delta}\right)$$

補足：単純なアルゴリズムにて，硬貨を投げる回数 n は

$$n \geq \frac{p(1-p)}{\varepsilon^2}\frac{1}{\delta}$$

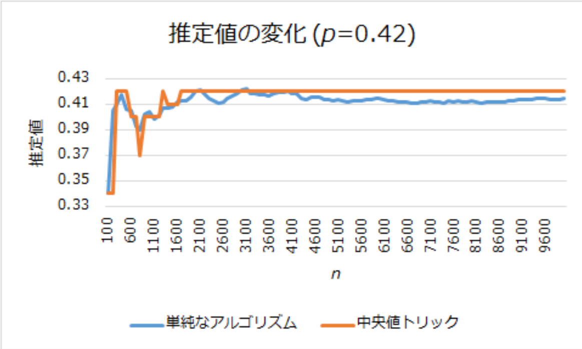
つまり，中央値トリックにより，硬貨を投げる回数が減った

実験してみた

- パラメータ
  - $p = 0.42$
  - $t = 100$
  - k を変化させることで，$n = (2k-1)t$ を変化させる
- Ruby 2.1.4 で実装
- n の増加に関する推定値の変化を図示してみた
- 横軸が n，縦軸が推定した p の値

注意：このパラメータ設定はとても恣意的なので，
他のパラメータ設定で追試をしてみるとよい

実験してみた：結果



中央値トリック：注意

注意

- 単純なアルゴリズムの出力 X/n に対して，$\mathrm{E}[X/n] = p$ が成り立つ
- 中央値トリックの出力 Y に対して，$\mathrm{E}[Y] = p$ が成り立つとは限らない

X/n は不偏推定量であるが，Y はそうではない

どういうことか？

- n を大きくすると，X/n は p に近づいていく
- n を大きくすると，Y は p の近似値に近づいていく

目次

今日の目標

今日の目標

典型的な乱択アルゴリズムの設計と解析ができるようになる

- 確率の推定 (中央値トリック)

残った時間の使い方

- 演習問題をやる
  - 相談推奨 (ひとりでやらない)
- 質問をする
  - 教員と TA は巡回
- 退室時，小さな紙に感想など書いて提出する ← 重要
  - 内容は何でも OK
  - 匿名で OK