確率の基礎の復習

植野真臣
電気通信大学
情報理工学研究科
情報数理工学プログラム

スケジュール（予定）

4月11日　授業の概要とガイダンス
4月18日　ベイズの定理
4月25日　ベイズの定理はどのように誕生したのか？
5月2日　ベイズはコンピュータ、人工知能の父である！！
5月9日　アランチューリングとベイズ
5月16日　ベイズから機械学習へ
5月23日　確率の基礎の復習
5月30日　ビリーフとベイズ
6月6日　尤度と最尤推定
6月13日　数値計算法による推定
6月20日　ベイズ推定と事前分布
6月27日　マルコフチェーンモンテカルロ(MCMC)法
7月4日　ベイジアンネットワーク
7月11日　ベイジアンネットワークと機械学習
7月25日　テストと総括

1．頻度論による確率

コインを何百回も投げて表が出た回数（頻度）を数えて，その割合を求めることを考えよう．いま，投げる回数をn とし，表の出た回数n_1とすると，$n\to\infty$のとき，$\frac{n_1}{n}\to\frac{1}{2}$

となることが予想される．このように，何回も実験を繰り返してn 回中，事象A が n_1 回出たとき，$\frac{n_1}{n}$ をA の確率と解釈するのが頻度主義である．

しかし，この定義では真の確率は無限回実験をしなければならないので得ることは不可能である．

2．主観確率

例えば，以下のような主観確率の例がある．

1. 第三次世界大戦が20XX 年までに起こる確率が0.01
2. 明日，会社の株式の価格が上がる確率が0.35
3. 来年の今日，東京で雨が降る確率が0.5

ベイズ統計では，これらの主観確率は個人の意思決定のための信念として定義され，ビリーフ（belief）と呼ばれる．当然，頻度論的確率を主観確率の一種とみなすことができるが，その逆は成り立たない．

3. 条件付き確率

定義3 (条件付き確率)
$A\in\mathcal{A}$,$B\in\mathcal{A}$について，事象B が起こったという条件の下で，事象A が起こる確率を条件付き確率（conditional probablity）と呼び，

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)}$$

で示す．

4. 同時確率

$P(A\mid B)=\frac{P(A\cap B)}{P(B)}$より以下の乗法公式が成り立つ．

定理5 (乗法公式)
$P(A\cap B)=P(A\mid B)P(B)$
このとき，$P(A\cap B)$ をA とB の**同時確率**（joint probability）と呼ぶ．

5. 独立性

定義4 (独立)
ある事象の生起する確率が，他のある事象が生起する確率に依存しないとき，二つの事象は独立（independent）であるという．すなわち事象 A と事象 B が独立とは $P(A \mid B) = P(A)$ であり，

$P(A \cap B) = P(A)P(B)$

が成り立つことをいう．

6. チェーンルール

さらに乗法公式を一般化すると以下のチェーンルールが導かれる．
$P(A \cap B \cap C) = P(A \mid B \cap C)P(B \mid C)P(C)$
3 個以上の事象にも拡張できるので，チェーンルール（chain rule）は
以下のように書ける．

6. チェーンルール

定理6 チェーンルール
N 個の事象 $\{A_1, A_2, \cdots, A_N\}$について
$P(A_1 \cap A_2 \cap \cdots \cap A_N)$
$= P(A_1 / A_2 \cap A_3 \cap \cdots \cap A_N)P(A_2 / A_3 \cap A_4 \cap \cdots \cap$

7. 全確率の定理

定理7 (全確率の定理（total probability theorem）)

たがいに背反な事象 $A_1, A_2, \cdots, A_n$ $(A_i \in \mathcal{A})$ が全事象 Ω を分割しているとき，
事象 $B \in \mathcal{A}$ について，
$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$ が成り立つ．

8. ベイズの定理

定理8 (ベイズの定理（Bayes' theorem）)
たがいに背反な事象 $A_1, A_2, \cdots, A_n$ が全事象 Ω を分割しているとする．
このとき，事象 $B \in \mathcal{A}$ について，

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

が成り立つ．

例題1

キリストの弟子たちはキリストの復活を望んでいました。あまりに臨みが強すぎて少し似ているだけの人でもキリストに見えてしまうことがあります。弟子がキリストの復活を見たと証言する事象をA, 実際にキリストが復活したという事象をBとする．$\mathrm{P}(A|\mathrm{B}) = 1.0$, $\mathrm{P}(A|\neg\mathrm{B}) = 0.5$, $\mathrm{P}(B) = 0.000001$ とする．ある弟子がキリストの復活を見たと証言したとき、本当にキリストが復活した確率を求めてみよう．

例題1-2

この後、30人の弟子が独立にキリストの復活を見たと証言した。本当にキリストが復活した確率を求めてみよう.

$\mathrm{P}(A|\mathrm{B}) = 1.0, \mathrm{P}(A|\neg\mathrm{B}) = 0.5, \mathrm{P}(B) = 0.000002$とする.

例題1-3

実際は11人の弟子がキリストの復活を見たと証言した。本当にキリストが復活した確率を求めてみよう.

$\mathrm{P}(A|\mathrm{B}) = 1.0, \mathrm{P}(A|\neg\mathrm{B}) = 0.5, \mathrm{P}(B) = 0.000001$とする.

9. ビリーフ(信念)

つぎの二つの賭けを考えよう.

1. もしキリストが復活していれば1 万円もらえる.

2. 赤玉n 個，白玉$100-n$ 個が入っている合計100 個の玉が入っている壺の中から一つ玉を抜き出し，それが赤玉なら1 万円もらえる.

どちらの賭けを選ぶかといわれれば，2 番目の賭けで赤玉が100 個ならば，誰もが迷わず2 番目の賭けを選ぶだろうし，逆に$n = 0$ ならば，1 番目の賭けを選ぶだろう. この二つの賭けがちょうど同等になるようにn を設定することができれば，$\frac{n}{100}$があなたの「キリストが復活した」ビリーフになる. このように，ベイズ統計における確率の解釈「ビリーフ」は頻度主義の確率で扱える対象を拡張でき，個人的な信念やそれに基づく意思決定をも合理的に扱えるツールとなる.

例1 では，もともとのキリストが復活する確率$P(B)$ が，弟子の報告により$P(B \mid A)$ にビリーフが更新されていることがわかる. すなわち，弟子の証言によって事前のビリーフが事後のビリーフに更新されたのである. このとき，ベイズ統計では，

弟子の証言を「エビデンス」（evidence）と呼び，

事前のビリーフを「**事前確率**」（prior probability），事後のビリーフを「**事後確率**」（posterior probability）と呼ぶ

例題2

被害者Xはある日狙撃された。この事象をEとしよう。

命中率8割のスナイパーAと2割のスナイパーBのどちらかが犯人であることが分かっている。今、どちらが犯人かは全くわからない。

それぞれが犯人である確率を求めよ。

例題つづき

そのあとさらに2発Xに銃弾が打たれたが2発とも外れた。この事象をEとしてそれぞれが犯人である確率を求めよ。

例題つづき

新たな容疑者としてスナイパーCが浮上してきた。Cの命中率は4割である。A,B,Cの誰が犯人かわからない。最初に命中、そのあと2回外れたデータより、それぞれが犯人である確率を求めよ。

尤度

スナイパーA,B,Cのデータパターン E=(命中、外れ、外れ)が出る確率 $P(E|A)$, $P(E|B)$, $P(E|C)$ を求めた。

これらを「尤度」と呼ぶ。事前確率を考えず、尤度だけを考えるフィッシャーたちの学派を尤度派と呼ぶ。

例題3 (3 囚人問題)

ある監獄にアラン，バーナード，チャールズという3 人の囚人がいて，それぞれ独房に入れられている．3 人は近く処刑される予定になっていたが，恩赦が出て3 人のうち1人だけ釈放されることになったという．誰が恩赦になるかは明かされておらず，それぞれの囚人が「私は釈放されるのか?」と聞いても看守は答えない．囚人アランは一計を案じ，看守に向かって「私以外の2 人のうち少なくとも1 人は死刑になるはずだ．その者の名前が知りたい．私のことじゃないんだから教えてくれてもよいだろう?」と頼んだ.すると看守は「バーナードは死刑になる」と教えてくれた．それを聞いたアランは「これで釈放される確率が1/3 から1/2に上がった」とひそかに喜んだ．果たしてアランが喜んだのは正しいのか？

事前分布を変えてみよう

アランのそれぞれの事前確率は

$P(A)=\frac{3}{5}$, $P(B)=\frac{1}{5}$, $P(C)=\frac{1}{5}$

であった。この時、$P(A|E)$を求めよ。

10. 確率変数

定義5

頻度論

これから試行する実験の結果、実験結果として取り得る値

主観確率

確率法則に従う不確かな変数すべて。

11. 同時確率分布

定義6

いま，m個の確率変数をもつ確率分布$p(x_1, x_2, \cdots, x_m)$ を変数$x_1, x_2, \cdots, x_m$ の**同時確率分布**（joint probability distribution）と呼ぶ.

12. 周辺確率分布

定義7

x_i のみに興味がある場合，同時確率分布から x_i の確率分布は，離散型の場合，

$p(x_i) = \sum_{x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_m} p(x_1, x_2, \cdots, x_m)$

12. 周辺確率分布

定義7

連続型の場合，

$p(x_i) = \int p(x_1, x_2, \cdots, x_m)\, dx_1, \cdots, dx_{i-1}, dx_{i+1}, \cdots, dx_m$

で求められ，$p(xi)$ を離散型の場合，**周辺確率分布**（marginal probability distribution），連続型の場合，**周辺密度関数**（marginal probability density function）と呼ぶ．

後で詳しく学びますが、イメージを持つために先出ししておきましょう！！

同時確率分布による
確率推論

例：性別、髪の長さ、背の高さの同時確率分布

データより性別、髪の長さ、背の高さの同時確率分布が以下であることが分かっているとする。

P(男、髪短い、背高い）=0.2
P(男、髪長い、背高い）=0.125
P(男、髪長い、背低い）=0.05
P(男、髪短い、背低い）=0.125
P(女、髪短い、背高い）=0.05
P(女、髪長い、背高い）=0.125
P(女、髪長い、背低い）=0.2
P(女、髪短い、背低い）=0.125

P(男）=0.5, P(女)=0.5

同時確率分布からの確率推論

「その人は髪が短い」ことがわかった

P(男、髪短い、背高い）=0.2

P(男、髪短い、背低い）=0.125
P(女、髪短い、背高い）=0.05

P(女、髪短い、背低い）=0.125

男の確率=0.325/0.5=0.65

同時確率分布からの確率推論

さらに「その人は背が高い」ことがわかった

P(男、髪短い、背高い）=0.2

P(女、髪短い、背高い）=0.05

男の確率=0.2/0.25=0.8

確率推論の数学的定式化

データx_dが得られたときのx_iの確率は

$$p(\mathrm{x}_i|\mathrm{x}_d) = \sum_{j\neq i} p(\mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_N|\mathrm{x}_d)$$

世界中のすべての変数の同時確率分布を知ればなんでも推論できる！！

問題

変数が増えると状態数が指数的に増える

すべて2値しかとらなくても　10変数で1024パターンの同時確率を推定しないといけない。

100変数で

1267650000000000000000000000000000

パターンの同時確率を推定しなければならない。

二つの問題

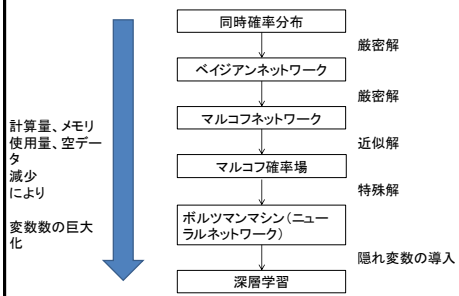
計算量が指数的に爆発する。

データ数よりパターン数のほうが多くなってしまうと、各パターンを推定するためのデータが0になるものが大量発生。

（大量のデータがあっても空データだらけになる）

ビッグデータ問題の課題は、スパースデータ（空データの増加）と計算量(メモリ、計算速度）

解決のための数理モデル



13. 確率分布とパラメータ

定義8（パラメータ空間と確率分布）

k 次元パラメータ集合を$\boldsymbol{\Theta} = \{\theta_1, \theta_2, \cdots, \theta_k\}$ と書くとき，確率分布は以下のような関数で示される．

$$f(x|\boldsymbol{\Theta})$$

すなわち，確率分布$f(x|\boldsymbol{\Theta})$ の形状はパラメータ$\boldsymbol{\Theta}$ のみによって決定され，パラメータΘ のみが確率分布$f(x|\boldsymbol{\Theta})$ を決定する情報である．

13. 確率分布とパラメータ

定義**8** (パラメータ空間と確率分布)

k 次元パラメータ集合を$\boldsymbol{\Theta} = \{\theta_1, \theta_2, \cdots, \theta_k\}$と書くとき，確率分布は以下のような関数で示される.

$$f(x|\boldsymbol{\Theta})$$

すなわち，確率分布$f(x|\boldsymbol{\Theta})$ の形状はパラメータ$\boldsymbol{\Theta}$ のみによって決定され，パラメータΘ のみが確率分布$f(x|\boldsymbol{\Theta})$ を決定する情報である.

ベイズ統計

1. 頻度論の統計学では　パラメータは確率変数でない
2. ベイズ統計学では　パラメータも確率変数

14. 確率分布とパラメータ

例 コインをn 回投げたとき，表が出る回数を確率変数x とした確率分布は以下の二項分布に従う.

$$f(x \mid \theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

ここで，θ は，コインの表が出る確率のパラメータを示す.

15. 尤度原理(フィッシャー)

定義**9** (尤度) $X = (X_1, \cdots, X_i, \cdots, X_n)$ が確率分布$f(X_i|\theta)$ に従うn個の確率変数とする.

n 個の確率変数に対応したデータ $\boldsymbol{x} = (x_1, \cdots, x_n)$ が得られたとき，

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

を尤度関数（likelihood function）と定義する（Fisher, 1925）.

尤度の例

例　コインをn 回投げたとき，表が出た回数がx 回であったときのコインの表が出るパラメータθの尤度は

$$L(\theta|n,x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

もしくは，

$$L(\theta|n,x) \propto \theta^x (1-\theta)^{n-x}$$

でもよい.

尤度は，データパターンが観測される確率に比例するパラメータθの関数である.

尤度は確率の定義を満たす保証がないために確率とは呼べないが，これを厳密に確率分布として扱うアプローチが後述するベイズアプローチである.

最尤推定法

尤度を最大にするパラメータθ を求めることは，データを生じさせる確率を最大にするパラメータθ を求めることになり，その方法を**最尤推定法**（maxmimum likelihood estimation，**MLE**）と呼ぶ.

最尤推定値

定義10 (最尤推定量)
データx を所与として，以下の尤度最大となるパラメータを求めるとき，

$$L(\theta|x) = \max\{L(\theta|x): \boldsymbol{\theta} \in C\}$$

$\hat{\theta}$を最尤推定量（maximum likelihood estimator）と呼ぶ（Fisher 1925）.
ただし，C はコンパクト集合を示す.

対数尤度とスコア関数

$$l = \ln L(\theta|x)$$

実際には　対数尤度を最大化する
以下のθについてlを偏微分したスコア関数=0となるθを求める.

$$\frac{\partial}{\partial\theta} l = \frac{\partial}{\partial\theta} \ln L(\theta|x) = \frac{1}{L(\theta|x)} \frac{\partial L(\theta|x)}{\partial\theta}$$

例題4　スコア関数の期待値

$$\mathrm{E}\left(\frac{\partial}{\partial\theta} l\right)$$

を求めよ。

例題4　　回答

$$\mathrm{E}\left(\frac{\partial}{\partial\theta} l\right) = \mathrm{E}\left(\frac{1}{L(\theta|x)} \frac{\partial L(\theta|x)}{\partial\theta}\right)$$

$$= \int_x \frac{1}{L(\theta|x)} \frac{\partial L(\theta|x)}{\partial\theta} L(\theta|x)\partial x = \frac{\partial}{\partial\theta} \int_x L(\theta|x)\partial x$$

ここで　$\int_x L(\theta|x)\partial x = 1$

より

$\frac{\partial}{\partial\theta} \int_x L(\theta|x)\partial x = 0$　　■

例題5　スコア関数の分散を求めよ

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta} l\right)$$

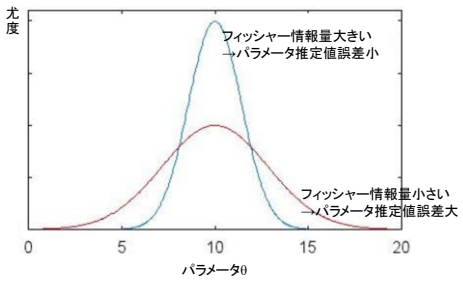
例題5　　回答

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta} l\right) = \mathrm{E}\left(\frac{\partial}{\partial\theta} l - \mathrm{E}\left(\frac{\partial}{\partial\theta} l\right)\right)^2 = \mathrm{E}\left(\frac{\partial}{\partial\theta} l - 0\right)^2$$

$$= \mathrm{E}\left(\frac{\partial}{\partial\theta} l\right)^2 = \mathrm{E}\left(\frac{1}{L(\theta|x)} \frac{\partial L(\theta|x)}{\partial\theta}\right)^2$$

これをフィッシャー情報量と呼ぶ.

フィッシャー情報量は推定値の信頼性



例題6

例6 (二項分布の最尤推定)

コインを投げてn 回中x 回表が出たときの確率θの最尤推定値を求めよ.

例題6　解答

$$L(\theta|n,x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}\propto\theta^x(1-\theta)^{n-x}$$

$$l=x\log\theta+(n-x)\log(1-\theta)$$

$$\frac{\partial l}{\partial\theta}=\frac{x}{\theta}-\frac{n-x}{1-\theta}=\frac{n-x\theta-(n\theta-x\theta)}{\theta(1-\theta)}$$

$$=\frac{x-n\theta}{\theta(1-\theta)}$$

$\theta\neq0,1$より

$\frac{\partial l}{\partial\theta}=0$となるのは　$\hat{\theta}=\frac{x}{n}$

例題7 (正規分布)

$$f(x_i|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}$$

について，データ$(x_1,\cdots,x_n)$を得たときの平均値パラメータμ，および分散パラメータσ^2 の最尤推定値を求めよ.

例題7　回答

データ$(x_1,x_2,\cdots,x_n)$を得たときの尤度は

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$l=n\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$\frac{\partial l}{\partial\mu}=0,\frac{\partial l}{\partial\sigma}=0$のとき，$l$は最大となるので

$$\frac{\partial l}{\partial\mu}=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}=\frac{\sum_{i=1}^{n}x_i-n\mu}{\sigma^2}=0 \quad\Rightarrow\quad \hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}$$

$$\frac{\partial l}{\partial\sigma}=-\frac{n}{\sigma}+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3}=0 \quad\Rightarrow\quad -n+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}=0$$

$$\widehat{\sigma^2}=\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n}$$

例題8

母集団の確率分布がポアソン分布

$$p(x)=e^{-\lambda}\frac{\lambda^x}{x!}\quad(\lambda>0,\quad x=0,1,\cdots)$$

についてn 回の観測を行ったところ

データ $\{x_1,x_2,\cdots,x_n\}$

を得た。λを最尤推定せよ。

回答

対数尤度は$l = log\left[\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\right]$

$= log\left[e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}\right]$

$= -n\lambda + (\sum_{i=1}^{n} x_i)\log\lambda - \log(\prod_{i=1}^{n}(x_i!))$

$\frac{dl}{d\lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$

より

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

強一致性

定義11 (強一致性)

推定値$\hat{\theta}$が真のパラメータθ^*に概収束するとき，$\hat{\theta}$は強一致推定値（strongly consistent estimator）であるという.

$P(\lim_{n\to\infty} \hat{\theta} = \theta^*) = 1.0$

つまり，データ数が大きくなると推定値が必ず真の値に近づいていくとき，その推定量を強一致推定値と呼ぶ.

最尤推定値の一致性

定理9 (最尤推定値の一致性)

最尤推定値$\hat{\theta}$は真のパラメータθ^*の強一致推定値である（Wald,1949）.

最尤推定値の漸近正規性

定義12

θ^*の推定値 $\hat{\theta}$が**漸近正規推定量**（asymptotically normal estimator）であるとは，$\sqrt{n}(\hat{\theta} - \theta^*)$ の分布が正規分布に分布収束することをいう. すなわち，任意の$\theta^* \in \Theta^*$ と任意の実数に対して

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\hat{\theta} - \theta^*)}{\sigma(\theta^*)} \le x\right) = \Phi(x)$$

このことを，$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{as} N(0, \sigma^2(\theta^*))$と書く.

$\sigma^2(\theta^*)$ を漸近分散（asymptotic variance）という.

最尤推定値の漸近正規性

定理10

確率密度関数が正則条件（regular condition）の下で，微分可能のとき，

最尤推定量は漸近分散$I(\theta^*)^{-1}$をもつ漸近正規推定量である.

$$I(\theta^*) = E_\theta\left[\left(\frac{\partial}{\partial\theta} lnL(\theta|\mathbf{x})\right)^2\right]$$

をフィッシャー（Fischer）の情報量と呼ぶ.

より複雑なモデル

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2}^2 + \varepsilon_i,$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

入力　(x_{i1}, x_{i2}, y_i)(i=1, …n)

データファイル　の読み込み

パラメータw_0, w_1, w_2, σ^2を最尤推定せよ。

尤度は

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}\right)$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}\right)$$

対数尤度は

$$l = n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}$$

非線形モデルは解析的に解けない

数値計算法

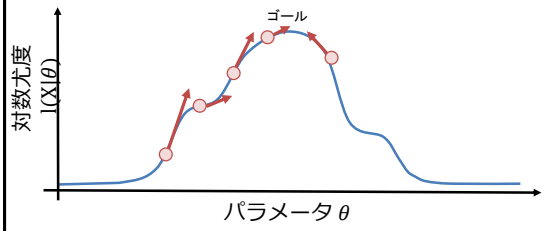
パラメータ推定値が解析的に求まらない場合には数値計算によって求める

代表的な手法

- 勾配上昇法
- ニュートン・ラフソン法

勾配上昇法（最急上昇法）

適当な初期値から、勾配方向にパラメータを更新することで極値（勾配0）を求める
傾きが正ならパラメータを正の方向へ、傾きが負ならば負の方向へ



最小値を求める問題の場合は勾配降下法（最急降下法）と呼ばれる

勾配上昇法のアルゴリズム

パラメータθ , 対数尤度関数 $l(X|\theta)$

アルゴリズム

1. パラメータθに適当な初期値を付与
2. 対数尤度関数の偏微分方向に微分値の η 倍更新
$$\theta_{n+1} = \theta_n + \eta \frac{\partial l(X|\theta)}{\partial \theta} : \forall n$$
3. 以下の収束条件を満たす（全てのパラメータ更新量が十分小さくなる = ϵ 以下になる）まで2.を反復
$$\eta \frac{\partial l(X|\theta)}{\partial \theta} \leq \epsilon : \forall n$$

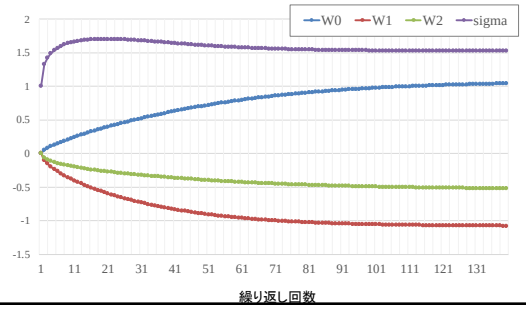
一階偏微分

$$l = n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial w_0} = \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)}{\sigma^2}$$
$$\frac{\partial l}{\partial w_1} = \sum_{i=1}^{n} \frac{x_{i1}(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)}{\sigma^2}$$
$$\frac{\partial l}{\partial w_2} = \sum_{i=1}^{n} \frac{x_{i2}^2(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)}{\sigma^2}$$
$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{\sigma^3}$$

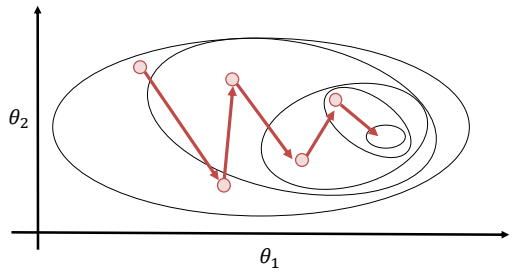
推定例

$\eta = 0.0001, \epsilon = 0.001$ サンプルサイズ1000
真値：$w_0 = 1.0, w_1 = -1.0, w_2 = -0.5, \sigma = 1.5,$



勾配上昇法の問題

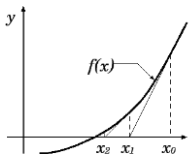
勾配情報のみで更新方向を決定するため効率が悪い
勾配以外の情報を使用 ⇨ ニュートン・ラフソン法



ニュートンラフソン法

方程式$f(x) = 0$を解く手法。
最大値問題の場合は、偏微分$f'(x) = 0$となるxを求める方程式を解けばよい。

ニュートン　ラフソン法



$f(x) = 0$を解く。
図のように適当な初期値 x_0において$f(x)$ に接線を引けば、接線の方程式は

$$y - f(x_0) = f'(x_0)(x - x_0)$$

X軸との交点は

$$x_1 = x_0 - f(x_0)/f'(x_0)$$

次にx_1での$f(x)$への接点とx軸との交点を求める。これを繰り返す。

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

ニュートン法はテーラー近似

非線形関数の方程式$f(x_n) = 0$ を解きたい。
$f(x_n)$をx_{n-1}のまわりでテーラー展開すると

$$f(x_n) = f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + O((x_n - x_{n-1})^2)$$

$f(x_n) = 0$ より

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0$$

これより、

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

例 （1次元の場合）

$f(x) = x^2 - 2 = 0$ を解け（初期値 1.0とする）
ニュートンラフソン法を用いて 横軸に繰り返し数、縦軸にxの推定値を書け.

例

$f(x) = x^2 - 2 = 0$ を解け（初期値　1.0とする）

$$f'(x) = 2x$$

より

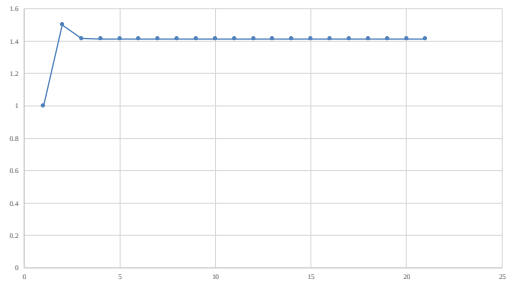
$$x_{n+1} = x_n - \frac{f(x)}{f'(x)}$$

$$x_{n+1} = x_n - \frac{x^2 - 2}{2x_n} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right)$$

初期値　1.0とする

数値例

推定値の遷移



繰り返し回数

多次元の場合の最尤法でのニュートン・ラフソン法

勾配（1階微分）に加えて、曲率（2階微分）を利用

パラメータ集合 $\boldsymbol{\theta} = \{\theta_1 \cdots \theta_N\}$, 対数尤度関数 $l(X|\boldsymbol{\theta})$とするとき、対数尤度関数の勾配行列 $g(\boldsymbol{\theta})$と2階微分行列:ヘッセ行列 $H(\boldsymbol{\theta})$をそれぞれ以下で表す

$$g(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l(X|\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial l(X|\boldsymbol{\theta})}{\partial \theta_n} \end{bmatrix}, \quad H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l(X|\boldsymbol{\theta})^2}{\partial^2 \theta_1} & \cdots & \frac{\partial l(X|\boldsymbol{\theta})^2}{\partial \theta_1 \partial \theta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial l(X|\boldsymbol{\theta})^2}{\partial \theta_n \partial \theta_1} & \cdots & \frac{\partial l(X|\boldsymbol{\theta})^2}{\partial^2 \theta_n} \end{bmatrix}$$

ニュートン・ラフソン法のアルゴリズム

パラメータ集合 $\boldsymbol{\theta} = \{\theta_1 \cdots \theta_N\}$, 対数尤度関数 $l(X|\boldsymbol{\theta})$

<u>アルゴリズム</u>

1. 各パラメータ$\{\theta_1 \cdots \theta_N\}$に適当な初期値を付与
2. 対数尤度関数の偏微分方向に微分値のη倍更新

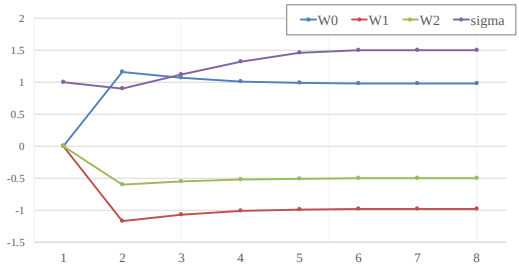
$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta\, H(\boldsymbol{\theta})^{-1} g(\boldsymbol{\theta})$$

3. 収束条件を満たす（全てのパラメータ更新量が十分小さくなる$= \epsilon$ 以下になる）まで2.を反復

推定例

$\eta = 1.0, \epsilon = 0.001$ サンプルサイズ1000

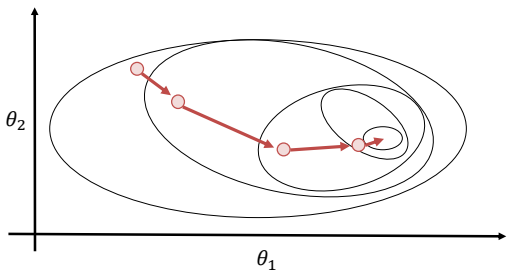
真値:$w_0 = 1.0, w_1 = -1.0, w_2 = -0.5, \sigma = 1.5,$



繰り返し回数

ニュートン・ラフソン法のイメージ

曲率（勾配の変動）が大きい場所では更新幅を小さくし、曲率が小さい場所では大きく更新



数値計算法の注意点

初期値依存
- 初期値によって推定値が発散することがある
- 発散したと判断される場合にはランダムに初期値を振り直して再スタートするなどの工夫が必要

学習率 η の設定
- 小さすぎると1ステップあたりの更新幅が小さくなり、収束に時間がかかる
- 大きすぎると極値を飛び越えてしまい収束しにくくなる。また、発散の可能性も高まる
- 適切な値を経験的に設定する必要がある

収束判定閾値 ϵ の設定
- 十分に小さく取るべき（例えば、0.001）だが、小さくするほど収束に時間がかかる