

a t m a r k I T

# AI・機械学習のための 数学超入門 ～第2部 偏微分～

羽山博[著]

01. [AI・機械学習の数学] 微分法の基本を身につけて「変化」を見極めよう

02. [AI・機械学習の数学] 微分法を応用して、回帰分析の基本を理解する

03. [AI・機械学習の数学] 偏微分の基本（意味と計算方法）を理解する

04. [AI・機械学習の数学] 偏微分を応用して、重回帰分析の基本を理解する

05. [AI・機械学習の数学] 合成関数の微分（連鎖律）とニューラルネットワーク初歩の初歩

# [AI・機械学習の数学]微分法の基本を身につけて「変化」を見極めよう

式の値が最小になるときの  $x$  の値を求めたり、値がどのように変化していくかを見たりするためには微分が活用できる（回帰分析・重回帰分析につながる基礎知識）。今回は平均変化率から始め、微分の基本を一步一步じっくりと追いかける。

羽山博, 著 (2020 年 05 月 14 日)

機械学習では、値を少しずつ変えていき適切な値に近づけていくという計算をよく行います。最初は適切な値とは大きく離れていますが、徐々に近づいていくでしょうから、変化率を見れば適切な値に近づいているかどうか分かります。変化率が無視できるほど小さくなれば、そのときの値が最適な値と見なせるはずです。

ところで、そもそも**変化率**というのはどういうものでしょうか。今回は変化率の意味を確認し、さらに微分の考え方と計算方法を学びます。

## 目標【その 1】：平均変化率を理解する

**平均変化率**とは、一定の間隔の中でどれだけ値が変わったか、ということです。

$$\text{平均変化率} = \frac{\text{増分値}}{\text{間隔}}$$

で表されます。平均変化率が**正**なら増えていることを表し、**0** なら変化していないことを表します。**負**なら減っているということが分かります。また、ある時点での平均変化率よりも次の時点の平均変化率が大きければ、ますます増加する傾向にある、ということも分かります。

## 解説：平均変化率の意味と利用例を知る

目標のところでは簡単な式を記しただけですが、平均変化率はさまざまな場面で使われます。例えば、車が **3 時間** で **240km** 進んだとか、株価が **6 か月** で **300 円** 上がったとか、日常の例を挙げればキリがありません。最近（2020 年 5 月現在）の深刻な話題としては新型コロナウイルスの感染者が 1 日でどれだけ増えたか、といったことも平均変化率と考えられます。平均変化率が大きくなると拡大の傾向にあり、小さくなると収束に向かっているということも分かります。

平均変化率は単なる割り算で求められます。実際のところ、車の例なら途中でサービスエリアに寄ったりすることもあるでしょうし、株価は秒刻みで上下するので、「間隔」の取り方によって変化率は変わりますが、「ならして」みるとこうなるという意味で「**平均変化率**」と呼びます **\*1**。



\*1 ちなみに、この場合の「ならして」は「均して」と書きます。

$240 \div 3$  ... 3時間で240km進んだ場合の平均変化率(= 平均速度)  
 $300 \div 6$  ... 6か月で300円株価が上がった場合の平均変化率(= 上昇率)

繰り返しになりますが、平均変化率は、変化した値を、時間や期間などの間隔で割れば求められることが分かります (km/h などの単位は省略してあります)。以下の図で見てみましょう。100km の地点から 340km の地点まで移動するのに、1 時から 4 時まで時間がかかっています。

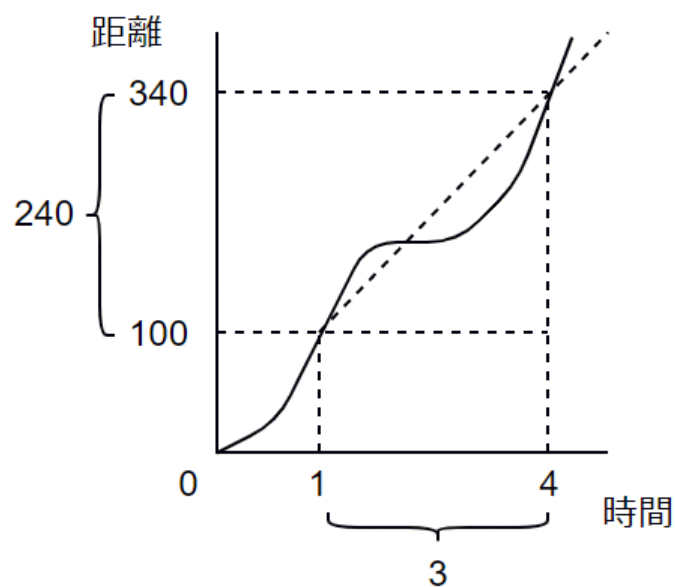


図1 平均変化率を求める

この場合、時間は  $4 - 1 = 3$  で、距離は  $340 - 100 = 240$  なので、平均変化率は、

$$240 \div 3 = 80$$

になるわけです。平均変化率は、図中の斜めの点線の傾きに当たることも分かりますね。当然のことながら、間隔が異なると平均変化率が変わることもあります。

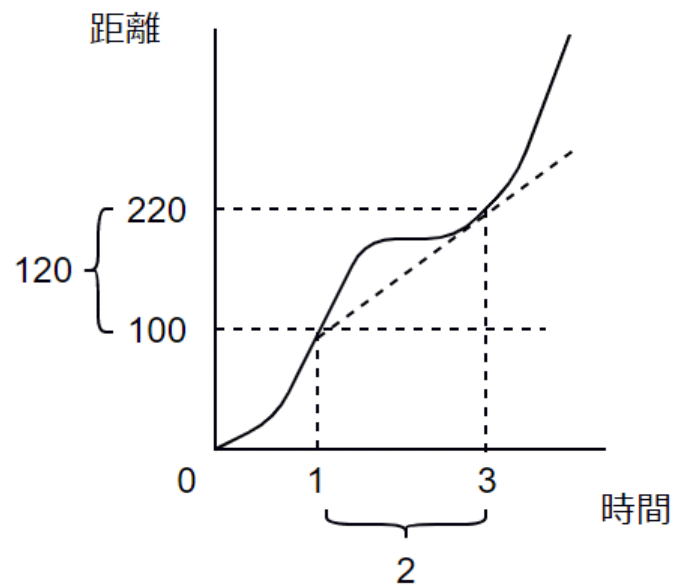


図2 間隔が異なると平均変化率も変わる

この場合、時間は  $3 - 1 = 2$  で、距離は  $220 - 100 = 120$  なので、平均変化率は、

$$120 \div 2 = 60$$

となります。平均変化率の計算は単なる引き算と割り算ですね。では、練習問題として、もう少し抽象化した例にチャレンジしてみましょう。

## 練習問題

第一部の中学数学からのおさらい編では距離の二乗和の最小値を求めるために二次式を使いました。二次式は中学・高校で慣れ親しんでいるので、ここでも簡単な二次式（二次関数）を例に平均変化率を計算してみましょう。以下のグラフは、

$$y = 3x^2 - 5x + 1 \quad (1)$$

のグラフです \*2。



\*2 ボールなどを投げ上げたときの時間  $t$  と位置  $y$  の関係も  $y = -1/2gt^2 + at$  という二次式で表されます（ $g$  は重力加速度と呼ばれ、地球上では約  $9.81$  です。 $a$  は初速です。なお、投げ上げたボールは下に落ちてくるので、下の図とは反対の上に凸なグラフになります）。

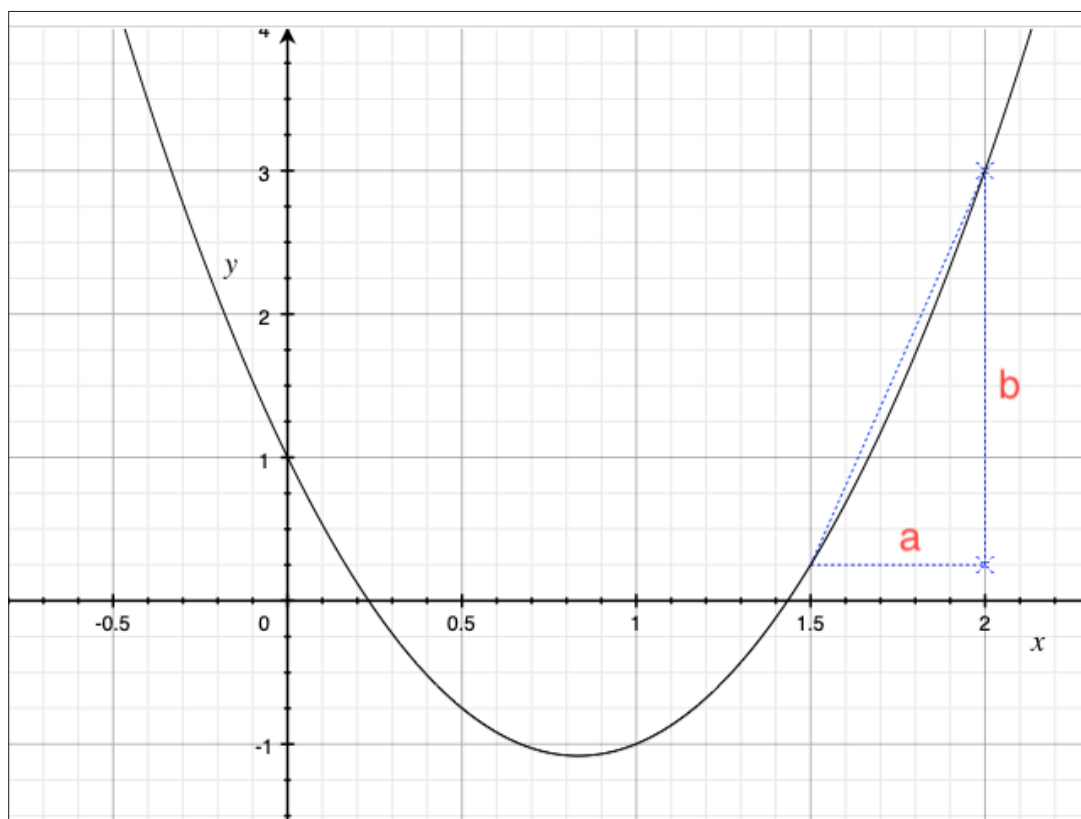


図3 二次式のグラフ

$x$  の値が  $1.5$  から  $2$  まで変わるときの  $y$  の値の変化率を求めてみましょう。つまり、図中の斜めの点線の傾きを求めます。図中の  $b$  の長さを  $a$  の長さで割ればいいですね。 $a$  の長さは、 $x$  の値の  $1.5$  と  $2$  から求められます。これはカンタン、 $2 - 1.5 = 0.5$  です。一方、 $b$  の長さを求めるためには、 $1.5$  や  $2$  に対応する  $y$  の値を求めておく必要があります。

では、 $x$  が 1.5 のとき、 $y$  の値はいくらになるでしょうか。これは、(1) 式の  $x$  に **1.5** を代入すれば求められます。計算が必要な箇所は電卓を使っちゃいましょう。

$x = 1.5$  のとき、

$$\begin{aligned} y &= 3 \cdot 1.5^2 - 5 \cdot 1.5 + 1 \\ &= 0.25 \end{aligned}$$

$1.5^2 = 2.25$  なので、 $y = 3 \cdot 2.25 - 5 \cdot 1.5 + 1$  を計算すればいい、ということが分かりますね。答えは **0.25** でした。

同様にして、

$x = 2$  のとき、

$$\begin{aligned} y &= 3 \cdot 2^2 - 5 \cdot 2 + 1 \\ &= 3 \end{aligned}$$

という値も得られます。

いったん整理しておきましょう。

$$x = 1.5 \text{ のとき、} y = 0.25$$

$$x = 2 \text{ のとき、} y = 3$$

です。これで平均変化率が求められますね。

$$\begin{aligned} \text{平均変化率} &= \frac{3 - 0.25}{2 - 1.5} \\ &= 5.5 \end{aligned}$$

はい、平均変化率が求められました。この **5.5** という値は  $x$  が **1** 増えると  $y$  は **5.5** 増えるという意味です。

## 目標【その2】：平均変化率を文字式で表してみる

$x$  の値に対して、何らかのルールで  $y$  の値が対応する、ということを、

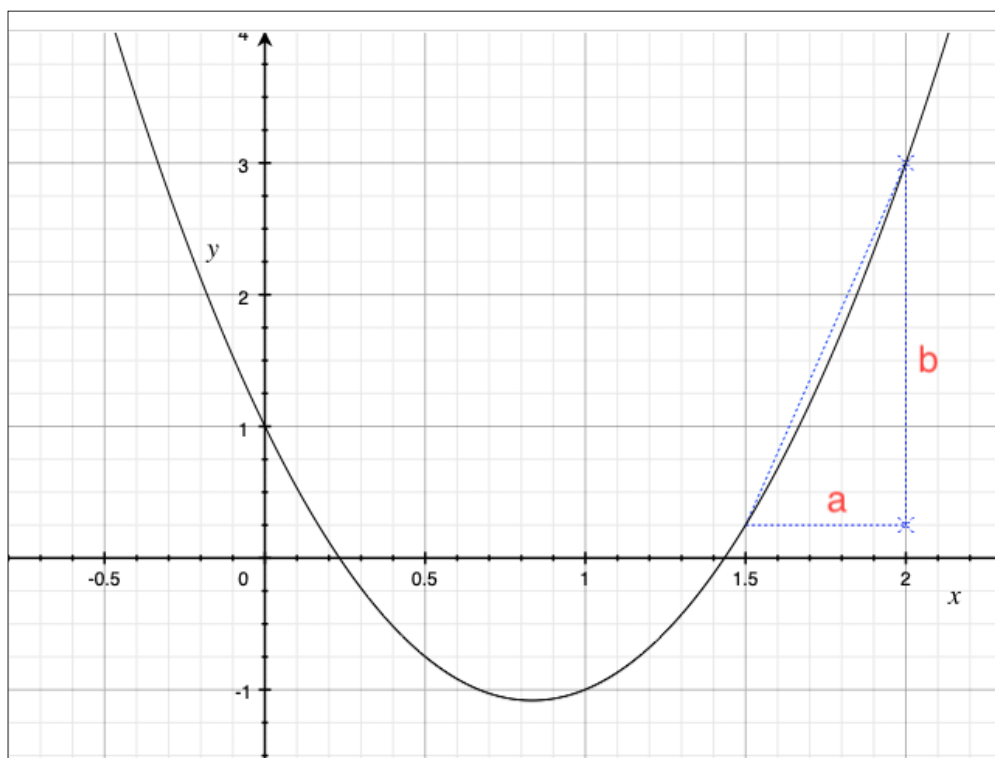
$$y = f(x)$$

と表してみることにします。このとき、 $x$  が  $h$  だけ増えたときの平均変化率は以下の式で表されます。

$$\frac{f(x+h) - f(x)}{h}$$

### 解説：平均変化率を文字式で表してみる

平均変化率を文字式で表すに当たって、ちょっと前置きのお話をしておきましょう。目標に示した「 $x$  の値に対して、何らかのルールで  $y$  の値が対応する」というのは、要するに図3（下に再掲）のようなグラフが書けるということです。



【再掲】図3 二次式のグラフ

そのルールを「 $f$ 」という文字で表し、「 $f$ 」というルールに  $x$  を与えるということを、

$$f(x)$$

と表してみましょう。すると、 $y$  の値が決まるので、

$$y = f(x)$$



となります。前掲の図 3 の例なら、

$$f(x) = 3x^2 - 5x + 1$$

です。なんだか回りくどい言い方をしているようですが、このように、あるルールに従って値が対応することやそのルールを表す式のことを「関数」と呼びます。ルール（関数）が複数あるときには **g** という文字を使うこともあります。**f(x)** や **g(x)** というように区別して表すわけです。回りくどさついでに、別の図で「関数」を表しておきましょう。

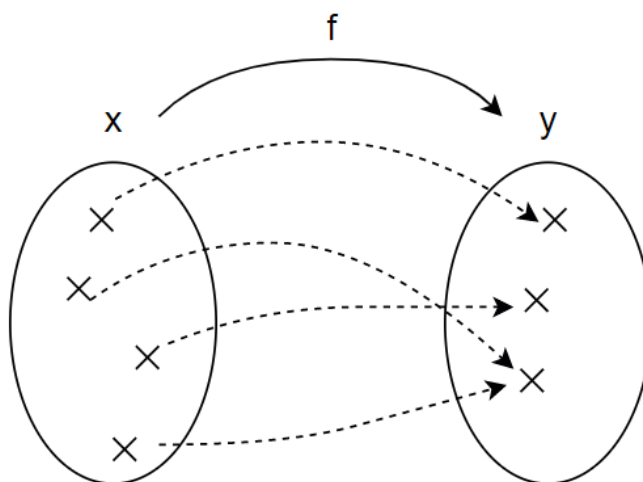
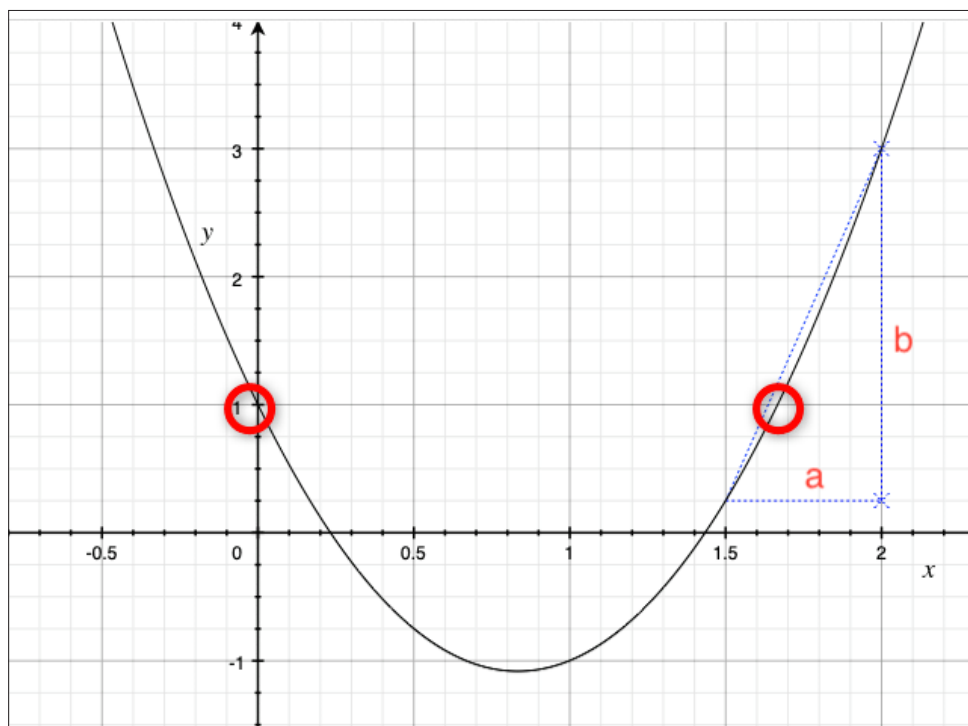


図 4 関数とは値の対応のこと

この図は **x** の値が **y** の値に対応していることを表しています。**f** というのが対応を表すルール（関数）ですね。図からも読み取れますが、異なる **x** の値に対して、**y** の値が同じになることもあります。図 3（二次式のグラフ、下に再掲）の例でいえば、**x = 0** のときにも、**x = 5/3** のときにも **y** の値が **1** になります。



【再掲】図 3 二次式のグラフ

先ほど見た二次関数  $y = 3x^2 - 5x + 1$  では、係数は **3** や **5** といった具体的な数値でしたが、これを一般的に表すために文字で表してみましょう。次のような感じですね。

$$y = ax^2 + bx + c$$

ここで注意すべき点は、**a, b, c** はいずれも定数であるということです。一方、**y, x** は変数です。**a, b, c** は何か分からないけれども決まった値であり、**y, x** はさまざまな値を取るということです。慣れれば特に意識しなくても分かりますが、どれが定数でどれが変数なのかをちょっと意識しておくことで式の意味合いや取り扱いの方法がよく分かり、見通しがよくなります。

では、「ルール **f** によって **x** の値が **y** の値に対応していますよ」ということが分かるように、**y** の代わりに **f(x)** を使って書いてみてください。

$$f(x) = ax^2 + bx + c$$

カンタンでしたね。**x** と **y** だけだと「**x** と **y** に関係がある」ということが伝わりませんが、**x** と **f(x)** なら、「**x** に何かのルールを適用したものが **f(x)** である」という関係がよく分かります。この方が、**関数** というコトバのイメージに合っていますね。こういう書き方、ちょっとカッコイイですし。

続いて、**平均変化率** です。実は、 $ax^2 + bx + c$  に限らず、どんな関数でも同じことなので、ここからはもう **f(x)** だけで話を進めます。

**x** の値が **h** だけ増えたとします。このときの平均変化率を求めてみましょう。

分母は  $x$  から  $x + h$  まで変わる ( $h$  だけ変わる)

分子は  $f(x)$  から  $f(x + h)$  まで変わる

というわけですから、平均変化率は以下の式になります。

$$\text{平均変化率} = \frac{f(x+h) - f(x)}{h}$$

この式と図 5 をじっくりと見て、式と図の対応関係を確認しておいてください。次の「微分のお話」で使うためです。

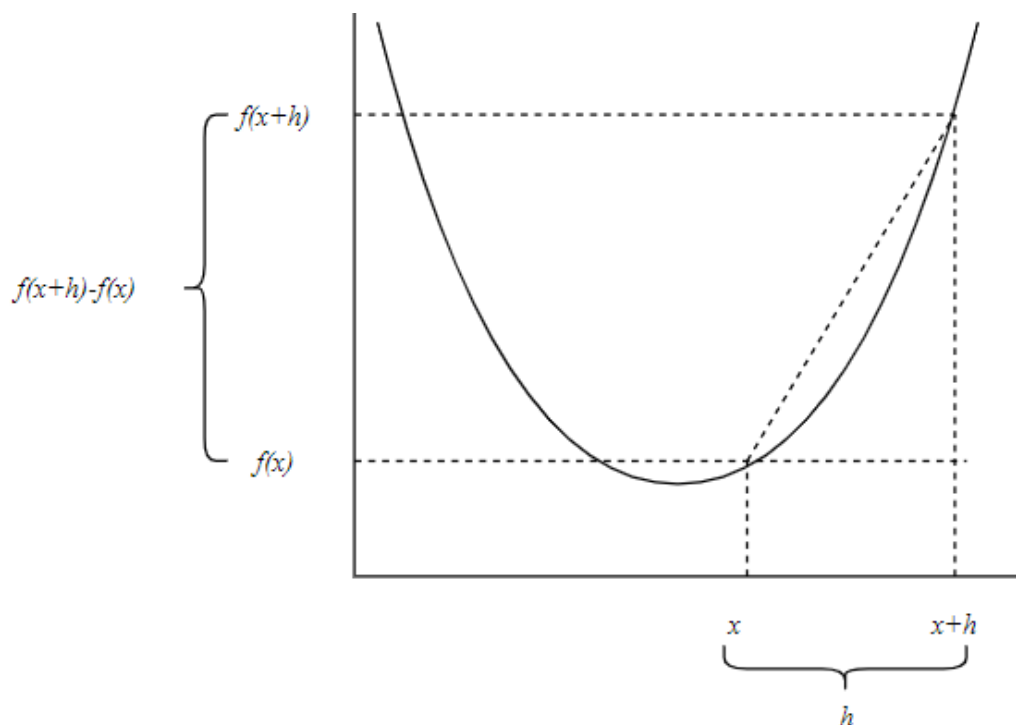


図 5 平均変化率を一般的に表す

これで、平均変化率が一般的に表せました。上でも述べましたが、どんな関数でも同じです。

### 目標【その 3】：いよいよ微分のお話（導関数を理解する）

さて、いよいよ微分のお話です。平均変化率の式で、間隔  $h$  をどんどん小さくしていくと接線の傾きが求められます。**微分**とは、その「**接線の傾き**」を表す式を求めることとも考えられます（厳密な定義ではありませんが、最初はそういうイメージで十分です）。まずは以下の式を見てください。

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

この  $f'(x)$  は、「関数  $f(x)$  の**導関数**（どうかんすう）」と呼ばれます。微分とは導関数を求めることに他なりません。つまり、この式が微分の定義です。

ここでの目標は、導関数を表す式（つまり上記の式における右辺）の意味を理解することです。

## 解説：微分の考え方と書き方について

いきなりものすごい式が出てきてハードルが急に上がったように思った人がいるかもしれませんが、右辺の式に見覚えはないでしょうか。右端の部分は平均変化率の式そのものでよね。つまり、平均変化率をどうこうしてやれば、微分ができるということです。

ここで、新しく出てきた記号の読み方をまず確認しておきましょう。新しく出てきたのは、左辺の  $f'(x)$  と右辺の  $\lim$  の2つだけです。

左辺の  $'$  は「プライム」と読みます。従って、 $f'(x)$  は「エフプライムエックス」や「エフプライムかっこエックス（かっこことじ）」のように読みます（※「エフダッシュ」ではありません）。

右辺の  $\lim$  は「リミット」と読みます。ある変数の値を何らかの値に限りなく近づけていく、といった意味なのですが、これが理解できれば微分も理解できるので、ここから詳しく見ていくことにしましょう。

というわけで、ここから平均変化率がどう微分に結び付いてくるのかというお話をします。最終的には、超カンタンな計算方法にたどり着くのですが、考え方が大事なのでゆっくりと読み進めてください。

平均変化率は図6の左に示した「斜めの点線の傾き」のことでしたね。このような状態から、間隔  $h$  をどんどん小さくしていくと、いったいどうなるのでしょうか。

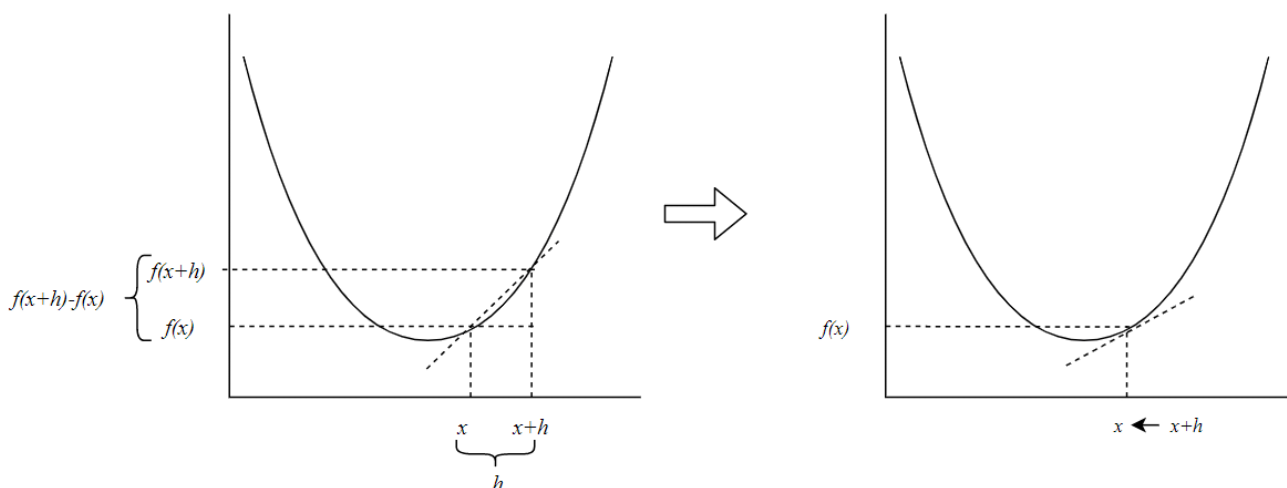


図6 接線の傾き

直感的に分かと思いますが、図6の右のように**接線の傾き**に近づいていきます（念のため、この動きを追いかけた動画も用意しておきます）。



動画 1 接線の傾き

平均変化率の「間隔」が限りなく **0** に近づいていくので、接線の傾きは「ある瞬間の変化率」とも表現できますね。

そこで、接線の傾きを求めるために、平均変化率の式で、間隔 **h** をどんどん小さくしていき、限りなく **0** に近づけてみましょう。そのときに、平均変化率がどのような値に近づくかを求めるといいですね。

ただ「変数の値を限りなく何らかの値に近づけていったとき、式の値が限りなく近づく値はいくらですよ」と言うのも書くのも長ったらしいので、**lim** という記号を使って以下のように表すことにします。なお、「限りなく近づく値」のことを「**極限值**」と呼びます。

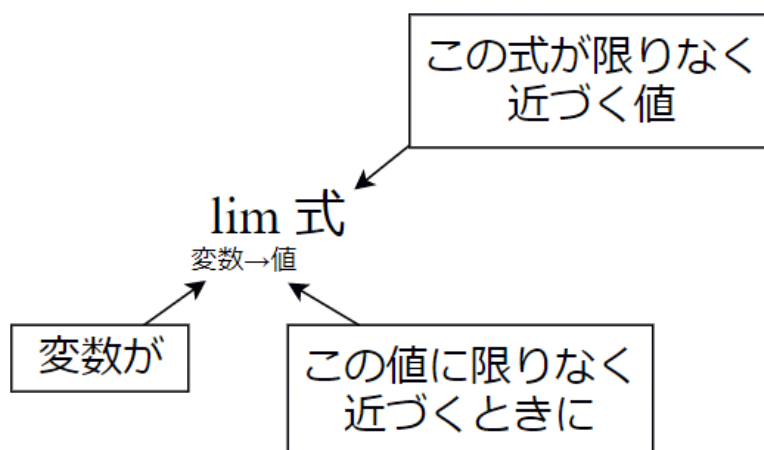


図 7 lim の書き方と意味

この書き方だと簡潔に表せますね。読み方は「リミット、変数が、値に近づくときの、式の値」といった感じになります。例えば、

$$\lim_{h \rightarrow \infty} \frac{1}{n}$$

なら、「リミット、エヌが、無限大に近づくときの、エヌ分のイチ」です（ちなみにこの式の値は **0** です）。lim の書き方も大事なので、ちょっとだけ問題をやっておきましょうか。x が限りなく **2** に近づくとき、 $x^2 + 3x - 4$  はどのような値に近づくかは、以下のように表されますね。

$$\lim_{x \rightarrow 2} (x^2 + 3x - 4)$$

ちなみに、この値は  $2^2 + 3 \cdot 2 - 4$ 、つまり、**6** に限りなく近づくので lim の式の値は **6** になります。

話を元に戻しましょう。h が限りなく **0** に近づくとき、平均変化率が**どの値**に近づくかを求めるのでしたね。それを lim を使って表すと、

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

のようになります。この式は目標のところで見た、「f(x) の導関数」を表す式そのものです。そして、この式の値が接線の傾きになっています。

繰り返しになりますが、導関数を求めることを**微分する**と言います（大事なことなので 2 回言いました）。

また、いちいち「f(x) の導関数」というのも面倒なので、こちららも記号を使って表しましょう。実は、導関数にはいろいろな書き方があり、

$$y' \qquad f'(x)$$

$$\frac{dy}{dx} \qquad \frac{d}{dx} f(x)$$

といった表し方をします。手で書いたり、キーボードから入力したりするときの手間を考えると、

$$y' \text{ や } f'(x) \text{ という書き方}$$

が簡単ですが、発展的な計算をするときには、

$$\frac{d}{dx} f(x) \text{ という書き方}$$

の方が便利になります \*3。



\*3

$$\frac{dy}{dx}$$

は分数の形になっていますね。式を変形するときに分数と同じように扱うこともできるのですが、「ディーエックス分のディーワイ」とは読まずに「ディーワイ、ディーエックス」と順に読むのが一般的です。

今のところ、簡単に書けて、関数だということが分かる  $f'(x)$  で十分なので、以下のように書くことにします。

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad \dots \text{これが微分の定義！}$$

というわけで、これが微分の定義です。右辺の意味をもう一度確認しておくと「間隔  $h$  をどんどん小さくしていき、限りなく  $0$  に近づけたときに、平均変化率がどの値に近づくか」ということになります。それが導関数の式であり、 $x$  における  $f(x)$  の接線の傾き（を表す関数）になるわけです \*4。



\*4  $x$  軸のある点  $a$  における導関数の値は、 $f'(a)$  と表されます。この値のことを「微分係数」と呼びます。高校の教科書では、まず微分係数を学んで、次に導関数について学ぶことが多いようです。ここでは、一気に導関数から見てきました。 $f'(a)$  の  $a$  は定数で、微分係数  $f'(a)$  は何らかの値（ある点における接線の傾き）であることに注意しましょう。一方、 $f'(x)$  の  $x$  は変数で、導関数  $f'(x)$  は関数であることに注意しましょう。 $x$  の値を変えると、それに対応した接線の傾きが求められます。



長々と抽象的なお話をしてきたように思われるかもしれませんが、実際のところ、「 $h$ 」を限りなく  $0$  に近づける」とか「関数  $f(x)$  の導関数」といった言葉をできるだけ簡単な記号で表すとどうなるかというお話と、その記号をどんなふうを読むか、というお話をしてきただけです \*5。



\*5 「限りなく近づく」というのは「等しい」ということではありません。ちょっと違う例で見てみましょう。例えば、

$$\lim_{h \rightarrow 0} (h + 1) = 1$$

ですが、これは「 $h$ 」を限りなく  $0$  に近づけていけば、 $h + 1$  という式が限りなく近づく値は  $1$  ですよ」ということです。 $h = 0$  のとき  $h + 1 = 1$  になるというのは意味が異なります。 $\lim$  の意味や公式については、ここでは詳しく説明しませんが、とりあえずそういうものだ、と思っておいください。なお、式によっては「限りなく近づく値」が無限に大きくなったり、無限に小さくなったりすることや（発散）、一定の値に決まらず発散もしない場合もあります（振動）。

そういうわけで、微分の定義は分かったものの、それだけでは、どういう計算をすればいいのか見当が付きませんね。実は、簡単に計算を行う方法があるのです。次は、導関数がどのように計算されるかを見ていきます。

## 目標【その 4】：微分の計算方法（導関数の求め方）

単刀直入に答えを先に示しておきます。多項式の微分は以下の式さえ覚えておけば簡単にできます。

$$f(x) = ax^n$$

のとき、

$$\begin{aligned} f'(x) &= (ax^n)' \\ &= anx^{n-1} \end{aligned}$$

### 解説：微分の計算方法（導関数の求め方）

微分の方法（導関数の求め方）は次のような単純な手順です。「微分」は、難しい数学の代表のように思われていますが、多項式の微分は極めて簡単な計算です。

手順：

1. 指数を係数に掛ける
2. 指数を 1 減らす
3. 定数項は消す



たったのこれだけです。式の項ごとにこのルールを適用します。以下の図でも確認しておきましょう（動画も用意してあります）。



動画 2 微分の計算方法と計算例

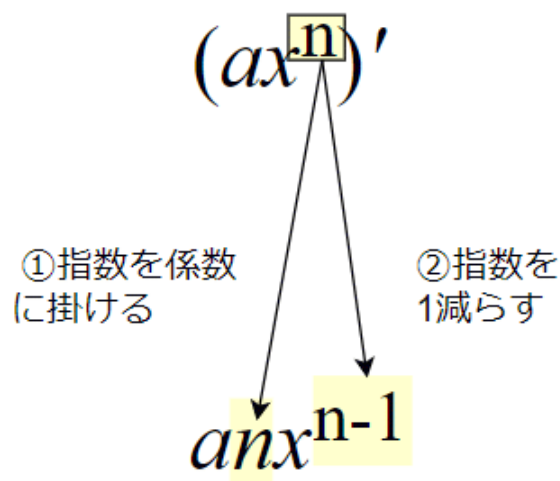


図 8 微分の計算方法

具体的な例でも見ておきましょう。以下の図は、 $2x^3$  を微分する例と、 $x^3 - 5x + 6$  を微分する例を、上の手順通りにやってみたものです。

(例)

$$(2x^3)'$$

① 指数を係数に掛ける      ② 指数を1減らす

$$2 \cdot 3x^{3-1} = 6x^2$$

(例)

$$(x^3 - 5x + 6)'$$

① 指数を係数に掛ける      ③ 指数を係数に掛ける      ⑤ 定数項は消す

② 指数を1減らす      ④ 指数を1減らす

$x$ とは $x^1$ のこと

定数項は $x^0$ の項なので、指数の0を係数に掛けると、この項が0になる(「定数項は消す」というルールは「指数を係数に掛ける」というルールを適用しただけのこと)

$$3x^{3-1} - 5 \cdot 1x^{1-1} = 3x^2 - 5$$

$x^0 = 1$

図9 微分の計算例

$x$  は  $x^1$  なので、 $x$  の指数は  $1$  です。また、 $x^0 = 1$  なので、定数項は  $x^0$  の項であると考えられます。手順3の「定数項は消す」というルールは、実際には、手順1の「指数を係数に掛ける」というルールを適用して、指数である  $0$  を係数に掛けただけのことです。

いかがでしょう。そういえば、高校のときにこういうやり方を覚えたなあ、と思い出した方も多いのではないのでしょうか。思い出したついでに、もうひと押しです。練習問題にも取り組み、計算方法を確実に身につけておきましょう。

### 練習問題

以下の関数を微分してみましょう。

- (1)  $f(x) = 7x^3$  を  $x$  で微分する
- (2)  $f(x) = 10$  を  $x$  で微分する
- (3)  $f(x) = 2x^4 + 3x^2 + 5$  を  $x$  で微分する
- (4)  $f(t) = t^2 - 4t + 4$  を  $t$  で微分する

## 解答

(1)  $f'(x) = 21x^2$

(2)  $f'(x) = 0$

(3)  $f'(x) = 8x^3 + 6x$

(4)  $f'(t) = 2t - 4$

(1) は、指数が **3** なのでそれを係数に掛けて、 $7 \times 3 = 21$  が導関数の係数になりますね。指数の方は **1** つ減らすと **2** です。よって、 $21x^2$  となります。

(2) は、ちょっとひっかけっぽいですが、定数項しかないので、**0** になります。 $f(x)$  の値は常に **10** なので、全く変化しませんね。だから変化率は **0** です。

(3) は、手順通り各項を計算していけば求められます。

(4) は変数として  $t$  を使っただけです。やはり手順は同じです。

## 解説（続き）：微分の定義と計算方法について

微分の意味や定義も分かった、計算方法も分かった、めでたしめでたし……と言いたところですが、「いや、ちょっと待ってくれ」と思った人はいないでしょうか。微分の定義は以下のようなものでした。

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

そして、計算方法は以下のようなものでした。

$$f(x) = ax^n$$

のとき、

$$\begin{aligned} f'(x) &= (ax^n)' \\ &= anx^{(n-1)} \end{aligned}$$

これでは、定義と計算方法がどうつながっているのかが分かりませんね。意味もやり方も分かったから、なんだか分からない面倒な途中の話はいいや、と投げ出さずに、上の定義と下の計算方法がつながっている（＝導き出せる）ことを確認してみましょう。ただし、厳密に証明するとかなり長くなるので、簡単な例で答えが合っていることだけを確かめるにとどめます。

では、例として、 $f(x) = ax^2$  を取りあげます。この導関数を微分の定義に当てはめて計算してみましょう。簡単な例とはいえ、式の変形が少し細くなるので動画も用意しておきました。



動画 3 導関数を求める（二次関数の例）

まず、微分の定義に、関数の式を代入してみましょう。

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \dots [A] \\ &= \lim_{h \rightarrow 0} \frac{a(x+h)^2 - ax^2}{h} \dots [B] \end{aligned}$$

[A] …これが導関数の定義

[B] …  $f(x+h) = a(x+h)^2$  と  $f(x) = ax^2$  を代入した

次に、分子の式を展開して、整理してみます。

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{a(x^2 + 2hx + h^2) - ax^2}{h} \dots [A] \\ &= \lim_{h \rightarrow 0} \frac{ax^2 + 2ahx + ah^2 - ax^2}{h} \dots [B] \\ &= \lim_{h \rightarrow 0} \frac{2ahx + ah^2}{h} \dots [C] \\ &= \lim_{h \rightarrow 0} (2ax + ah) \dots [D] \end{aligned}$$

[A] …二乗の展開の公式を使って式を展開した

[B] …分配法則を使ってカッコを外した

[C] …  $ax^2$  と  $-ax^2$  の計算をした (0 になるのでこの項が消える)

[D] …  $h$  で約分した

$ax^2$  と  $-ax^2$  が出てくるので、この項が消えてくれます。式を変形しているときに、こういうふうにきれいに消えてくれるのはうれしいですね。

0 で割り算することは禁止されていますが、 $h$  は 0 ではなく、0 に限りなく近い値なので約分ができます。また、

$$\lim_{h \rightarrow 0} h = 0$$

と定義されており、

$$\lim_{h \rightarrow 0} ah = a \lim_{h \rightarrow 0} h = 0$$

で、さらに、多項式の場合は各項に分けて計算できるので、結局のところ、上の式は以下ようになります。

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} (2ax + \underbrace{ah}_{\rightarrow 0}) \\ &= 2ax \end{aligned}$$

$2ax$  の項は  $h$  の値が変わっても  $2ax$  のままですね。定義通りに計算していくと、

$$(ax^2)' = 2ax$$

となりました。一方、目標【その4】では、

$$(ax^n)' = anx^{(n-1)}$$

という簡単な計算方法だけを示しましたが、この式の  $n$  に 2 を当てはめると、

$$(ax^2)' = 2ax^{(2-1)} = 2ax$$

となり、上で計算した結果と同じになっていることが分かります。というわけで、目標【その4】で見た計算方法は、ちゃんと微分の定義から導き出せるようだ、ということが分かりました（「ようだ」と言葉を濁したのは、これがちゃんとした証明になっておらず、 $(ax^2)'$  の場合には答えが合っていることを確認しただけからです）\*6。



\*6 とはいえ、ちゃんとした証明になっていないとスッキリしないという人のために、式だけ書いておきます（※下記の証明内容を知らなくても実用上の問題はありません）。 $n$ 乗の式の展開の説明などは省き、ざっくりとした道筋だけしか書いていませんが、 $(x^n)' = nx^{n-1}$  を証明すればいいので、以下のようになります。

$$\begin{aligned}
 f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} \\
 &= \lim_{h \rightarrow 0} \frac{x^n + nhx^{n-1} + \overbrace{\dots + h^n}^{\text{ここはすべて } h^2 \text{ 以上の項}} - x^n}{h} \quad \dots \boxed{A} \\
 &= \lim_{h \rightarrow 0} (nx^{n-1} + \overbrace{\dots + h^{n-1}}^{\text{ここにはすべて } h \text{ が含まれる}}) \quad \dots \boxed{B} \\
 &= nx^{n-1} \quad \dots \boxed{C}
 \end{aligned}$$

[A]  $\dots x^n$  は+と-で消える

[B]  $\dots h$  で約分した結果、 $h$  を含まないのは  $nx^{n-1}$  だけになる

[C]  $\dots h$  が限りなく 0 に近づくので  $nx^{n-1}$  だけが残る

要するに、 $n$ 乗の項はプラスとマイナスで消え、他の項には  $h$  が掛けられているので、結局  $n-1$  乗の項しか残らないということです。

では、最後に少しだけ応用的な例題でおさらいをして終わりにしましょう。といっても、同じような式ばかりではつまらないので、これまでに学んだ  $\Sigma$  にも登場してもらいましょう。

$$f(\bar{x}) = n\bar{x}^2 - 2\left(\sum x_i\right)\bar{x} + \sum x_i^2$$

を、

$\bar{x}$  で微分

してみてください。

「なんじゃこりゃあ」と思った人もいるかもしれませんが、

### 変数として $\bar{x}$ を使っていること

に注目すれば、手順通りに計算できます（図 10）。 $\Sigma$ が含まれていますが、それらは単なる係数（定数）にすぎません。

The diagram illustrates the differentiation of the expression  $(n\bar{x}^2 - 2(\sum x_i)\bar{x} + \sum x_i^2)'$ . The original expression is at the top, with the terms  $n\bar{x}^2$ ,  $2(\sum x_i)\bar{x}$ , and  $\sum x_i^2$  highlighted in yellow. Arrows point from these terms to the differentiated result  $2n\bar{x} - 2(\sum x_i)$  at the bottom, which also has  $2n\bar{x}$  highlighted in yellow. The annotations are as follows:

- ① 指数を係数に掛ける (Multiply the exponent by the coefficient)
- ② 指数を1減らす (Decrease the exponent by 1)
- ③ 指数は1なので係数に掛けても係数は変わらない (Since the exponent is 1, multiplying by the coefficient does not change the coefficient)
- ④ 指数を1減らすと  $x^0$ つまり1になる (Decreasing the exponent by 1 results in  $x^0$ , which is 1)
- ⑤ 定数項は無視 (Constant terms are ignored)

図 10  $\Sigma$ を係数に含む関数を微分する

最初の項は、 $n\bar{x}^2$ ですね。

図10の①で指数を係数に掛けた後、

②で指数を1減らすと $2n\bar{x}$ になります。

次の項は、 $-2(\sum x_i)\bar{x}$ となっています。

$\bar{x}$ の指数は1なので、それを③で係数に掛けても $-2(\sum x_i)$ のままです。

④で指数を1減らした $\bar{x}^0$ は1なので、結局この項は $-2(\sum x_i)$ になります。

最後の $\sum x_i^2$ は⑤定数項なので無視します。

従って、答えは、

$$f'(\bar{x}) = 2n\bar{x} - 2(\sum x_i)$$

となります。ちなみに、

この式を0と置いて $\bar{x}$ についての方程式を解くと、

$$\bar{x} = \frac{\sum x_i}{n} \text{ となり、}$$

平均を求める式が導き出せます。



さて、ずいぶんといろんなお話をしてきましたが、微分の計算そのものはとても簡単だということが分かったと思います。とはいえ、単に計算ができるのと、意味をちゃんと理解しているのは大違いです。平均変化率のお話から微分の定義を学び、それを基にいくつかの例で導関数を求めたことにより、微分の意味がよく分かった（思い出せた）と思います。

実際には、関数が分かっている、導関数が簡単には求められない（ここで示した簡単な計算方法だけでは微分できない）関数や方程式もあります。そのような関数を取り扱うためには、コンピュータシミュレーションによって数値計算をすることがあります。その場合、定義に戻って  $h$  の値を少しずつ変えていくことにより、近似値を求めたりします。そういうわけで、微分の定義が役に立つ日がいずれやって来ることと思います。

## 次回は……

今回は微分の考え方と基本的な計算の方法について説明しました。毎回、基本のお話に加えて、いくつかの公式と機械学習につながる応用例についても触れているのですが、今回は学ぶべきことが多く、やや複雑なところもあったので、微分に関するいくつかの公式と、応用例は[次回に譲ること](#)とします。



# [AI・機械学習の数学]微分法を応用して 回帰分析の基本を理解する

微分の考え方と計算方法を理解したら、次は微分の公式を押さえて活用してみよう。幾つかの公式を紹介し、応用例として回帰分析を行うための最小二乗法について基本的な考え方を見ていく。

羽山博, 著 (2020 年 06 月 04 日)

微分の考え方と計算方法については、[前回](#)、一通り見ました。微分は、最小値を求めたり、最適な値に近づけていったりする計算に使えるというお話はしましたが、具体例はまだ紹介していませんでした。

そこで、今回は、**最小二乗法**を利用して**回帰分析**を行う方法を紹介します。最初に少し微分の公式をおさらいしつつ整理しておき、続いて回帰分析の簡単な例へと進みます。

## 目標【その 1】：微分の公式を整理しておく

以下に示す公式は微分の計算をするときに必ず使う公式です \*1。

$$\{a \cdot f(x)\}' = a \cdot f'(x) \quad (a \text{ は定数}) \quad (1)$$

$$\{f(x) + g(x)\}' = f'(x) + g'(x) \quad (2)$$



\*1 かっこが二重になって見づらいときには、外側のかっこを {} にして見やすくすることがあります。

公式 (1) を日本語で読み下すと、

「関数の **a 倍** を微分したものと、関数を微分したものを **a 倍** したものと、が等しい」という意味になります。

一方、公式 (2) は、

「関数の和の微分は、関数の微分の和と等しい」といった意味になります。これらの式は「**微分の線形性**」と呼ばれる性質を表したものです（※「線形性」の意味は後述します）。

さらに、公式 (1) と公式 (2) から、以下の公式 (3) も導き出せます。

$$\{a \cdot f(x) + b \cdot g(x)\}' = a \cdot f'(x) + b \cdot g'(x) \quad (a, b \text{ は定数}) \quad (3)$$

実は、これらの公式は直感的に理解できるので、これまで当然のように何度も使ってきました。微分の基本の基本でもあり、今回のテーマである最小二乗法を利用した回帰分析でも使うので、ここできちんと確認しておこうというわけです。

### 解説：微分の公式を整理しておく

目標のところに記した(1)(2)の公式については、極限(lim)に戻れば、簡単に証明できます。しかし、本筋のお話から外れるので証明は割愛し、ここでは公式の意味と具体例を確認するにとどめます。

まず、公式(1)の意味を見ると図1のようになります。

$$\{a \cdot f(x)\}' = a \cdot f'(x)$$

$a$ を  $f(x)$  に掛けたもの  
の微分  $=$   $a$ を  $f(x)$  の微分  
に掛けたもの

図1 定数倍は微分の外に出せる

要するに、関数の定数倍は、微分の外に出せるということです。「元の関数に係数  $a$  が掛けられているときには、微分して係数  $a$  を掛けてもいい」ということですね。

ただし、公式(1)では、 $a$  が定数であることに注意してください。 $x \cdot f(x)$  のように関数の中で使われている変数を掛けたものだと、この式は成り立ちません\*2。



**\*2** そのような場合は、式を展開してから微分してください。例えば  $f(x) = 2x^2 + 1$  のときは  $xf(x) = x(2x^2 + 1) = 2x^3 + x$  のように展開してから微分します。答えは  $6x^2 + 1$  ですね。この計算には、微分の積の公式も使えますが、それについては回を改めて紹介します。

ここで、微分の基本的な計算方法について、ちょっと細かく見てみます。分かり切った話を回りくどく説明する感じになるので、面倒に感じたら公式(2)の説明まで読み飛ばしてもらって結構です。

さて、ちょっと細かな話です。そもそも、微分の最もシンプルな形(公式)は以下のような式です。計算方法は、**(1)** 指数を係数に掛ける、**(2)** 指数を1減らす、でしたね。

$$\begin{aligned} f(x) &= x^n \text{ のとき、} \\ f'(x) &= (x^n)' = nx^{n-1} \end{aligned} \tag{0}$$

公式（1）と公式（0）を使えば、以下のような計算ができます。

$$\begin{aligned}
 f(x) = x^n \text{ のとき、} \\
 \{a \cdot f(x)\}' &= (a \cdot x^n)' \\
 &= a(x^n)' \quad \dots \boxed{A} \\
 &= anx^{n-1} \quad \dots \boxed{B}
 \end{aligned}$$

[A] …… 公式（1）を適用して  $a$  を外に出した

[B] …… 公式（0）を適用して  $x^n$  を微分した

これまで、しごく当たり前のように、

$$(2x^3)' = 2 \cdot 3x^2 = 6x^2$$

のような計算をしてきましたが、上で見たように、細かくステップを分けて公式を当てはめると以下のような計算になります。

$$\begin{aligned}
 (2x^3)' &= 2 \cdot (x^3)' \quad \dots \boxed{A} \\
 &= 2 \cdot 3x^2 \quad \dots \boxed{B} \\
 &= 6x^2
 \end{aligned}$$

[A] …… 公式（1）を適用して  $2$  を外に出した

[B] …… 公式（0）の公式を適用して  $x^3$  を微分した

続いて、公式（2）についても図2で確認しておきましょう。

$$\begin{array}{ccc}
 \{f(x) + g(x)\}' & = & f'(x) + g'(x) \\
 \swarrow & & \swarrow \\
 \text{関数を足したもの} & = & \text{微分したもの} \\
 \text{の微分} & & \text{を足したもの}
 \end{array}$$

図2 足してから微分しても、微分してから足しても同じ

これまで「多項式の場合は各項を微分すればいい」とお話してきましたが、この公式がそういった計算を行う根拠です。具体例で問題を解きながら確認しておきましょう。

$$\begin{aligned}
 f(x) &= 4x^3 + 4x + 1 \\
 g(x) &= 5x^2 + 2x + 3
 \end{aligned}$$

のとき、公式（2）の左辺は以下の通りになります。

$$\begin{aligned}
& \{f(x) + g(x)\}' \\
&= \{4x^3 + 4x + 1 + 5x^2 + 2x + 3\}' \cdots [A] \\
&= (4x^3 + 5x^2 + 6x + 4)' \cdots [B] \\
&= 12x^2 + 10x + 6 \cdots [C]
\end{aligned}$$

[A] ……  $f(x)$  と  $g(x)$  を当てはめた

[B] …… 次数の降順に整理した

[C] …… 各項の微分を計算した

計算は単なる掛け算と足し算だけなので難しくはないですね。何気なく式を変形して計算をしています、実をいうと C 式への変形のところで、しれっと公式 (2) そのものを使っています (各項を個別に微分していますよね)。ですが、まあ、ここは式の確認と計算練習ということで目をつむりましょう。

公式 (2) の右辺にも  $f(x)$  と  $g(x)$  を当てはめてみます。

$$\begin{aligned}
& f'(x) + g'(x) \\
&= (4x^3 + 4x + 1)' + (5x^2 + 2x + 3)' \cdots [A] \\
&= (3 \cdot 4x^2 + 4) + (2 \cdot 5x + 2) \cdots [B] \\
&= 12x^2 + 10x + 6 \cdots [C]
\end{aligned}$$

[A] ……  $f(x)$  と  $g(x)$  を当てはめた

[B] …… 各項の微分を計算した

[C] …… 次数の降順に整理した

正しく計算できていれば、左辺の結果と右辺の結果が一致するはずですが (毎度のことですが、これは証明ではなく、1 つの例で単に答えが合うことを確認しただけです)。

さて、目標のところに記したように、これらの公式は「微分の線形性」を表しています。

「線形性」というのは、公式 (1) と公式 (2) のような性質が成り立つことを言います (まとめて言えば公式 (3) のような性質が成り立つことです)。つまり、

- (1) 「項を何倍かしたものに何らかの計算をしたもの」と「項に何らかの計算をしたものを何倍かしたもの」が等しくなる
- (2) 「各項を足したものに対して何らかの計算 (例えば微分) をしたもの」と「何らかの計算 (微分) を個々の項にしたものを足したもの」とが等しくなる

という性質と考えればいいでしょう。今回は「何らかの計算」は「微分」でしたが、極限 (lim) や総和 (Σ) に関しても線形性が成り立ちます。

ちなみに、公式 (3) は以下のように導き出せます。公式 (3) の記述とは逆に、右辺から左辺を導き出していきますね。

$$\begin{aligned} a \cdot f'(x) + b \cdot g'(x) &= \{a \cdot f(x)\}' + \{b \cdot g(x)\}' \quad \dots \boxed{A} \\ &= \{a \cdot f(x) + b \cdot g(x)\}' \quad \boxed{B} \end{aligned}$$

[A] …… 公式 (1) を逆に適用して **a** と **b** を微分の中に入れた

[B] …… 公式 (2) を逆に適用して個々の微分をまとめた

公式 (3) 式は、公式 (2) に定数 **a** と **b** を掛けただけなので、さほど理解に苦しむことはないと思います。

さらに、微分の計算をしっかりと身に付けるための練習問題を以下に用意しておきますが、早く回帰分析をやりたい、という人は練習問題を飛ばして次に進んでもらっても構いません。

## 練習問題

(1) 微分の公式を書いてみましょう

$$\{a \cdot f(x) + b \cdot g(x)\}' = a \cdot f'(x) + b \cdot g'(x) \quad (a, b \text{ は定数})$$

(2) 以下の計算を行ってみましょう

$$\begin{aligned} f(x) &= x^3 - 2x^2 + 1 \\ g(x) &= 5x^2 + x + 3 \end{aligned}$$

のとき、(1) で書いた公式の **a** に当たる値を **3**、**b** に当たる値を **2** とします。このとき、公式の左辺での計算と右辺での計算をやってみてください。

左辺は以下の通りです。

$$\begin{aligned} &\{a \cdot f(x) + b \cdot g(x)\}' \\ &= \{3 \cdot (x^3 - 2x^2 + 1) + 2 \cdot (5x^2 + x + 3)\}' \quad \dots [A] \\ &= (3x^3 - 6x^2 + 3 + 10x^2 + 2x + 6)' \quad \dots [B] \\ &= (3x^3 + 4x^2 + 2x + 9)' \quad \dots [C] \\ &= 9x^2 + 8x + 2 \quad \dots [D] \end{aligned}$$

- [A] …… **a** に **3** を代入し、**b** に **2** を代入した
- [B] …… **3** と **2** を分配して式を展開した
- [C] …… 次数の降順に項をまとめた
- [D] …… 微分した

右辺は以下の通りです。

$$\begin{aligned}
 &af'(x) + bf'(x) \\
 &= 3(x^3 - 2x^2 + 1)' + 2(5x^2 + x + 3)' \quad \cdots [A] \\
 &= 3(3x^2 - 4x) + 2(10x + 1) \quad \cdots [B] \\
 &= 9x^2 - 12x + 20x + 2 \quad \cdots [C] \\
 &= 9x^2 + 8x + 2 \quad \cdots [D]
 \end{aligned}$$

- [A] …… **a** に **3** を代入し、**b** に **2** を代入した
- [B] …… 各項を微分した
- [C] …… **3** と **2** を分配して式を展開した
- [D] …… 次数の降順に項をまとめた

## 目標【その2】：回帰分析とは何かを知る

では、微分の応用例として回帰分析への適用について見てみましょう。まず、回帰分析とは何かを簡単に説明しておきます。**回帰分析**とは、データを何らかの式に当てはめること、と考えていいでしょう。そのような式のことを回帰式と呼びます。**回帰式**としては、直線の式が最も単純な例です。例えば、賃貸住宅の面積と家賃が直線的な関係になっているものとしましょう。このとき回帰式（直線の式）は、

$$y = ax + b$$

となります。

この場合、面積が **x**、家賃が **y** に当たることが分かりますね。**a** は係数（直線の傾き）、**b** は定数項（切片、つまり **x = 0** のときの **y** の値）です。

回帰分析は表計算ソフト「Excel」などでも簡単にできます。

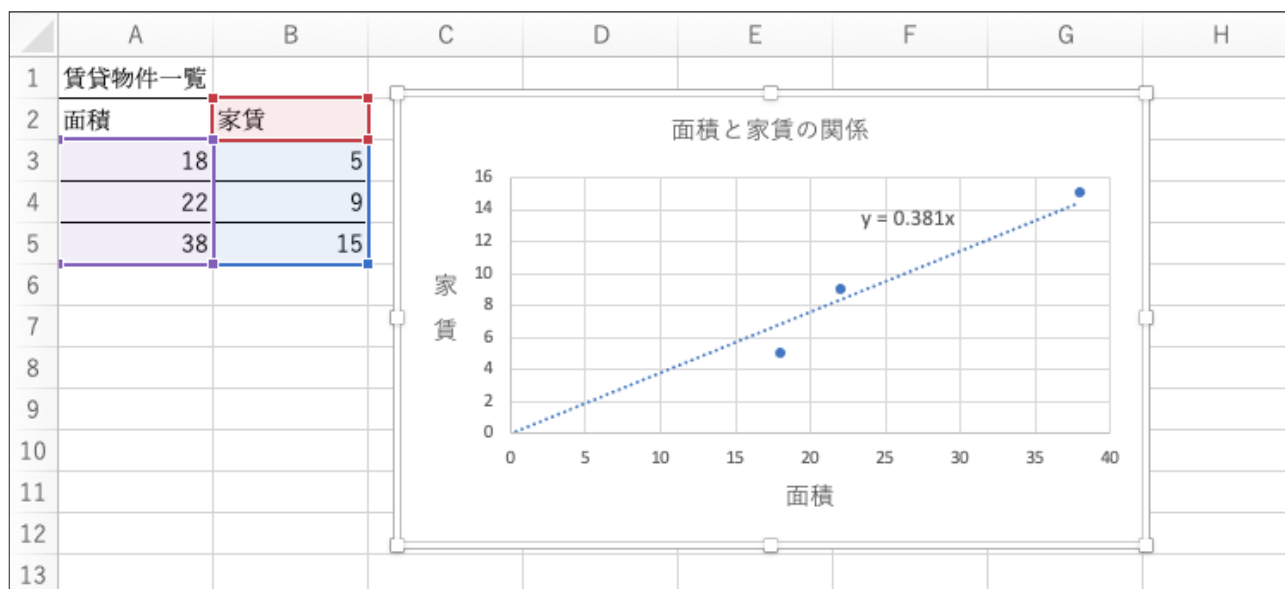


図3 回帰分析のために散布図と近似曲線を描いた例

図3では、セルA3には18、セルB3には5という値が入っていますが、これは面積が18 m<sup>2</sup>の物件の家賃が5万円だという意味です。同様に、22 m<sup>2</sup>の物件なら9万円、38 m<sup>2</sup>の物件なら15万円という3つのデータがありますね。グラフはそれらのデータを基に作った散布図です。散布図では、回帰式を表す直線やその式を表示できるので、併せて示してあります。

グラフの中にも記されているように、回帰式の係数aは0.381となっていることが分かります（ここでは、定数項のbは0として計算しています）。また、回帰式が分かれば、元のデータにはなかった面積の物件でも、家賃が予測できることが分かりますね。例えば、面積が30 m<sup>2</sup>なら、家賃は0.381×30=11.43万円と予測できます。

なんだ、Excelで求められるのなら別に微分の話とか、機械学習の話とかは知らなくてもコトは足りるんじゃないの、と思われるかもしれませんが、しかし、回帰式が複雑になるとExcelで計算するのが難しい場合もあります。そのような状況を考えると、どのような仕組みで定数項や係数が求められるのかを知っておくことは重要です。式さえ立てられれば、あとはコンピュータ（統計パッケージやプログラムなど）に計算させればよいというわけです。

## 解説：回帰分析とは何かを知る

回帰分析の方法は、ズバリ、測定されたデータの最も近くを通る回帰式を求める、という方法です。回帰式が直線の場合は、もちろん直線の式  $y = ax + b$  (の  $a$  と  $b$ ) を求めるわけです。

ところで、「最も近くを通る」というのはどういうことでしょうか。図 4 で考えてみましょう。

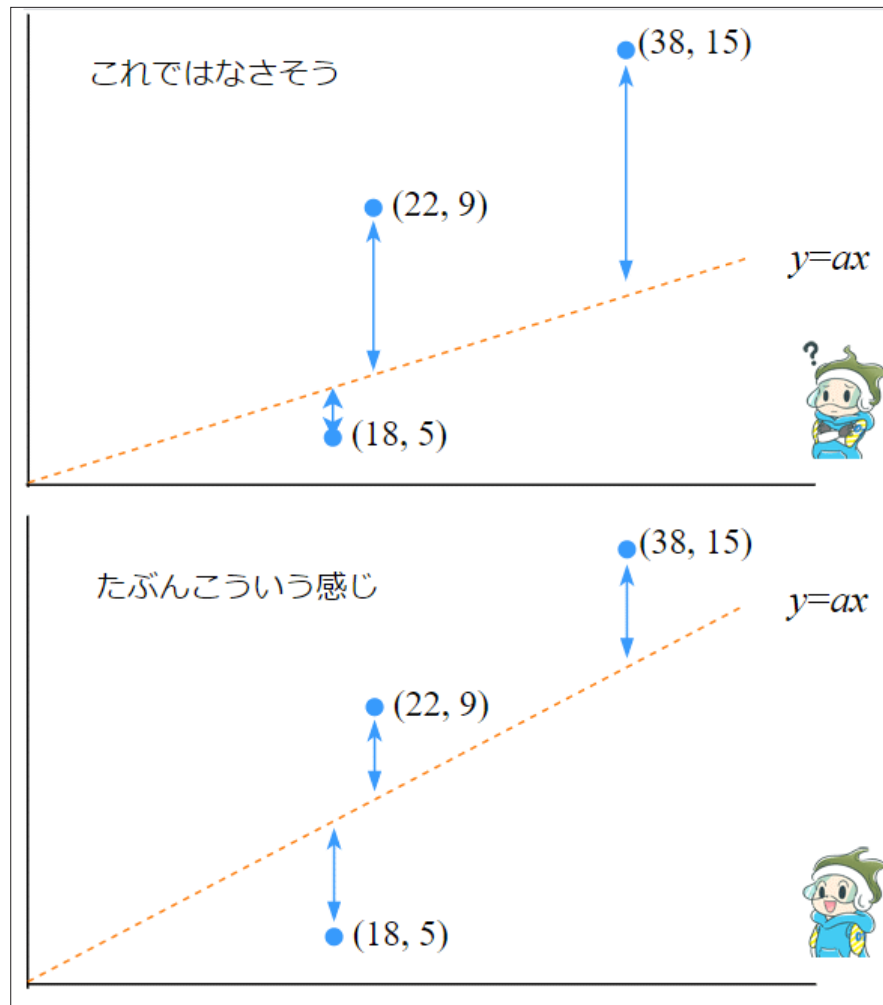


図 4 データの最も近くを通る直線を探す



データの近くを通りそうな直線をざっくりと引いてみましたが、図4の上のグラフでは直線があまりデータの近くを通ってはいないようです。一方、下のグラフはデータの近くを通っていそうですね。図からも想像が付くと思いますが、「最も近くを通る回帰式（直線の式）を求める」というのは、各データと直線の距離の二乗和が最も小さくなるように、傾き  $a$  を決めるということです \*3。



\*3 一般に「距離」というとユークリッド距離のことを表すので、各点から直線に引いた垂線の長さが距離となりますが、ここで「距離」と呼んでいるのは  $y$  方向の距離のことです。これは、実際の値と、 $y = ax$  という式に従うならこうなるはずだという理論的な値との差なので、「誤差」と考えられますね。従って、これ以降は「誤差」と呼ぶことにします。

単なる和ではなく二乗和を使う理由は連載の第2回でお話しました。単なる和だとプラスとマイナスで相殺されてしまうから、ですね。また、絶対値の和だと、

**最小値になるときの  $\bar{x}$  の値（ここでは  $a$  の値）**

が1つに決まらないこともあるので、二乗和を使う、ということでした。

そうと分かれば、さっそく回帰式を求めてみたいと思うのが人情ですが、その前にもう少しだけ回帰分析について説明しておきます。その方が意味合いもよく分かりますし、先を見通すこともできるので。

既にお気づきかもしれませんが、当然のことながら、家賃は面積だけで決まるわけではありません。築年数や駅からの距離、階数などによっても変わるでしょう。そういうわけで、 $x$  に当たる変数が複数ある場合もあります。すると、直線ではなく平面やさらにそれ以上の次元（超平面）の、ちょっと図では表しにくい式になります。図5は  $x$  に当たる変数が2つある場合のイメージで、回帰式はこのような3次元空間の平面で表現できます。

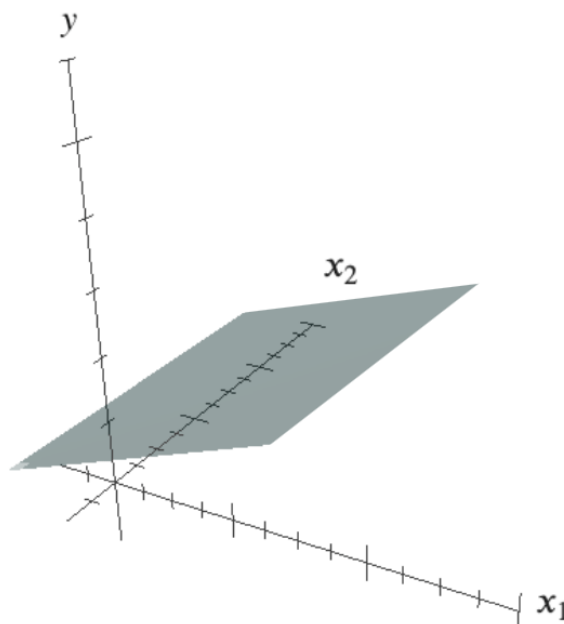


図5 重回帰分析のイメージ（説明変数が2つある場合はデータの最も近くを通る平面を探す）

このとき、軸を表す文字をたくさん使うのはわずらわしいので、添字を使って変数を区別するのが普通です。例えば、上の「家賃計算」の例では、家賃を  $y$ 、面積を  $x_1$ 、築年数を  $x_2$ 、駅からの距離を  $x_3$  のように表します（係数や定数項も添字で区別します）。 $x_1$  や  $x_2$ 、 $x_3$  によって  $y$  を説明するので、これらの変数のことを**説明変数**と呼びます。一方、説明される方の  $y$  は**目的変数**と呼ばれます。また、説明変数が1つの場合の回帰分析のことを**単回帰分析**、説明変数が複数の場合の回帰分析のことを**重回帰分析**と呼びます **\*4**。



**\*4** 目的変数は「従属変数」、説明変数は「独立変数」などとも呼ばれます。

新しい用語がたくさん出てきたので、ちょっとまとめておきましょう。図6の式が回帰分析で使われる回帰式の例です。ここでは、項の順序を変えて、定数項を最初に書いてあります。

単回帰分析

$$y = b + ax$$

目的変数                  説明変数

重回帰分析

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$$

目的変数                  説明変数

図6 単回帰分析と重回帰分析の回帰式

単回帰にしろ重回帰にしろ、与えられたデータを基に回帰式の定数項や係数を求めることに変わりはありません。重回帰の場合、定数項や係数がたくさんある、というだけのことです。ここでは定数項や係数を  $a_i$  という文字で表していますが、機械学習では、定数項や係数を  $w_i$  や  $\theta_i$  という文字で表すことがあります。 $w$  は重み (weight) の略です。 $\theta$  は角度を表すのにも使われますが、パラメータ（何かを特徴付ける値など）を表すときにも使われます。なお、ここでは全ての項が一次式である場合の例を示していますが、二次の項 ( $x_i^2$ ) などを含めた方が、当てはまりのいい場合も考えられます。

### 目標【その3】：最小二乗法による回帰分析の超簡単な例

既にお話したように、回帰分析では各データの最も近くを通る回帰式の定数項と係数を求めます。そのためには「各データと回帰式の  $y$  方向の距離（誤差）の二乗の総和」を求め、それが最小になるように定数項と係数を決めればいいということでした（このような方法のことを「最小二乗法」と呼びます）。ここからは、実際にそういった計算をしていきます。

とはいえ、最初なのでできるだけ簡単な例にしましょう。目標その1で見た面積と家賃の例を使うことにします。面積が0なら家賃も0だと考えていいでしょうから、定数項は0とします。つまり、この単回帰分析の回帰式は、前のページで解説した  $y = b + ax$  に当てはめると以下ようになります。

$$y = ax$$

$y$  が家賃、 $x$  が面積、 $a$  は係数です（ $b$  は0なので消えます）。簡単すぎますね。でも、最初から複雑な例に取り組もうとしても訳が分からなくなるだけなので、あえてここからスタートします。

答えを先に示しておきましょう。求めたい値は、各データを  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  で表したときに、各データと回帰式との間の誤差の二乗和が最小になるときの  $a$  の値です。 $a$  を求める式は最終的に、

$$a = \frac{\sum (x_i \cdot y_i)}{\sum x_i^2}$$

となります。うひゃあ、簡単な例でもこんなに難しいの？ と驚いた人もいるかもしれませんが、日本語で読み下せば恐るるに足らずです。

分子は「各データの  $x$  と  $y$  を掛けたものの総和」ですね。それを、分母の「 $x$  の二乗和」で割った値、ということになります。

具体例を当てはめると、分子は「各データの面積 × 家賃を総合計したもの」で、分母は「各データの面積の二乗を総合計したもの」です。

というわけで、式の計算内容（どういう計算なのか）は分かったと思います。謎なのは、どうやってこの式が導き出されるのかということですね。

今回の目標は、この式の計算内容が理解できること（もうできましたね）と、ちゃんと導き出せることです。一歩ずつ確実に進めていけば、必ずこの式にたどり着けるのでご心配なく。一緒に見ていきましょう。

解説：最小二乗法による回帰分析の超簡単な例（具体的な例）

いきなり、Σ満載の式をお見せしましたが、実感が湧くように、具体例を使って答えを導き出す流れを見ておきましょう（ここからの説明には動画も用意してあります）。



動画 1 最小二乗法による回帰分析（具体的な例）

まず、図 7（図 1 の再掲）の表から面積と家賃の最初（＝表中の 3 行目）のデータを見てください。

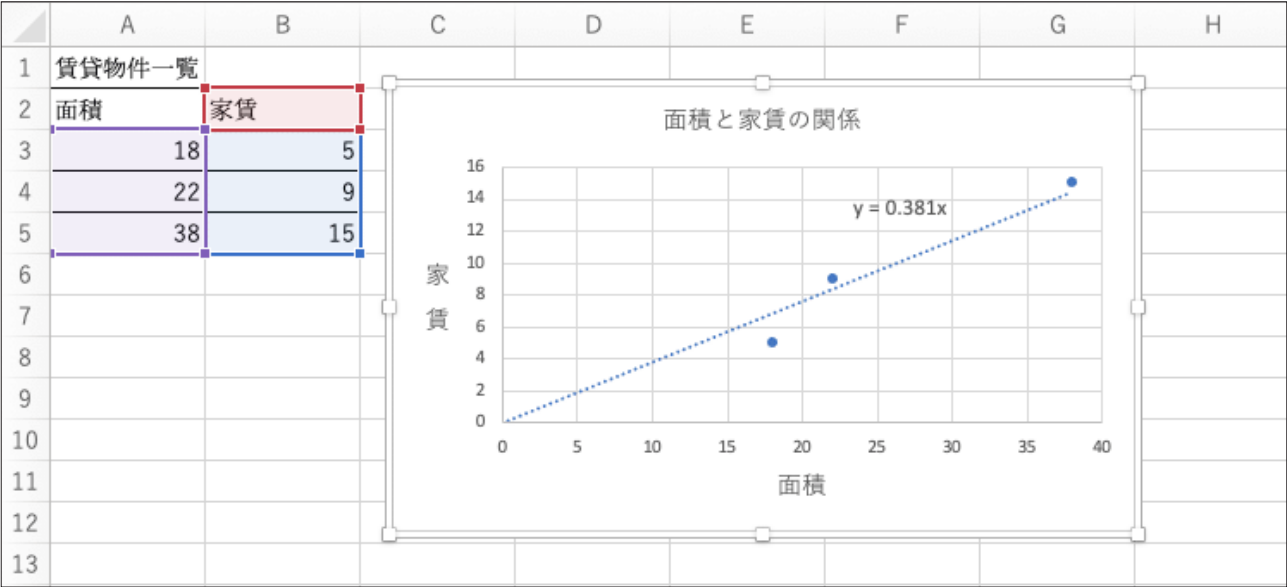


図 7 面積と家賃のデータとグラフ（図 1 の再掲）

最初のデータは面積  $x$  が 18、家賃  $y$  が 5 です。もし、家賃が（前述のように） $y = ax$  で求められるなら、直線上の点の  $y$  座標は  $18a$  です。

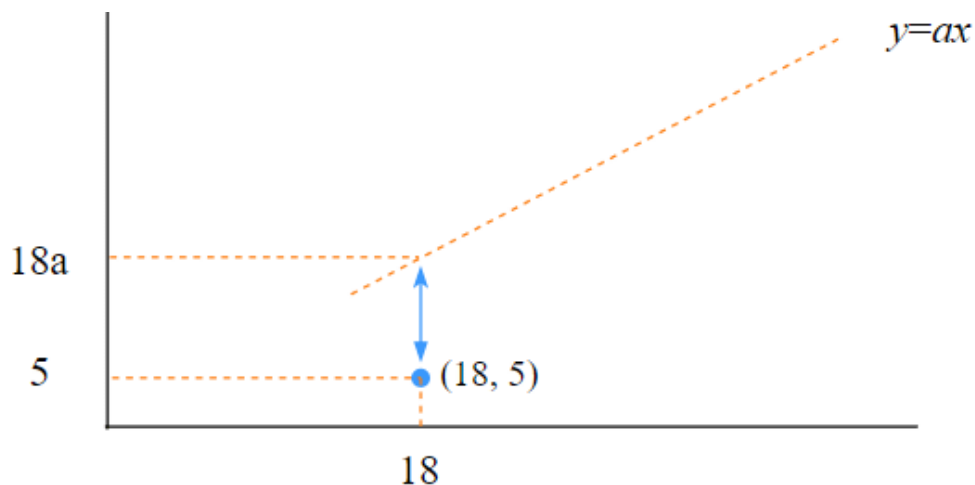


図 8 測定値と回帰式（直線）との間の誤差

図 8 の青い矢印がデータと直線との距離だということが分かりますね。従って、「実際のデータの y 座標 **(5)**」と「直線上の y 座標 **(18a)**」との距離の二乗は、以下の式で求められます。

$$(5 - 18a)^2$$

ただし、**a** が先にあった方が計算が楽になるので、以下のように表すことにします。二乗すれば必ず正になるので、上の式を以下のように書き換えても結果は同じです。

$$(18a - 5)^2$$

他のデータについても、「実際のデータの y 座標」と「直線上の y 座標」との距離、つまり誤差の二乗は図 9 のように求められます。

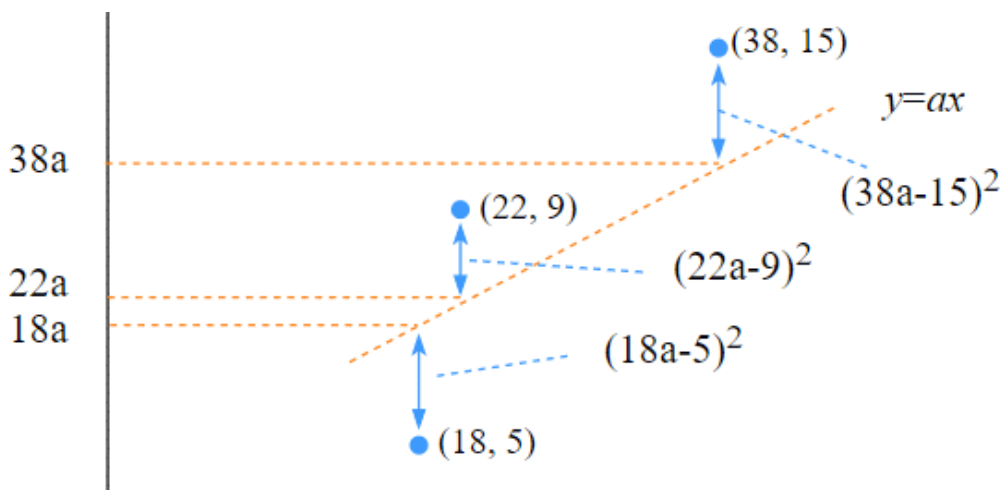


図 9 測定値と回帰式（直線）との間の誤差の二乗を求める

私たちがやりたいことは、誤差の二乗和を最小にする  $a$  の値を求めることでしたね。そこで、まずはこれらの誤差の二乗和を求めてみましょう。誤差の二乗を全部足せばいいので、以下のようになります。

$$(18a - 5)^2 + (22a - 9)^2 + (38a - 15)^2 \quad (1)$$

二乗の展開公式、つまり  $(a - b)^2 = a^2 - 2ab + b^2$  を使って、(1) 式を展開しておきましょう（説明を聞きながら流れを追いかけてみたい人は[前掲の動画 1](#) を利用してみてください）。

$$\begin{aligned} & (18a - 5)^2 + (22a - 9)^2 + (38a - 15)^2 \\ &= (18a)^2 - 2 \cdot 18 \cdot 5a + 5^2 \\ & \quad + (22a)^2 - 2 \cdot 22 \cdot 9a + 9^2 \\ & \quad + (38a)^2 - 2 \cdot 38 \cdot 15a + 15^2 \quad \cdots [A] \\ &= (18^2 + 22^2 + 38^2)a^2 - 2(18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15)a + (5^2 + 9^2 + 15^2) \quad \cdots [B] \end{aligned}$$

[A] …… 二乗の展開の公式を適用した

[B] …… 次数の降順にまとめた

数値の計算は後回しにします。例えば、 $18 \cdot 5$  などの計算は後でやるということです。もしかすると後で約分して式を簡単にできたりするかもしれませんが、また、微分すると定数項が消えるので、いちいち計算するのは無駄だというわけです（というわけで、次に微分を使うことがバレてしまいましたね）。

続けましょう。測定値の一番近くを通る直線を求めるには、誤差の二乗和を最小にする  $a$  の値を求めるということですね。ひと目見て分かると思いますが、上の式は  $a$  の二次式で、 $a^2$  の係数は正です。グラフは下に凸な形になります。図 10 のような感じです。

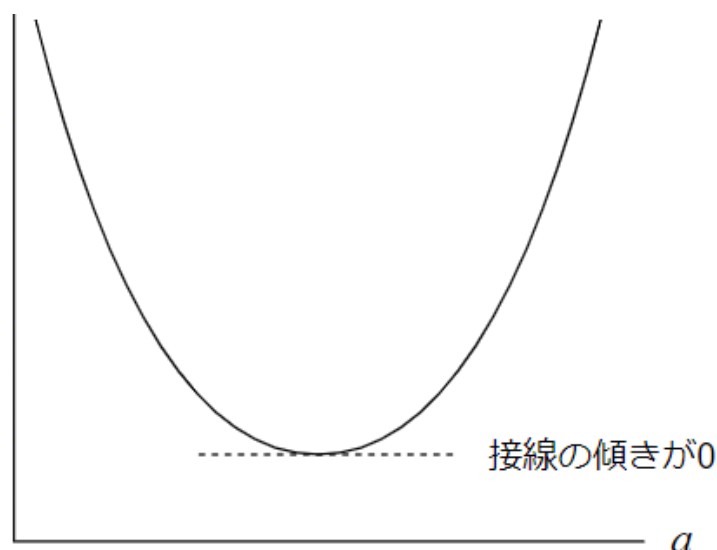


図 10 下に凸な二次関数の値が最小になるのは接線の傾きが 0 になったとき



このグラフを見れば、二次式が最小の値になるのは接線の傾きが **0** のときだということが分かりますね。接線の傾きはどうすれば求められたかというところ……そう、前回やったとおり、微分すれば求められます **\*5**。



**\*5** この連載の **2** 回目を見たように、平方完成でも求められますね。ちなみに、二乗のことは「平方」とも呼ばれるので、誤差の二乗和は「誤差平方和」とも呼ばれます。

では、(1) 式を展開したものを **a** で微分し、**0** と置いてみましょう。以下の式で、アンダーラインを引いた部分は定数なので、簡単に微分できます。

$$\frac{\{(18^2 + 22^2 + 38^2)a^2 - 2(18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15)a + (5^2 + 9^2 + 15^2)\}}{2} = 0 \quad \dots [A]$$

$$2(18^2 + 22^2 + 38^2)a - 2(18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15) = 0 \quad \dots [B]$$

[A] …… 微分して **0** とおいた

[B] …… 微分の計算を行った

最初の項は  $\{a^2\}' = 2a$  を利用すれば求められます。また、2 番目の項は  $a' = 1$  を利用すれば求められます。定数項は消えてしまうので、ずいぶん簡単になりましたね。B 式は **a** についての簡単な方程式であることも分かります。あとは一気に解いていきましょう。

$$2(18^2 + 22^2 + 38^2)a - 2(18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15) = 0 \quad \dots [B]$$

$$2(18^2 + 22^2 + 38^2)a = 2(18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15) \quad \dots [C]$$

$$(18^2 + 22^2 + 38^2)a = (18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15) \quad \dots [D]$$

$$a = \frac{18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15}{18^2 + 22^2 + 38^2} \quad \dots [E]$$

$$a = \frac{90 + 198 + 570}{324 + 484 + 1444} \quad \dots [F]$$

$$a = \frac{858}{2252} \quad \dots [G]$$

$$a = 0.381 \quad \dots [H]$$

[B] …… 上の式（再掲）

[C] …… 移項した ( $2a - 2b = 0 \rightarrow 2a = 2b$  のようにした)

[D] …… 両辺を **2** で割った ( $2a = 2b \rightarrow a = b$  のようにした)

[E] …… 両辺を  $18^2 + 22^2 + 38^2$  で割って「 $a = \bigcirc\bigcirc$ 」の形にした

[F] ～ [H] …… 分母と分子を計算して、答えを求めた

というわけで、係数  $a$  の値が求められました。回帰式は  $y = 0.381x$  ということになります。最初に見た Excel での例（図 1）と一致していますね。

### 解説：最小二乗法による回帰分析の超簡単な例（一般的な例）

今度は、一般的な場合で考えてみましょう。いきなり文字式を使って計算していてもいいのですが、せっかく具体例を見たので、ちょっと「アタリ」を付けておきましょう。具体例のデータは以下のようなものでした。

x (面積)	y (家賃)
18	5
22	9
38	15

表 1 面積と家賃のデータ

具体例の計算途中にあった E 式を思い出してみてください（図 11）。

$$\begin{array}{c}
 x_1 \times y_1 + x_2 \times y_2 \dots\dots \Rightarrow \sum (x_i \cdot y_i) \\
 \swarrow \quad \downarrow \quad \downarrow \quad \swarrow \\
 a = \frac{18 \cdot 5 + 22 \cdot 9 + 38 \cdot 15}{18^2 + 22^2 + 38^2} \dots \boxed{E} \\
 \swarrow \quad \swarrow \\
 x_1^2 \quad x_2^2 \dots\dots \Rightarrow \sum x_i^2
 \end{array}$$

図 11 具体例と一般的な表し方を対応付ける

この式の分母は  $x$  の各要素の二乗和ですね。分子は  $x$  の各要素と  $y$  の各要素を掛けたものの総和です。なので、

$$a = \frac{\sum (x_i \cdot y_i)}{\sum x_i^2}$$

になりそうだ、ということが想像できます \*6。

**\*6** 計算練習って修行とか苦行のように思われますが、機械的に計算するのではなく、式の特徴や公式との関係を想像しながら進めると、一般的な式で考えるときにも実感が湧いてくると思います。



というわけで、本当にこうなるのかを次に見ていきます。動画でも解説しているので、ぜひ参照してみてください。



動画2 最小二乗法による回帰分析（一般的な場合）



動画2 最小二乗法による回帰分析（一般的な場合）

各データを  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  で表したとします。今のところ、回帰式は  $y = ax$  ですね（図 12）。まず、最初のデータについて考えます。実際のデータは  $(x_1, y_1)$  ですが、回帰式での  $y$  座標は  $y = ax$  の  $x$  に  $x_1$  を代入した値になります。つまり、 $ax_1$  ですね。

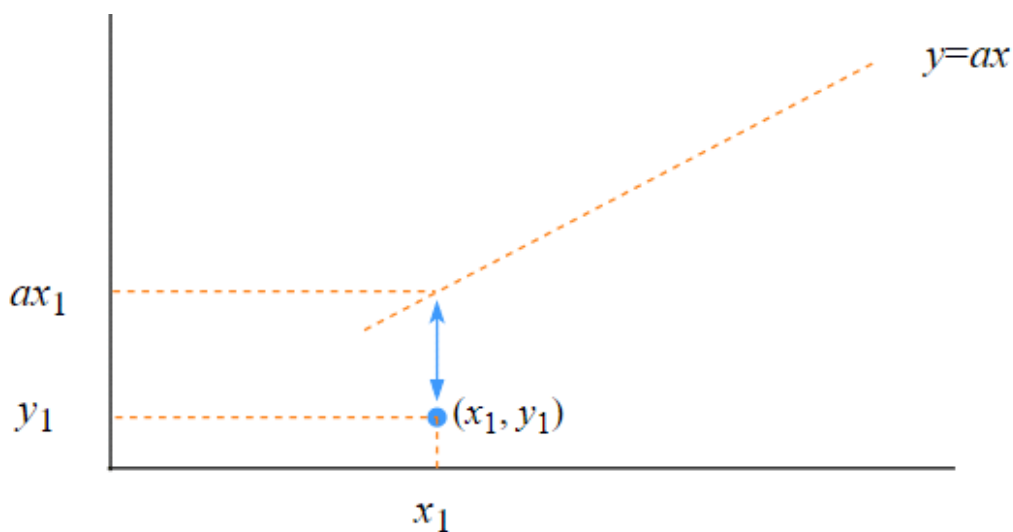


図 12 各データと回帰式（直線）との間の誤差を求める

誤差の二乗は以下の式で求められます。

となります。他のデータについても同様ですね。誤差の二乗の総和は以下ようになります。

$$(ax_1 - y_1)^2 + (ax_2 - y_2)^2 + \dots + (ax_n - y_n)^2 \\ = \sum (ax_i - y_i)^2$$

ですね。これを展開して、 $a$  について整理しましょう。

$$\begin{aligned}\sum (ax_i - y_i)^2 &= \sum (a^2 \cdot x_i^2 - 2a \cdot x_i \cdot y_i + y_i^2) \quad \dots [A] \\ &= \sum (a^2 \cdot x_i^2) - \sum (2a \cdot x_i \cdot y_i) + \sum y_i^2 \quad \dots [B] \\ &= (\sum x_i^2)a^2 - 2(\sum (x_i \cdot y_i))a + \sum y_i^2 \quad \dots [C]\end{aligned}$$

[A] …… 二乗の展開公式を使って展開した

[B] ……  $\Sigma$ を各項に分けた

[C] ……  $a$  について整理した

C 式への変形がちょっととまどうかもしれません。 $x_i$  や  $y_i$  は  $i$  によって異なる値となりますが、 $\Sigma$  の計算から見ると  $2$  や  $a$  は定数なので、 $\Sigma$  の外に出せます。

さて、 $a$  の係数は、 $x_i$  の二乗和なので正です。そして、 $a$  の二乗の式になっているので、これは下に凸な二次関数のグラフになります。この式が最小になるのは接線の傾きが  $0$  のときなので、微分して  $0$  と置いてみましょう。今度は  $a$  を変数として扱います。

$$\begin{aligned}\{(\sum x_i^2)a^2 - 2(\sum (x_i \cdot y_i))a + \sum (y_i^2)\}' &= 0 \quad \dots [A] \\ 2(\sum x_i^2)a - 2\sum (x_i \cdot y_i) &= 0 \quad \dots [B] \\ (\sum x_i^2)a &= \sum (x_i \cdot y_i) \quad \dots [C] \\ a &= \frac{\sum (x_i \cdot y_i)}{\sum x_i^2} \quad \dots [D]\end{aligned}$$

[A] …… 微分して  $0$  と置いた

[B] …… 微分の計算をした ( $a$  についての方程式になった)

[C] …… 移項して両辺を  $2$  で割った

[D] …… 両辺を  $\sum x_i^2$  で割った

A 式から B 式に変形するときに、微分の線形性の公式を使って各項の計算をしているのに気づいたでしょうか (ごく自然にやっているの、意識に上らなかったかもしれませんね)。式が複雑になっても、ちゃんと公式の形に当てはまっていることが分かります。

いかがでしょう。ちゃんと目標のところで示した式になりましたね。メデタシメデタシです！

ところで、ここまでは回帰式が  $y = ax$  という単純なものだったので、誤差の二乗和を微分して  $0$  と置くだけで係数  $a$  が求められました。しかし、定数項  $b$  がある場合（つまり  $y = b + ax$  の場合）はどうすればいいのでしょうか。また、重回帰分析の場合、つまり説明変数が  $x_1, x_2, \dots$  とたくさんある場合にはどうすればいいのでしょうか \*7。



\*7 ここでの  $x_1, x_2, \dots$  は、変数  $x$  の 1 番目の要素、2 番目の要素、……という意味ではありません。変数  $x_1$ 、変数  $x_2$ 、……という意味です。従って、変数  $x_1$  の最初の要素は  $x_{11}$ 、次の要素は  $x_{12}$  のように表されます。

実は、そのような場合には、**偏微分**と呼ばれる方法を使います。偏微分は高校の数学では出てこないかもしれませんが、単に 1 つの変数に注目し、他の変数は定数として微分するだけのことです。例えば、回帰式が  $y = b + ax$  の場合、上のように誤差の二乗和を求め、まず  $b$  で偏微分して  $0$  と置きます。次に  $a$  で偏微分して  $0$  と置きます。これで方程式が 2 つできる（ $a$  と  $b$  の連立方程式になる）ので、あとはそれを解くだけです。これは、変数が増えても同じことです。回帰式が、

$$y = a_0 + a_1x_1 + a_2x_2$$

であれば、 $a_0$ 、 $a_1$ 、 $a_2$  についての連立方程式ができるので、それらを解くだけです。

しかし、一度に詰め込んでも消化しきれないでしょうから、今回は偏微分についてはちょっとお預けにしておきたいと思います（次回を楽しみにしてください）。

## 次回は……

今回は微分の基本的な公式をおさらいした後、回帰分析への応用として最小二乗法の基本的な計算方法を見ました。回帰分析に関しては、定数項を求めたり、説明変数が増えたりすると偏微分が必要になります。偏微分については次回のお話で見ていくこととします。

# [AI・機械学習の数学]偏微分の基本(意味と計算方法)を理解する

「偏微分」って何？ いかにも難しそうな名前だが、微分を理解していれば意外に簡単。前回までの知識を踏まえて、今回は偏微分の意味と計算方法を理解しよう。

羽山博, 著 (2020 年 07 月 14 日)

微分は関数が最小値をとる時の  $x$  の値を求めるために使えます。前回はその具体的な利用例として、最小二乗法による回帰分析を行う方法を紹介しました。しかし、取り扱った回帰式は  $y = ax$  という単純なものだけでした。そこで「偏微分」を利用し、複数の説明変数があるときにも最小二乗法が使えるようにします。つまり、重回帰分析の方法を見ていこうというわけです。

そのために今回は、偏微分の見え方と計算の方法について簡単な例で見ておくことから始めます。続けて次回、偏微分を利用して重回帰分析を行う方法を紹介します。

## 目標：偏微分の意味と計算方法を理解する

複数の変数があるような関数（多変数関数）を微分するときに、1つの変数にだけ注目し、それ以外は定数として扱うというのが偏微分です。簡単な例を示しておきます。

$$z = 3x^2 + y^2 + 4x + 7y + 5$$

のとき、

- $x$  で偏微分すると以下ようになります。

$$\frac{\partial z}{\partial x} = 6x + 4$$

- $y$  で偏微分すると以下ようになります。

$$\frac{\partial z}{\partial y} = 2y + 7$$

偏微分は高校の数学では学びませんが、微分の延長線上にある話なのでさほど難しくはありません。「 $\partial$ 」という変わった記号を使うので抵抗を感じる人もいますが、話を聞けば「なーんだ、そんな簡単なものなのか」と思うはずです。では、「 $\partial$ 」の読み方と書き方から始めましょう。

## 解説：偏微分の意味と計算方法を理解する

偏微分のために使う「 $\partial$ 」という記号は、実は「d」を丸く書いたものです。「ラウンドディー」「デル」などと読めますが、文脈から偏微分であることが明らかな場合は、単に「ディー」と読むこともあります。書き方は図1の通りです。



一般的な筆順    こう書く人もいる

図1  $\partial$ の筆順

小学生じゃあるまいし、筆順なんてどちらでもいいといえばそうなのですが、文字を見て覚えるだけでなく、一度手を動かして書いてみると、それだけでぐっと身近に感じられます。なお、たいていの日本語入力システムでは「でる」と入力して変換すれば「 $\partial$ 」が候補に表示されるはずです。

ところで、多変数関数というのはどのようなものなのでしょうか。これまでは、

$$y = ax^2 + bx + c \cdots [A]$$

の[A]のような説明変数（独立変数）が1つだけの1変数関数しか扱ってきませんでした。[A]の場合、式の右辺を見ると、

- $x$  が変数
- $a$  と  $b$  は係数
- $c$  は定数項

であることが分かります。

多変数関数とは、

$$z = ax^2 + bxy + cy + d \cdots [B]$$

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n [C]$$

の[B]や[C]のような関数です。つまり、複数の説明変数がある関数だというわけです。[B]では、

- $x$  と  $y$  が変数
- $a, b, c$  は係数
- $d$  は定数項

ですね。[C] では、

- $x_1, x_2, \dots, x_n$  が変数
- $a_1, a_2, \dots, a_n$  は係数
- $a_0$  は定数項

です \*1。



\*1 [B] の  $z$  や [C] の  $y$  も「変数」ですが、これらは「目的変数」（従属変数）と呼ばれるものでしたね。目的変数（左辺）は説明変数の値を決めれば自動的に値が決まるので、関数における変数の「個数」には数えません。以下、説明変数のことを単に「変数」と呼びます。

変数が **3つ以上** になると図で表すことは難しいのですが、変数が **2つ** の場合は図 2 のような立体的なグラフになります。ただし、変数の次数や係数などにより、形はさまざまです。

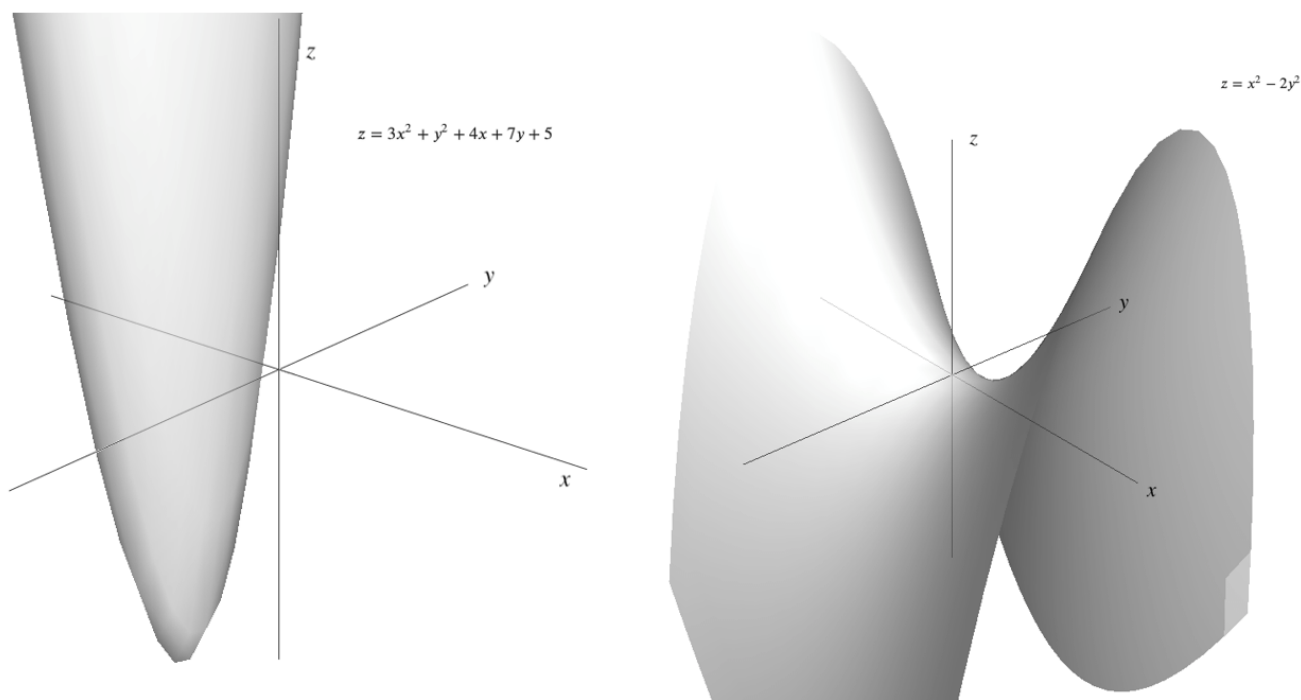


図 2 二変数関数のグラフの例

偏微分の意味は「目標」のところで触れた通り、1つの変数にだけ注目し、それ以外は定数として扱うということです。図3のように、目的変数や関数名を上書き、どの変数に注目するかを下に書きます。

$$\frac{\partial z}{\partial x}$$

図3 偏微分の書き方

読み方は「ラウンドディー・ゼット、ラウンドディー・エックス」のようになります（ちなみに、分数と同じ形をしていて、分数のように扱うこともできるので「ラウンドディー・エックス分のラウンドディー・ゼット」と読む人もいます）。これは、微分を以下のように表したときと同じですね。

$$\frac{dy}{dx}$$

もう少し例を示しておきましょう。

$$\frac{\partial f(x, y)}{\partial x} \dots \boxed{A}$$

$$\frac{\partial f}{\partial x_i} \dots \boxed{B}$$

[A] … $x$ と $y$ の2変数関数 $f$ を $x$ で偏微分する

[B] …関数 $f$ を $x_i$ で偏微分する

機械学習やディープラーニングでは、変数が多くなることが多いので、[B]のような書き方がよく使われます。また、以下のように「どの変数で微分するか」を小さく書いて表すこともあります。例えば、

$$f(x, y) = 3x^2 + 4xy + 3y^2 + 1$$

のとき、 $f(x, y)$ を $x$ で偏微分することを、

$$f_x(x, y) = (3x^2 + 4xy + 3y^2 + 1)_x$$

と表します。

というわけで、ここまでは「書き方」だけを見てきましたが、いよいよ、偏微分の計算方法を見ていくことにしましょう。例えば、

$$z = 3x^2 + y^2 + 4x + 7y + 5$$



を  $x$  で偏微分してみます。以下の図 4 を見ながら、計算方法を確認してください。

$$\begin{aligned}\frac{\partial z}{\partial x} &= (3x^2 + y^2 + 4x + 7y + 5)_x \\ &= (3x^2 + 4x + y^2 + 7y + 5)_x \\ &= 2 \cdot 3x + 4 \\ &= 6x + 4\end{aligned}$$

これは定数項として扱う

図 4 偏微分の計算方法

$x$  で偏微分するなら、まず、 $x$  の降べきの順、つまり次数の高い方から順に式を整理しておきます。次に、 $x$  以外の変数は「定数」と見なして微分します。上の例なら、 $y^2 + 7y + 5$  は定数と見なしました。つまり、 $3x^2 + 4x + \text{定数}$  を  $x$  で微分する、ということになります。今までの微分の知識で計算できますね \*2。



\*2 「偏微分可能かどうか」ということも数学的には重要ですが、機械学習やディープラーニングでは、偏微分可能な関数をモデルとして扱うのが普通なので、ここでは触れないこととします。

念のため、偏微分の図形的な意味合いも確認しておきましょう。例えば、 $z = x^2 - 3y^2$  を  $x$  で偏微分することは、 $y$  の値を固定して  $x$  で微分することなので、 $z = x^2 - 3y^2$  で表されるグラフを  $y = y_0$  ( $y_0$  は何らかの決まった値) で切った断面のグラフを微分していると考えられます。

図 5 には分かりやすい例として、グラフを回転させ  $y = 0$  での断面が見えるようにしたものを示してあります。ちょうど馬の鞍（くら）のような形のものを、馬の頭からお尻の方向に切った感じのグラフになります。グラフを回転させて見やすくしたアニメーションも用意してあるので、参考にしてください。



動画 1 2 変数関数 (グラフを回転させて見やすくしたアニメーション)



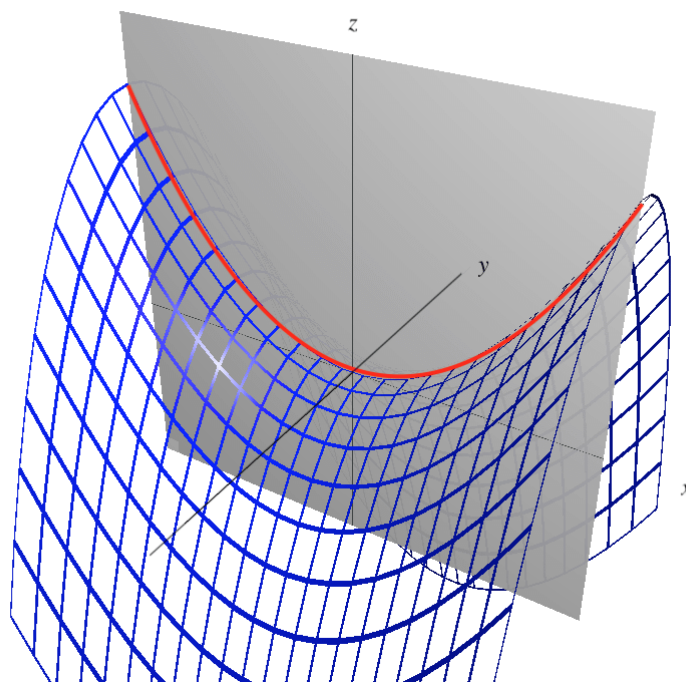


図5  $y$  の値を固定する

このとき、この断面にあるグラフの式は  $z = x^2$  になっていますね。偏微分とは、このような断面にあるグラフ（二次関数）で表されている式を微分することです。厳密な証明は省きますが、 $y$  の値が異なっても微分すれば結果は同じです。例えば、 $y = 2$  なら、 $z = x^2 - 3y^2$  は  $z = x^2 - 12$  になりますが、微分すればやはり  $z_x = 2x$  です。

逆に  $x$  の値を固定してみましょう。図6は  $x = 0$  での断面が見えるようにしたものです。こちらは、馬の鞍を輪切りにしたような感じのグラフです。

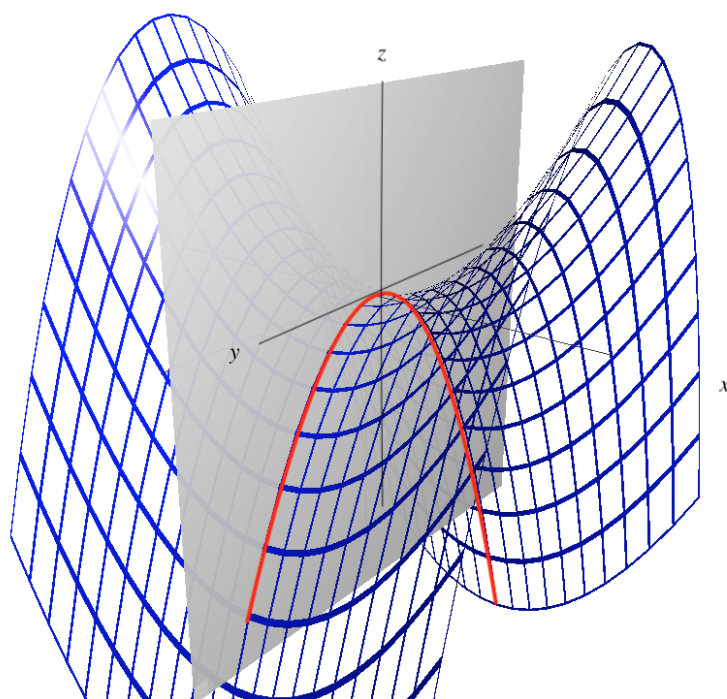


図6  $x$  の値を固定する

このとき、この断面にあるグラフの式は  $z = -3y^2$  です。やはり  $x$  の値がいくらであっても、微分すると  $z_y = -6y$  になります \*3。



\*3 全ての変数で同時に微分することを「全微分」と呼び、接平面の式を求める場合などに利用できます。が、ここでは詳細については触れません（いずれ機会があれば紹介したいと思います）。

偏微分という名前がいかに難しそうですし、「 $\partial$ 」というへんてこな記号を使いますが、全然難しくなかったですね。ただ、計算ミスをする可能性もあるので、以下の練習問題を解きながら偏微分の計算方法を確認しておきましょう。(3) はこれまでのおさらいもかねた応用問題です。ちょっと手ごわいので、後回しにしても構いません。なお、練習問題の解答については動画を用意してあります。偏微分の計算方法に慣れるためにぜひ参照してみてください。

## 練習問題



(1) 以下の関数を  $y$  で偏微分する手順を確認してみましょう

$$z = 3x^2 + y^2 + 4x + 7y + 5$$

まず、 $y$  の降べきの順に整理します。

$$z = y^2 + 7y + 3x^2 + 4x + 5$$

次に、 $y$  で偏微分します。

$$\begin{aligned}\frac{\partial z}{\partial y} &= (y^2 + 7y + \underbrace{3x^2 + 4x + 5})_y \\ &\quad \text{定数項として扱う} \\ &= 2y + 7\end{aligned}$$

(2) 以下の関数を  $x$  で偏微分する手順を確認してみましょう

$$f(x, y) = 4x^2 + 4xy + 2y^2 + 5$$

すでに  $x$  の降べきの順に整理されているので、そのまま偏微分します。

$$\begin{aligned}\frac{\partial f(x, y)}{\partial x} &= (4x^2 + 4x \underbrace{y}_{\text{定数}} + \underbrace{2y^2 + 5}_{\text{定数項として扱う}})_x \\ &= 2 \cdot 4x + 4y \\ &= 8x + 4y\end{aligned}$$

ここで注意すべき点は、 $4xy$  の部分です。 $x$  以外は定数として扱いますが「定数項」ではありません。従って、この部分を  $x$  で偏微分したときに間違えて  $y$  を消してしまい  $(4xy)_x$  を  $4$  としてしまわないように気をつけてください。 $(4xy)_x$  を  $((4y)x)_x$  と考えれば分かりやすいと思います。 $(4xy)_x$  は  $4y$  となります **\*4**。



**\*4** ここではさらっと説明していますが、偏微分でも（前回説明した）線形性が成り立ちます。つまり以下の式が成り立っています。

$$\frac{\partial}{\partial x}(a \cdot f + b \cdot g) = a \frac{\partial}{\partial x} f + b \frac{\partial}{\partial x} g \quad (a, b \text{ は定数})$$

以下のようにも書けますね。

$$(a \cdot f + b \cdot g)_x = a \cdot (f)_x + b \cdot (g)_x \quad (a, b \text{ は定数})$$

(3) 以下の関数を  $a_1$  で偏微分する手順を確認してみましょう

$$\sum (y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i}))^2$$

複雑に見えますが、手順は同じです。まず、 $a_1$  の降べきの順に整理し、定数項をまとめるために、 $a_1$  を前に移動しましょう。

$$\begin{aligned}
& \sum (y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i}))^2 \\
&= \sum (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2 \cdots \boxed{A} \\
&= \sum (-a_1 x_{1i} + y_i - a_0 - a_2 x_{2i})^2 \cdots \boxed{B} \\
&= \sum (a_1 x_{1i} - \underbrace{(y_i - a_0 - a_2 x_{2i})}_{C_i})^2 \cdots \boxed{C}
\end{aligned}$$

[A] …中のかっこをはずした

[B] …  $a_1$  の項を前に移動した

[C] …  $(-a + b)^2$  を  $(a - b)^2$  の形にした（全体が二乗されているのでプラスをマイナスにしても同じ）

[C] の  $a_1$  以外の部分を  $C_i$  と表すと以下のようなになるので、二乗を展開し、 $\Sigma$ の線形性を利用して各項に分けましょう。

$$\begin{aligned}
&= \sum (a_1 x_{1i} - C_i)^2 \\
&= \sum (a_1^2 x_{1i}^2 - 2C_i a_1 x_{1i} + C_i^2) \cdots \boxed{D} \\
&= \sum (x_{1i}^2) a_1^2 - 2 \sum (C_i x_{1i}) a_1 + \sum C_i^2 \cdots \boxed{E}
\end{aligned}$$

[D] …二乗を展開した

[E] … $\Sigma$ の線形性を利用して、 $\Sigma$ を各項に分けた

$\Sigma$ から見ると  $a_1$  には  $i$  という添字が付いていないので定数と見なせます。そこで  $\Sigma$ の外に出したというわけです。なお、 $C_i$  の中には  $i$  という添字が付いている項が含まれているので定数ではありません。従って  $n \cdot C_i$  にはできないこと注意してください（偏微分すると消えるので結果は同じなのですが）。

これでようやく  $a_1$  で偏微分できるようになりました。 $\Sigma$ の部分は定数として扱えるので、以下ようになります。最後に  $C_i$  を元に戻しておきます。

$$\begin{aligned}
& (\sum (x_{1i}^2) a_1^2 - 2 \sum (C_i x_{1i}) a_1 + \sum C_i^2)_{a_1} \\
&= 2 \sum (x_{1i}^2) a_1 - 2 \sum (C_i x_{1i}) \cdots \boxed{F} \\
&= 2 \sum (x_{1i}^2) a_1 - 2 \sum ((y_i - a_0 - a_2 x_{2i}) x_{1i}) \cdots \boxed{G}
\end{aligned}$$

[F] …  $a_1$  で偏微分した

[G] … $C_i$  を元に戻した

ちょっと複雑な結果になりましたが、できたでしょうか。ゆっくり丁寧に見ていってくださいね。

少し話は飛びますが、[E] 式は  $a_1$  についての二次式で、 $a_1$  の二乗の項の係数は正であることに気付いたでしょうか。つまり、下に凸なグラフになりますね。それを  $a_1$  で偏微分して 0 と置けば、[E] 式を最小にする  $a_1$  の値が求められます。が、ここでは両辺を 2 で割って式を簡単しておくにとどめます。

$$\begin{aligned} 2 \sum (x_{1i}^2) a_1 - 2 \sum ((y_i - a_0 - a_2 x_{2i}) x_{1i}) &= 0 \cdots [H] \\ \sum (x_{1i}^2) a_1 - \sum ((y_i - a_0 - a_2 x_{2i}) x_{1i}) &= 0 \cdots [I] \end{aligned}$$

[H] …  $a_1$  で偏微分した結果を 0 と置いた

[I] … 両辺を 2 で割った

[I] 式は、さらに以下のように整理できます。

$$\sum ((y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i})) x_{1i}) = 0$$

実は、この 3 番の問題は重回帰分析の計算の一部になっていて、 $a_0$  や  $a_1$  についても同じように計算すると、回帰式の定数項  $a_0$  や係数  $a_1, a_2$  の値が求められます。が、今回は偏微分の計算方法を身につける、ということで、次回そういった重回帰分析への応用例を見ていくこととしましょう。

## 次回は……

今回は、偏微分のかえ方と計算の方法について学びました。今回の内容と前回学んだ「説明変数が 1 つだけの回帰分析」の内容を踏まえて、次回は、「説明変数が複数ある重回帰分析」を行う方法を見ていきます。

# [AI・機械学習の数学]偏微分を応用して重回帰分析の基本を理解する

これまでに見てきた、説明変数が 1 つだけの回帰分析と偏微分の基本知識を踏まえて、複数の説明変数がある重回帰分析を行うための基本的な方法を理解しよう。

羽山博, 著 (2020 年 07 月 28 日)

前々回は「説明変数が 1 つだけの回帰分析」を、前回は偏微分の考え方と計算の方法について学びました。それらの内容を踏まえて、今回は偏微分の応用編として「説明変数が複数ある重回帰分析」を行う方法を見ていきます。

## 目標：偏微分を利用して重回帰分析を行う

重回帰分析の回帰式の例は以下のようなものでした。いくつかの値を基に、このような回帰式の定数項と係数を求めようというのがここでの目標です。

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

回帰式の求め方は前々回の例と同様で、観測値（実際に得られたデータ）と理論値（回帰式で求めた値）との差の二乗和が最小になるように定数項や係数を決めるという方法です。

## 解説：偏微分を利用して重回帰分析を行う

まずは、具体的な例で考えてみましょう。図 1 のような不動産データがあったとします。このデータを基に回帰式の定数項と係数を求めてみたいと思います。実はこの例であれば Excel でもできるので、ついでに Excel で回帰式を求めた結果も示してあります。

	A	B	C	D	E	F	G	H
1	No.	築年数	面積	家賃		面積	築年数	定数項
2	1	4	37	11		0.2478	-0.0243	1.9292
3	2	16	18	6				
4	3	24	43	12				

図 1 不動産データと重回帰分析の例

[A] 列から [D] 列のデータが不動産データです。このデータを基に回帰式を求めた結果が [F] 列から [D] 列になります。ちょっと注意しないといけないのは、係数の順序が元のデータと異なるということです。回帰式は、以下のようになります。

$$\text{家賃} = 1.9292 - 0.0243 \times \text{築年数} + 0.2478 \times \text{面積}$$



築年数が大きくなるとやや家賃が下がり、面積が大きくなると家賃が上がるということが分かりますね。定数項は面積も築年数も **0** の物件の家賃ということになるので、違和感があるかもしれませんが、不動産業務のための固定費と考えてもいいでしょう（[前々回](#)の「説明変数が1つだけの回帰分析」では、そういう場合は **0** としました）。

家賃を  $y$ 、築年数を  $x_1$ 、面積を  $x_2$  とすると、以下のように表せます。**1.9292** は定数項ですね。

$$y = 1.9292 - 0.0243x_1 + 0.2478x_2$$

説明変数が2つあるので、図にするとちょっと見つらいですが、図2のようなイメージです。回帰式は直線ではなく平面を表します。誤差がプラスになるのかマイナスになるのかはこの時点では分からないので、とりあえずイメージだけつかんでおいてもらうといいでしょう **\*1**。



**\*1** 平均や分散などを求めるときには、実際の測定値と推定値（理論値）との差を「誤差」と呼びますが、回帰分析などでは、実際の測定値と予測値（理論値）の差のことを「残差」と呼ぶのが一般的です。以降、測定値と回帰式で求めた値の差を「残差」と呼ぶことにします。例えば、そのような差の二乗和は、誤差二乗和ではなく「**残差二乗和**」（あるいは「**残差平方和**」）と呼びます。ちなみに、統計学の回帰分析とは異なり、機械学習のニューラルネットワークでは、正解値（学習用に正解としてラベル付けされた値）と推定値（理論値）との差を「誤差」と呼び、その差の二乗和を **2** で割ったものを「二乗和誤差」と呼ぶのが一般的なので、統計学の用語と機械学習の用語で混同しないよう注意しましょう。

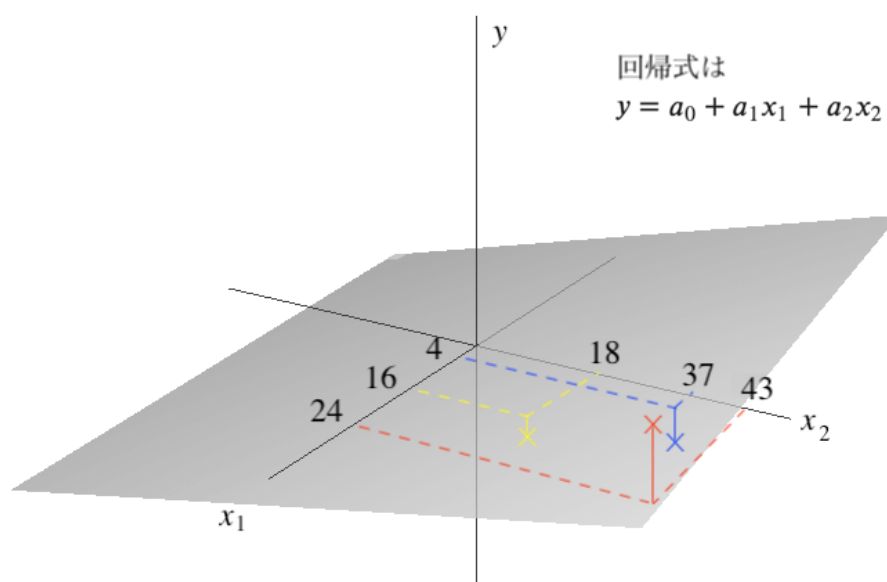


図2 回帰式で表される平面の例

この時点では、まだ定数項や係数は分からない（これから求める）ので、回帰式は以下ようになります。

$$y = a_0 + a_1x_1 + a_2x_2$$

## 残差二乗和を求める

私たちがやりたいことは、築年数と面積の各点の最も近くを通る回帰式（平面の式）を求めるということです。そこで、各データと回帰式で求めた値の差（残差）をそれぞれ求めてみましょう。ここからの説明については動画も用意してあります。ちょっと難しいな、と感じたらぜひ視聴してみてください。



動画 1 残差平方和の算出

1 件目のデータは、 $x_1$ （築年数）が 4、 $x_2$ （面積）が 37、 $y$ （家賃）が 11 なので、残差は、

$$\underbrace{11}_{\text{観測値}} - \underbrace{(a_0 + 4a_1 + 37a_2)}_{\text{回帰式による予測値}}$$

となります。同様に、2 件目のデータと 3 件目のデータの残差も以下のようになります。

$$\begin{aligned} \underbrace{6}_{\text{観測値}} &- \underbrace{(a_0 + 16a_1 + 18a_2)}_{\text{回帰式による予測値}} \\ \underbrace{12}_{\text{観測値}} &- \underbrace{(a_0 + 24a_1 + 43a_2)}_{\text{回帰式による予測値}} \end{aligned}$$

回帰式を求めるためには最小二乗法を利用します。前々回も説明しましたが、再度、説明すると、**最小二乗法**というのは、これらの残差の二乗の総和（残差二乗和）が最小になるように、定数項（ $a_0$ ）や係数（ $a_1$ 、 $a_2$ ）の値を決めるということでしたね。

では、上記の残差を全て二乗して足してみましょう。どのような式になるでしょうか。なお、いちいち残差二乗和と呼ぶのが面倒なので、残差二乗和を  $Q$  と表すことにします（特に  $Q$  という文字に意味はありません）。



$$Q = (11 - (a_0 + 4a_1 + 37a_2))^2 \\ + (6 - (a_0 + 16a_1 + 18a_2))^2 \\ + (12 - (a_0 + 24a_1 + 43a_2))^2$$

何だかものすごく複雑な式になっているように見えますが、この  $Q$  式を最小にする  $a_0$ 、 $a_1$ 、 $a_2$  を求めるということなので、まずは  $a_0$  で偏微分するとよさそうです。次に  $a_1$  で偏微分、最後に  $a_2$  で偏微分するというわけですね。しかし、ここでは、あえて  $a_1$  での偏微分を先にやることにします。別に  $a_0$  から順にやってもいいのですが、 $a_1$  での偏微分の方が、式の特徴がよく分かるという、ただそれだけの理由です。

$a_1$  で偏微分するということは、他の変数は定数として扱えるというわけですから、 $a_1$  で整理し、他の変数や定数を全て定数にしてみましょう。

$$Q = (-4a_1 + \underbrace{11 - a_0 - 37a_2}_{C_1 \text{と置く}})^2 \\ + (-16a_1 + \underbrace{6 - a_0 - 18a_2}_{C_2 \text{と置く}})^2 \\ + (-24a_1 + \underbrace{12 - a_0 - 43a_2}_{C_3 \text{と置く}})^2 \cdots [A] \\ = (-4a_1 + C_1)^2 + (-16a_1 + C_2)^2 + (-24a_1 + C_3)^2 \cdots [B]$$

[A]  $\cdots a_0$  について整理した（項の順序を変えただけ）

[B]  $\cdots a_0$  以外の部分を  $C_1$ 、 $C_2$ 、 $C_3$  と置いた

例えば、 $C_1 = 11 - a_0 + 37a_2$  と置くといった具合です。これらは定数として扱うといっても、同じ値ではないので全て同じ  $C$  にははいけません。そこで、 $C_1$ 、 $C_2$ 、 $C_3$  と置いたわけです。

さらに [B] 式を展開していきましょう。二乗の展開公式を使ってかっこを展開し、 $a_1$  の降べきの順に並べます \*2。

$$Q = (-4a_1)^2 - 2 \cdot 4C_1a_1 + C_1^2 \\ + (-16a_1)^2 - 2 \cdot 16C_2a_1 + C_2^2 \\ + (-24a_1)^2 - 2 \cdot 24C_3a_1 + C_3^2 \cdots [C] \\ = (4^2 + 16^2 + 24^2)a_1^2 \\ - 2(4C_1 + 16C_2 + 24C_3)a_1 \\ + (C_1^2 + C_2^2 + C_3^2) \cdots [D] \\ = 848a_1^2 - 2(4C_1 + 16C_2 + 24C_3)a_1 + \sum C_i^2 \cdots [E]$$

[C]  $\cdots$  二乗の展開公式を使ってかっこを展開した

[D]  $\cdots a_1$  の降べきの順に並べた。 $(-a_1)^2 = a_1^2$  なのでマイナスを取った

[E]  $\cdots C_1 + C_2 + C_3$  を  $\sum C_i$  のように表した。

**\*2** [C] 式から [D] 式への変形では、次のような計算間違いをしやすいので注意してください。実は、全て文字を使ったり  $\Sigma$  などの記号を使ったりすると間違えにくいのですが、具体的な数字を使って計算する場合には、筆者もうっかりと、してかしてしまうことがあります。文字や記号を使う意味はそういうところにもあるわけですね。



$$\begin{aligned} & (-4a_1)^2 + (-16a_1)^2 + (-24a_1)^2 \\ &= \underbrace{((-4) + (-16) + (-24))^2 a_1^2}_{\text{間違い}} \cdots \boxed{1} \end{aligned}$$

$$\begin{aligned} &= (-44)^2 a_1^2 \\ &= 1936 a_1^2 \cdots \times \end{aligned}$$

$$\begin{aligned} & (-2 \cdot 4C_1 a_1) + (-2 \cdot 16C_2 a_1) + (-2 \cdot 24C_3 a_1) \\ &= -2 \underbrace{(4 + 16 + 24)(C_1 + C_2 + C_3) a_1}_{\text{間違い}} \cdots \boxed{2} \end{aligned}$$

$$\begin{aligned} &= -2 \cdot 44(C_1 + C_2 + C_3) a_1 \\ &= -88(C_1 + C_2 + C_3) a_1 \cdots \times \end{aligned}$$

[1] 式では二乗してから足す必要があるのに、足してから二乗しています。

[2] 式では各項に掛けてから足す必要があるのに、足してから各項に掛けています。 $C_1$ 、 $C_2$ 、 $C_3$  が異なるものであることに注意しましょう（同じものであれば単純にくくり出せます）。

正しい計算は以下の通りです。

$$\begin{aligned} & (-4a_1)^2 + (-16a_1)^2 + (-24a_1)^2 \\ &= ((-4)^2 a_1^2) + ((-16)^2 a_1^2) + ((-24)^2 a_1^2) \\ &= (4^2 + 16^2 + 24^2) a_1^2 \\ &= 848 a_1^2 \end{aligned}$$

$$\begin{aligned} & -2 \cdot 4C_1 a_1 - 2 \cdot 16C_2 a_1 - 2 \cdot 24C_3 a_1 \\ &= -2(4C_1 + 16C_2 + 24C_3) a_1 \end{aligned}$$

次のページは……

では、いよいよ、次のページで  $Q$  を  $a_1$  で偏微分してみましょう。

## 残差二乗和を偏微分する

では、いよいよ、 $Q$  を  $a_1$  で偏微分してみましょう。また動画での解説も用意してあります。ぜひ視聴してみてください。



$$\begin{aligned}\frac{\partial Q}{\partial a_1} &= (848a_1^2 - \underbrace{2(4C_1 + 16C_2 + 24C_3)}_{\text{定数}})a_1 + \underbrace{\sum C_i^2}_{\text{定数項として扱う}})_{a_1} \\ &= 2 \cdot 848a_1 - 2(4C_1 + 16C_2 + 24C_3) \cdots [F]\end{aligned}$$

[F]  $\cdots a_1$  で偏微分した。定数項は消える

[F] 式は二次式で、 $a_1$  の二乗の係数は正なので、下に凸な関数ですね。従って、偏微分した [F] 式を 0 と置けば残差二乗和の式  $Q$  を最小にする  $a_1$  の値が求められるはずです。

$$2 \cdot 848a_1 - 2(4C_1 + 16C_2 + 24C_3) = 0$$

両辺を 2 で割ってから、 $C_i$  を元に戻すと以下ようになります。かっこを外すときに、かっこの前にマイナスがあることに注意してください。  $-(-a + b) = a - b$  のように、かっこの中のマイナスはプラスに、プラスはマイナスになります。

$$\begin{aligned}
848a_1 - (4C_1 + 16C_2 + 24C_3) &= 0 \cdots \boxed{G} \\
848a_1 - \underbrace{4(11 - a_0 - 37a_2)}_{C_1} \\
&\quad + 16\underbrace{(6 - a_0 - 18a_2)}_{C_2} \\
&\quad + 24\underbrace{(12 - a_0 - 43a_2)}_{C_3} = 0 \cdots \boxed{H} \\
848a_1 + (4 + 16 + 24)a_0 \\
&\quad + (4 \cdot 37 + 16 \cdot 18 + 24 \cdot 43)a_2 \\
&\quad - (4 \cdot 11 + 16 \cdot 6 + 24 \cdot 12) = 0 \cdots \boxed{I} \\
848a_1 + (4 + 16 + 24)a_0 \\
&\quad + (4 \cdot 37 + 16 \cdot 18 + 24 \cdot 43)a_2 \\
&\quad = 4 \cdot 11 + 16 \cdot 6 + 24 \cdot 12 \cdots \boxed{J}
\end{aligned}$$

[G] …両辺を **2** で割った。さらに両辺を **4** で割ることもできるが、後で仕組みを見たいので、この段階ではこのままにしておく

[H] …  $C_1$  を元に戻した

[I] …  $a_0$  と  $a_2$  で整理し、その順に項を並べ替えた

[J] …定数項を右辺に移項した

あとは [J] 式の数字を計算するだけです。結局、以下のような式が得られます。 $\therefore$  という記号は「ゆえに」「従って」という意味です。

$$\begin{aligned}
44a_0 + 848a_1 + 1468a_2 &= 428 \\
\therefore 11a_0 + 212a_1 + 367a_2 &= 107 \cdots \boxed{K}
\end{aligned}$$

[K] … $a_0$  から順に並べて、係数と定数項を計算した。両辺を **4** で割った。

はい、だいぶ簡単になりましたね。残差二乗和を最小にするような  $a_1$  の値は [K] 式を解けば求められるということです。ところが、[K] 式には  $a_0$  や  $a_2$  という未知数が含まれていますね。なので、これだけでは答えは求められません。そこで、 $a_0$  や  $a_2$  についても同じように偏微分して 0 と置く必要があります。……というのはまた大変ですね。そこで、ちょっとズルをします。[K] 式の前の段階の [J] 式で係数と定数項の特徴について、順を追って見てみましょう。

図 3 ( $a_1$  で微分する例) を見ると、以下のようになっていることが分かります。

### ステップ1： 偏微分する変数の係数を求める

$$Q = (11 - (a_0 + 4a_1 + 37a_2))^2 + (6 - (a_0 + 16a_1 + 18a_2))^2 + (12 - (a_0 + 24a_1 + 43a_2))^2 \dots\dots \text{残差二乗和}$$

①二乗和

$$848a_1 + (4 + 16 + 24)a_0 + (4 \cdot 37 + 16 \cdot 18 + 24 \cdot 43)a_2$$

$$4^2 + 16^2 + 24^2$$

図3 係数の特徴（偏微分する変数）

ステップ1： 偏微分する変数の係数は、(1) 元の式の係数の二乗和とする（図3）。例えば、 $a_1$  で偏微分する  
なら、元の式の  $a_1$  の係数である **4**、**16**、**24** の二乗和が、偏微分した後の  $a_1$  の係数になる。

### ステップ2： 他の変数の係数を求める（ $a_2$ で説明する）

$$Q = (11 - (a_0 + 4a_1 + 37a_2))^2 + (6 - (a_0 + 16a_1 + 18a_2))^2 + (12 - (a_0 + 24a_1 + 43a_2))^2$$

②偏微分する変数の係数と  
その変数の係数を掛けて合計

左辺

$$848a_1 + (4 + 16 + 24)a_0 + (4 \cdot 37 + 16 \cdot 18 + 24 \cdot 43)a_2$$

※  $a_0$  の係数は全て1なので  
( $4 \cdot 1 + 16 \cdot 1 + 24 \cdot 1$ ) と考えればよい

図4 係数の特徴（他の変数）

ステップ2： それ以外の係数は、(2) 偏微分する変数の係数と元の式の係数を掛けたものの合計とする（図  
4）。例えば、 $a_2$  の係数は、元の式の  $a_1$  の係数である **4**、**16**、**24** と、元の式の  $a_2$  の係数である **37**、**18**、**43**  
をそれぞれ掛けたものの合計になる。

### ステップ3：定数項を求める（定数項は右辺に書く）

右辺

$$Q = (11 - (a_0 + 4a_1 + 37a_2))^2 + (6 - (a_0 + 16a_1 + 18a_2))^2 + (12 - (a_0 + 24a_1 + 43a_2))^2$$

③偏微分する変数の係数と  $y$  の値を掛けて合計（これが定数項になる）

$$4 \cdot 11 + 16 \cdot 6 + 24 \cdot 12$$

図5 係数の特徴（定数項）

ステップ3：定数項もステップ2と同じ考え方で求める。(3) 偏微分する変数の係数と  $y$ （観測値）の値を掛けたものの合計とする。ただし、定数項は右辺に書く（図5）。例えば、元の式の  $a_1$  の係数である 4、16、24 と、元の式の  $y$  の値（家賃）である 11、6、12 をそれぞれ掛けたものの合計になる。

これらのルールを  $a_0$  や  $a_1$  に対して適用すれば、それぞれの式が立てられそうですね（実際に偏微分して 0 と置くという計算をすると、これらの式が得られます）。 $a_0$  について偏微分して 0 と置いたときの式は以下の通りです。

$$\begin{aligned} & (1^2 + 1^2 + 1^2)a_0 \\ & + (1 \cdot 4 + 1 \cdot 16 + 1 \cdot 24)a_1 \\ & + (1 \cdot 37 + 1 \cdot 18 + 1 \cdot 43)a_2 = (1 \cdot 11 + 1 \cdot 6 + 1 \cdot 12) \\ \therefore 3a_0 + 44a_1 + 98a_2 = 29 \cdots [L] \end{aligned}$$

[L] …  $a_0$  で偏微分して 0 と置いた場合。 $a_0$  の係数は全て 1 なので二乗和は 3

$a_2$  について偏微分して 0 と置いた場合は以下のような式になります。

$$\begin{aligned} & (37 \cdot 1 + 18 \cdot 1 + 43 \cdot 1)a_0 \\ & + (37 \cdot 4 + 18 \cdot 16 + 43 \cdot 24)a_1 \\ & + (37^2 + 18^2 + 43^2)a_2 = (37 \cdot 11 + 18 \cdot 6 + 43 \cdot 12) \\ \therefore 98a_0 + 1468a_1 + 3542a_2 = 1031 \cdots [M] \end{aligned}$$

[M] …  $a_2$  で偏微分して 0 と置いた場合。 $a_1$  の係数は 37、18、43 だったので、それらを各項の係数に掛ける。37、18、43 の二乗和は 3542



以上で、3つの式が出そろいました。これらの式は連立方程式になっていますね。この3つの式が同時に成り立つのであれば、残差二乗和（Q）の式が最小になるといえます。そのときの  $a_0$ 、 $a_1$ 、 $a_2$  の値を求めるために、この連立方程式を解くことにしましょう。なお、機械学習では、偏微分して0と置いて連立方程式を立て、それを解くという方法を採用せずに、偏微分係数を使って少しずつ値を変えていく方法を使います。

そういうわけで、連立方程式の解き方にあまり興味がないという人は、概略だけですが、機械学習での係数の求め方も示しておいたので、[そこまで読み飛ばして](#)もらっても結構です \*3。



\*3 連立方程式は、[Wolfram Alpha](#) などの[サイト](#)を使って解くこともできます。

## 連立方程式を解く

偏微分して0と置いた式をもう一度まとめて掲載しておきます。あまりに長い道のりで、使えるアルファベットがなくなってきたので、式の番号を [A] から振り直しますね。

$$\begin{aligned} 3a_0 + 44a_1 + 98a_2 &= 29 \cdots [A] \\ 11a_0 + 212a_1 + 367a_2 &= 107 \cdots [B] \\ 98a_0 + 1468a_1 + 3542a_2 &= 1031 \cdots [C] \end{aligned}$$

[A] 式は [L] 式を再掲したもの、[B] 式は [K] 式を再掲したもの、[C] 式は [M] 式を再掲したものです。これらの連立方程式を解けば、 $a_0$ 、 $a_1$ 、 $a_2$  が求められますね。

まず、[A] 式を使って [B] 式と [C] 式から  $a_0$  を消去しましょう。[B] 式を3倍したものから [A] 式を11倍したものを引けば [B] 式から  $a_0$  が消せます。

$$\begin{aligned} & 3 \cdot 11a_0 + 3 \cdot 212a_1 + 3 \cdot 367a_2 = 3 \cdot 107 \\ - & 11 \cdot 3a_0 + 11 \cdot 44a_1 + 11 \cdot 98a_2 = 11 \cdot 29 \\ \hline & (3 \cdot 212 - 11 \cdot 44)a_1 + (3 \cdot 367 - 11 \cdot 98)a_2 = 3 \cdot 107 - 11 \cdot 29 \\ & \therefore 152a_1 + 23a_2 = 2 \cdots [D] \end{aligned}$$

同様に、[C] 式を3倍したものから [A] 式を98倍したものを引けば [C] 式から  $a_0$  が消せます。

$$\begin{aligned} & 3 \cdot 98a_0 + 3 \cdot 1468a_1 + 3 \cdot 3542a_2 = 3 \cdot 1038 \\ - & 98 \cdot 3a_0 + 98 \cdot 44a_1 + 98 \cdot 98a_2 = 98 \cdot 29 \\ \hline & (3 \cdot 1468 - 98 \cdot 44)a_1 + (3 \cdot 3542 - 98 \cdot 98)a_2 = 3 \cdot 1031 - 98 \cdot 29 \\ & \therefore 92a_1 + 1022a_2 = 251 \cdots [E] \end{aligned}$$

次は、[D] 式を使って [E] 式から  $a_1$  を消します。[E] 式を 152 倍したものと [D] 式を 92 倍したものを引けば、 $a_1$  が消えます

$$\begin{aligned} & 152 \cdot 92a_1 + 152 \cdot 1022a_2 = 152 \cdot 251 \\ - & \underline{92 \cdot 152a_1 + 92 \cdot 23a_2 = 92 \cdot 2} \\ & (1525 \cdot 1022 - 92 \cdot 23)a_2 = 152 \cdot 251 - 92 \cdot 2 \\ & \therefore 153288a_2 = 37968 \\ & \therefore a_2 = \frac{524214}{129081} = \frac{28}{113} \simeq 0.2478 \dots \quad [F] \end{aligned}$$

$a_2$  の値が求められました。0.2478 です。

次に、[D] 式に  $a_2$  の値を代入して、 $a_1$  の値を求めます。 $a_2$  の値は、

$$[F] \text{ 式の } \frac{151872}{612912} \text{ を約分した } \frac{28}{113}$$

を使うのがいいでしょう（0.2478 は小数点以下を四捨五入した近似値なので）。

$$\begin{aligned} & 152a_1 + 23 \frac{28}{113} = 2 \\ a_1 &= \left( 2 - \frac{23 \cdot 28}{113} \right) \div 152 \\ &= \frac{2 \cdot 113 - 23 \cdot 28}{113 \cdot 152} = \frac{-418}{17176} = \frac{-11}{452} \simeq -0.0243 \dots \quad [G] \end{aligned}$$

$a_1$  の値も求められました。-0.0243 です。最後に、 $a_1$  と  $a_2$  を [A] 式に代入して  $a_0$  を求めます。

$$\begin{aligned} & 3a_0 + 44 \frac{-11}{452} + 98 \frac{28}{113} = 29 \\ a_0 &= \left( 29 - \frac{44 \cdot (-11)}{452} - \frac{98 \cdot 28}{113} \right) \div 3 \\ &= \left( 29 - \frac{11 \cdot (-11)}{113} - \frac{98 \cdot 28}{113} \right) \div 3 \\ &= \frac{29 \cdot 113 - 11 \cdot (-11) - 98 \cdot 28}{113 \cdot 3} \\ &= \frac{654}{113 \cdot 3} = \frac{218}{113} = 1.9292 \dots \quad [H] \end{aligned}$$



ようやく答えにたどり着きました。[F] [G] [H] より、

$$\begin{aligned}a_0 &= 1.9292 \\a_1 &= -0.0243 \\a_2 &= 0.2478\end{aligned}$$

となり、回帰式は、

$$y = 1.9292 - 0.0243x_1 + 0.2478x_2$$

となりました。これは最初に見た結果とちゃんと一致していますね。この式を使えば、さまざまな「築年数 ( $x_1$ )」と「面積 ( $x_2$ )」から「家賃 ( $y$ )」が予測できるというわけです。

最小二乗法による重回帰分析の手順は、

1. 観測値と予測値の差（残差）の二乗和を求める
2. 各係数を変数と見なして偏微分する
3. それによって導き出された連立方程式を解いて各係数を求める

というものでした。しかし、連立方程式は変数が多くなると、コンピュータを使って解くための計算量が増えるので、かなり時間がかかります。

そこで、機械学習では、ここで見たように連立方程式を解いて答えを導き出すのではなく、偏微分した式に適当な値を入れて、微分係数の符号を見るという方法を使うのが一般的です。微分係数の符号が正であれば、最小値はもっと左にあるはずで、つまり、もう少し小さな  $x_i$  を与えてあげれば最小値に近づきます。逆に、微分係数の符号が負であれば、最小値はもっと右にあるはずで、この場合は、もう少し大きな  $x_i$  を与えてあげれば最小値に近づきます。つまり、微分係数の符号と反対の方向に動かせばいい、というわけです（図 6）。

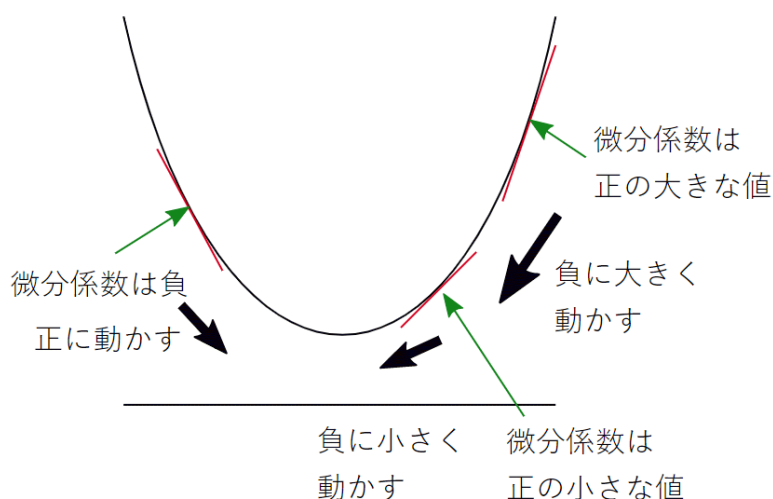


図 6 最小値を見つける

図の右側を見てください。最小値から遠く離れていると、微分係数は正の大きな値になるので、それを引いてあげればぐっと最小値に近づきますね。最小値に近いと、微分係数も小さい値なので、それを引いてあげると少しずつ最小値に近づきます。図の左側は、微分係数が負の場合です。その場合は微分係数を引く（負の値を引くので、足すことに絶対値を足すことになります）と最小値に近づきます。

ただし、微分係数をそのまま足したり引いたりするのではなく、微分係数に小さな値を掛けて使います。例えば、勾配降下法などでは、直前の  $x_i$  から「適当に選んだ小さな値 × 微分係数」を引くという計算を繰り返します。そして、値がほとんど変化しなくなったとき（＝収束したとき）の  $a_i$  を求めるというわけです。この「適当に選んだ小さな値」のことを「学習率」などと呼び、 $\eta$ （イータと読みます）などの文字を使って表します。学習率の値を大きく設定すると、正と負の間を行ったり来たりして収束しなくなることがあるので注意が必要です。また、学習率の値が小さすぎると収束に時間がかかります。

## まとめと補足

最後にまとめと補足です。

### 残差二乗和の偏微分（一般的な場合）

ここまでは、具体的な数値を使って偏微分を利用した回帰分析の例を見てきました。これを一般化するとどうなるかということですが、実は、[前回の練習問題の3番](#)がその計算の一部になっています。そういうわけで、3番を後回しにした人もいったん戻ってぜひチャレンジしてみてください。答えは、以下の通りでした。

回帰式が、

$$y = a_0 + a_1x_1 + a_2x_2$$

のとき、いくつかのデータを基に求めた残差二乗和の式は以下のようになります。 $y_i$  や  $x_{1i}$  などがそれぞれのデータを表します **\*4**。

$$\sum (y_i - (a_0 + a_1x_{1i} + a_2x_{2i}))^2$$



**\*4** 変数を区別するための添字と、個々のデータを区別するための添字が紛らわしくなるので、個々のデータを区別するための添字を右肩にかっこ書きで示すこともよくあります。上の例であれば、以下のようになります。

$$\sum (y^{(i)} - (a_0 + a_1x_1^{(i)} + a_2x_2^{(i)}))^2$$

これを  $a_1$  で偏微分すると以下ようになります。

$$\begin{aligned} & \frac{\partial}{\partial a_1} \sum (y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i}))^2 \\ &= 2 \sum (x_{1i}^2) a_1 - 2 \sum ((y_i - a_0 - a_2 x_{2i}) x_{1i}) \end{aligned}$$

$a_0$  や  $a_2$  で偏微分する場合でも同じですし、回帰式の変数が増えただけでも同じです。なお、 $a_0$  を  $a_0 x_0$  と表し、 $x_0$  が常に 1 であると考えれば、以下のように書けます（形式をそろえただけです）。

$$\begin{aligned} & \frac{\partial}{\partial a_1} \sum (y_i - (a_0 x_{0i} + a_1 x_{1i} + a_2 x_{2i}))^2 \\ &= 2 \underbrace{\sum (x_{1i}^2) a_1}_{a_1 \text{ の係数}} - 2 \underbrace{\sum ((y_i - a_0 x_{0i} - a_2 x_{2i}) x_{1i})}_{a_1 \text{ 以外全て}} \end{aligned}$$

つまり、すでに見てきた通り、

1. 偏微分する変数（例えば  $a_1$ ）については、係数の二乗和を係数とする
2. それ以外の項（例えば  $y_i$  や  $a_0$ 、 $a_2$ ）については、偏微分する変数の係数（例えば  $x_{1i}$ ）とそれぞれの変数の係数（例えば  $y_i$  そのものや  $x_{0i}$ 、 $x_{2i}$ ）を掛けたものの合計を係数とする

というわけです。

さらに、回帰式の変数が  $x_0$  から  $x_m$  までであるものとし、それぞれのデータの個数が  $n$  個 ( $1 \sim n$  まで) であるとき、残差二乗和の式を  $a_k$  について偏微分する場合の一般的な式を書くことで以下ようになります。ただ、以下の式は、全てを一般化し、しかもできるだけ簡潔に書いたものなので、ぱっと見て「理解できた」というわけにはいかないかもしれません。あくまで参考用のものなので、今はあまり気にしなくても大丈夫です。具体的なイメージとして、不動産データの回帰分析の方法が理解できていれば十分です。

回帰式が、

$$y = a_0 x_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \quad (\text{ただし、} x_0 = 1)$$

のとき、残差二乗和は以下ようになります。

$$\begin{aligned} & \sum_{i=1}^n (y_i - \underbrace{(a_0 x_{0i} + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_m x_{mi})}_{\text{この部分は } \sum \text{ を使って書ける}})^2 \\ &= \sum_{i=1}^n (y_i - \underbrace{\sum_{j=0}^m a_j x_{ji}}_{\sum \text{ を使って書き換えただけ}})^2 \cdots \boxed{A} \end{aligned}$$

[A] …残差二乗和の式の中の、 $a_0x_{0i} \sim a_mx_{mi}$  の和を $\Sigma$ で書いた

これを  $a_k$  で偏微分すると以下ようになります（式の変形は前回の練習問題の3番と同じなので、結果だけを示しています）。

$$\begin{aligned} & \frac{\partial}{\partial a_k} \sum_{i=1}^n (y_i - \sum_{j=0}^m a_j x_{ji})^2 \\ &= 2 \underbrace{\sum_{i=1}^n (x_{ki}^2)}_{a_k \text{の係数の二乗和}} a_k \\ &\quad - 2 \left( \sum_{i=1}^n \left( \underbrace{y_i}_{y_i \text{や}} - \sum_{j=0, j \neq k}^m \underbrace{(a_j x_{ji})}_{\text{各項の係数に}} \right) \underbrace{x_{ki}}_{a_k \text{の係数を掛ける}} \right) \\ &= 2 \left( \sum_{i=1}^n \left( \left( \sum_{j=0}^m (a_j x_{ji}) - y_i \right) x_{ki} \right) \right) \cdots \boxed{B} \end{aligned}$$

ちょっとやりすぎな感じかもしれませんが、これが一般的な形式になるというわけです。**\*5**。

**\*5** ごく単純な計算であれば、具体的な数値を使った方がイメージをつかみやすいのですが、変数が増えてくると、むしろ文字や記号を使った方が式の意味や計算の手順がよく分かりますし、計算を間違うことも少なくなりますね。



## 次回は……

というわけで、今回は最小二乗法による重回帰分析に偏微分を利用するお話を見てきました。偏微分は回帰分析だけでなくニューラルネットワークなどでも使われるので、機械学習やディープラーニングには必須の数学です。

といっても、偏微分の計算そのものは簡単でしたね。難しく思えるのは係数や変数、重みなどにさまざまな文字を使ったり、 $\Sigma$ などを使って一般的に式を記述したりすることがあるからです。そういうときには、できるだけ簡単な例で具体的な数値を使って試しに計算してみると少しずつ意味が分かってくると思います。

次回は、微分の積の公式や合成関数の微分（連鎖律）などについて見ていきたいと思います。

# [AI・機械学習の数学] 合成関数の微分(連鎖律)とニューラルネットワーク初歩の初歩

微分法は回帰分析だけでなく、機械学習のさまざまなタスクで使われる。特に、合成関数の微分（連鎖律）はニューラルネットワークの学習において必須となる。今回はそのための第一歩として、合成関数がどのようなものであるかを見た後、合成関数の微分法の公式とその計算方法を紹介する。

羽山博, 著 (2020 年 09 月 03 日)

微分法の基本については、その考え方や計算方法、多変数関数の偏微分など、これまでの連載で一通り見てきました。今回は、ニューラルネットワークの学習に使われる合成関数の微分（連鎖律）について、考え方と計算方法を説明します。ただし、ニューラルネットワークそのものについては、線形代数の知識も必要になってくるので、今回は詳しくは触れません。しかし、合成関数の微分ができるようになれば、ニューラルネットワークの学習を理解するにあたっての、はじめの一步が踏み出せます。

まずは、合成関数がどのようなものであるかというところから見ていきましょう。

## 目標【その 1】：合成関数を理解する

ある関数  $f(x)$  に対して、さらに別の関数  $g$  を適用する場合、関数  $g$  は以下のように表されます。

$$g(f(x))$$

この  $g(f(x))$  のように、ある関数を適用した結果に別の関数を適用したものを**合成関数**と呼びます。なお、以下のように、関数  $f$  と関数  $g$  の合成関数がある種の演算として考え、小さな  $\circ$  を使って表すこともあります（この連載では  $\circ$  を使った書き方は使いませんが、合成関数の交換法則が成り立たないことや結合法則が成り立つことなども簡潔に表せます）。

$$g \circ f$$

小さな  $\circ$  は、英語では「circle」や「of」などいろんな読み方がありますが、日本語では普通に「まる」と読む人が多いようです。

## 解説【その1】：合成関数を理解する

合成関数のイメージは、以下の図を見れば一目瞭然だと思いますが、最初が肝心なので丁寧に見ていきましょう。図1は、 $x$ に対して関数 $f$ を適用し、その結果に対して関数 $g$ を適用した様子です。それを数式で書くと、

$$g(f(x))$$

となります。

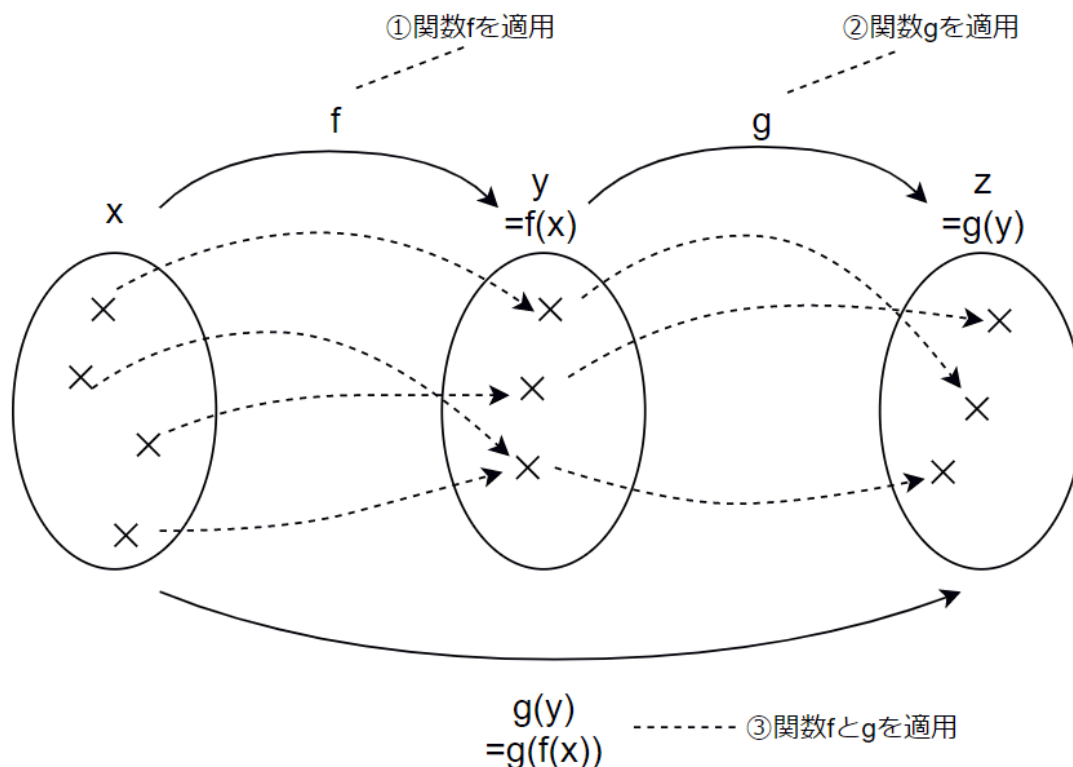


図1 合成関数のイメージ

$x$ の値に対して関数 $f$ を適用することは $f(x)$ と書けますね。これを $y$ と置きましょう。つまり、 $y = f(x)$ です。これはすでに何度も登場している関数にすぎません。次に、そうして得られた $y$ の値に対して関数 $g$ を適用することは $g(y)$ と書けます。これを $z$ と置きます。つまり、 $z = g(y)$ です。

ここで、 $g(y)$ に $y = f(x)$ を代入すると、 $g(f(x))$ になりますね（図2）。

$$\begin{array}{l} y = f(x) \\ \quad \searrow \text{代入する} \Rightarrow z = g(f(x)) \\ z = g(y) \end{array}$$

図2 合成関数を作る

上のようにわざわざ  $y$  とか  $z$  という文字を使わなくても簡単に理解できると思いますが（むしろ回りくどいかもしれませんね）、文字で表した方が数式が簡単になることもあるので、念のため、ということで。

では、具体的な例で合成関数の計算を幾つかやってみましょう。最初は、 $f(x)$  とか  $g(y)$  を使わずに書いてみます。

## 合成関数の例 1

---

$$y = 2x + 4 \quad (1)$$

$$z = 3y - 1 \quad (2)$$

のとき、 $z$  を  $x$  で表すには、 $y$  に  $2x+4$  を代入して以下のようにすればいいですね。

$$\begin{aligned} z &= 3y - 1 \\ &= 3(2x + 4) - 1 \cdots [A] \\ &= 6x + 12 - 1 \\ &= 6x + 11 \end{aligned}$$

[A]  $\cdots$  (2) 式に (1) 式を代入した

## 合成関数の例 2

---

次は、 $f(x)$  や  $g(y)$  を使って書いた別の例を見てみます。書き方が違うだけで、やるべきことは同じです。

$$y = f(x) = x - 1 \quad (3)$$

$$z = g(y) = 2y^2 + 3 \quad (4)$$

のとき、合成関数  $g(f(x))$  を求めてみます。この場合も  $g(y)$  の  $y$  に  $f(x)$  つまり  $x-1$  を代入すればいいですね。

$$\begin{aligned} z &= g(y) \\ &= 2(x - 1)^2 + 3 \cdots [B] \\ &= 2(x^2 - 2x + 1) + 3 \cdots [C] \\ &= 2x^2 - 4x + 2 + 3 \\ &= 2x^2 - 4x + 5 \end{aligned}$$

[B]  $\cdots$  (4) 式に (3) 式を代入した

[C]  $\cdots$  2 乗を展開した





**\*1** ここでは仕組みが分かるように、 $\bar{x}$ 、 $\bar{y}$ 、 $\bar{z}$ という文字を使いましたが、以下のように $\bar{x}$ という文字だけを使って表すこともあるので、注意が必要です。

$$\begin{aligned}f(x) &= 2x + 4 \\g(x) &= 3x - 1\end{aligned}$$

のとき、

$$\begin{aligned}g(f(x)) &= 3(2x + 4) - 1 \\&= 6x + 12 - 1 \\&= 6x + 11\end{aligned}$$

この場合、 $f(x)$ の $x$ と $g(x)$ の $x$ は同じものではなく、単に「何らかの値を表すもの」といった意味合いで使われています。 $g(f(x))$ は $g(x)$ の $x$ のところに $f(x)$ を代入したもの、と考えるといいでしょう。

なお、 $g(f(x))$ と $f(g(x))$ とは等しいとは限りません。上の例であれば $g(f(x)) = 6x + 11$ ですが、次の計算式で示すように $f(g(x)) = 6x + 2$ になります。

$$\begin{aligned}f(g(x)) &= 2(3x - 1) + 4 \\&= 6x - 2 + 4 \\&= 6x + 2\end{aligned}$$

### 合成関数の例 3

合成関数の考え方も計算方法も簡単でしたね。複数の変数を使ったやや複雑な関数についても見ておきましょう。といっても、ここまでの計算と同様、代入と四則演算、べき乗だけしか使いません。以下のような関数があるものとします。

$$y_1 = -0.5x_1 - 0.5x_2 + 0.8 \quad (5)$$

$$y_2 = 0.5x_1 + 0.5x_2 - 0.2 \quad (6)$$

$$z = 0.5a(y_1) + 0.5a(y_2) - 0.8 \quad (7)$$

ただし、 $a(y)$ は $y$ の値によって0か1かを返す以下のような関数とします。 $a$ は定数を表すためによく使われる文字ですが、ここでは関数名として使われていることに注意してください。

$$a(y) = \begin{cases} 0 & (y \leq 0) \\ 1 & (y > 0) \end{cases}$$



このとき、 $z$  を  $x_1$  と  $x_2$  で表すと、以下のようになります。

$$\begin{aligned} z &= 0.5a(y_1) + 0.5a(y_2) - 0.8 \\ &= 0.5a(-0.5x_1 - 0.5x_2 + 0.8) \\ &\quad + 0.5a(0.5x_1 + 0.5x_2 - 0.2) - 0.8 \cdots [A] \end{aligned}$$

[A] … $y_1$  に (5) 式の  $-0.5x_1 - 0.5x_2 + 0.8$  を、 $y_2$  に (6) 式の  $0.5x_1 + 0.5x_2 - 0.2$  を代入した

さらに、 $a(z)$  を  $x_1$  と  $x_2$  で表すと、以下のようになります。

$$\begin{aligned} a(z) &= a(0.5a(-0.5x_1 - 0.5x_2 + 0.8) \\ &\quad + 0.5a(0.5x_1 + 0.5x_2 - 0.2) - 0.8) \cdots [B] \end{aligned}$$

[B] … $z$  に [A] 式を代入した

念のため、合成関数がどのように適用されているかを図にしておくと、以下のようになります。

$$a(z) = a(\underbrace{0.5a(\underbrace{-0.5x_1 - 0.5x_2 + 0.8}_{y_1}) + 0.5a(\underbrace{0.5x_1 + 0.5x_2 - 0.2}_{y_2})}_{z} - 0.8)$$

図3 合成関数が適用された様子

ところで、(7) 式などの形を見ると、 $a$  でくくって、式を簡単にできそうに見えますが、 $a$  は定数や変数ではなく関数であることに注意してください。実は、 $a$  が線形性のある関数であれば、 $a$  で「くくる」こともできるのですが、線形ではない（非線形な）関数なので以下のようにはできません \*2。

[C] …間違った例



**\*2 関数の線形性**というのは、 $a$  と  $b$  を定数としたとき、以下のような性質が成り立つことでしたね。

$$\begin{aligned} f(ax + by) &= f(ax) + f(by) \\ &= a \cdot f(x) + b \cdot f(y) \end{aligned}$$

つまり「2つの値を足してから関数を適用した値と、それぞれの値に関数を適用した結果を足した値が等しい」「定数を掛けてから関数を適用した値と関数を適用してから定数を掛けた値が等しい」という性質です（線形性については、この連載の第5回の記事を参照してください）。

## 簡単なニューラルネットワークの例

実は、ニューラルネットワークの順伝播（＝入力から出力に向かう信号の流れ）は合成関数を使ってそのまま表すことができます。例えば、合成関数の例 3 で見た式は、XOR（エクスクルーシブ・オア）を求めるためのニューラルネットワークを表しています。**XOR** とは、入力が等しければ **0** を返し、等しくなければ **1** を返すような論理演算のことです。つまり、表 1 のような計算になります。

入力1 ( $x_1$ )	入力2 ( $x_2$ )	出力 ( $a(z)$ )
0	0	0
0	1	1
1	0	1
1	1	0

表 1 XOR の演算

簡単な Python プログラムを作って、合成関数の例 3 の計算を実行させると、以下（図 4）のようになります。ここでは、プログラムの内容については特に解説しませんが、ちゃんと XOR の計算ができていることが分かります。

```
def a(y):
    if y <= 0:
        return 0
    else:
        return 1

def f(w1,x1,w2,x2,b):
    return w1*x1 + w2*x2 + b

def xor(x1,x2):
    y1 = f(-0.5, x1, -0.5, x2, 0.8)
    y2 = f(0.5, x1, 0.5, x2, -0.2)
    z = f(0.5, a(y1), 0.5, a(y2), -0.8)
    return a(z)

print(xor(0,0),xor(0,1),xor(1,0),xor(1,1))

0 1 1 0
```

図 4 XOR を求めるニューラルネットワークのプログラム例（Python）

合成関数の例 3 を図で表してみると図 5 のようになります。簡単なものですが、これを見るとニューラルネットワークそのものであることが分かります。以下の図と式の関係は動画でも説明しているので、ぜひとも参照してみてください。

動画1 合成関数

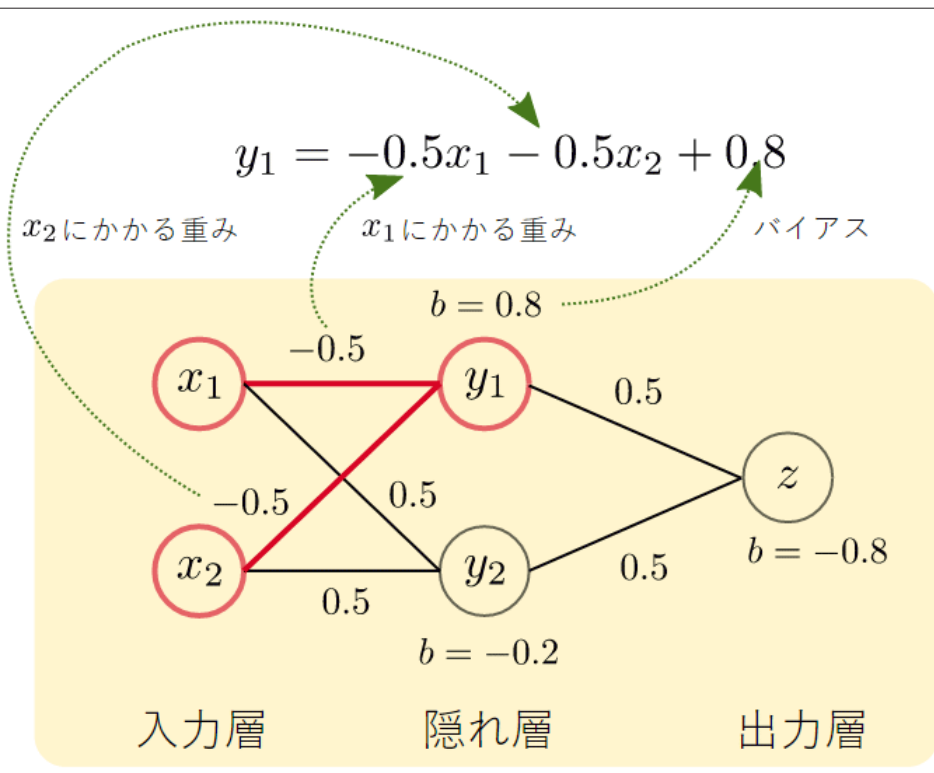


図5 XORを求めるニューラルネットワークの図

赤い太線で示した部分が以下の式に対応しています。

$$y_1 = -0.5x_1 - 0.5x_2 + 0.8 \quad \dots \boxed{A}$$

[A] …合成関数の例3の(5)式を再掲したもの

-0.5 という値は、 $x_1$  や  $x_2$  に掛ける「重み」です。0.8 という値はバイアスと呼ばれる値で、関数  $a$  を使って出力を決めるときに、0 以下かそうでないかという判定ができるようにするための調整用の値です。つまり、

$$\begin{cases} 0 & (-0.5x_1 - 0.5x_2 \leq -0.8) \\ 1 & (-0.5x_1 - 0.5x_2 > -0.8) \end{cases}$$

とする代わりに

$$\begin{cases} 0 & (-0.5x_1 - 0.5x_2 + 0.8 \leq 0) \\ 1 & (-0.5x_1 - 0.5x_2 + 0.8 > 0) \end{cases}$$

となるように、判定の基準となる値（閾値：しきいち）を左辺に移項したものです。

ちなみにニューラルネットワークの順伝播において、関数  $a$  のように、入力に対して出力する値を決める関数のことを**活性化関数**（Activation function）と呼びます。 $a$  という関数名は **A**ctivation の頭文字です。

$$y_1 = -0.5x_1 - 0.5x_2 + 0.8 \quad \dots \boxed{A} \text{ (再掲)}$$

[A] 式は、 $y_1$  に入力される値を表す式ですが、 $y_1$  から出力される値は活性化関数  $a$  を適用した結果です。つまり、 $a(y_1)$  が出力され、図 6 の右端にある  $z$  に入力されます。その部分と式との対応も見ておきましょう。

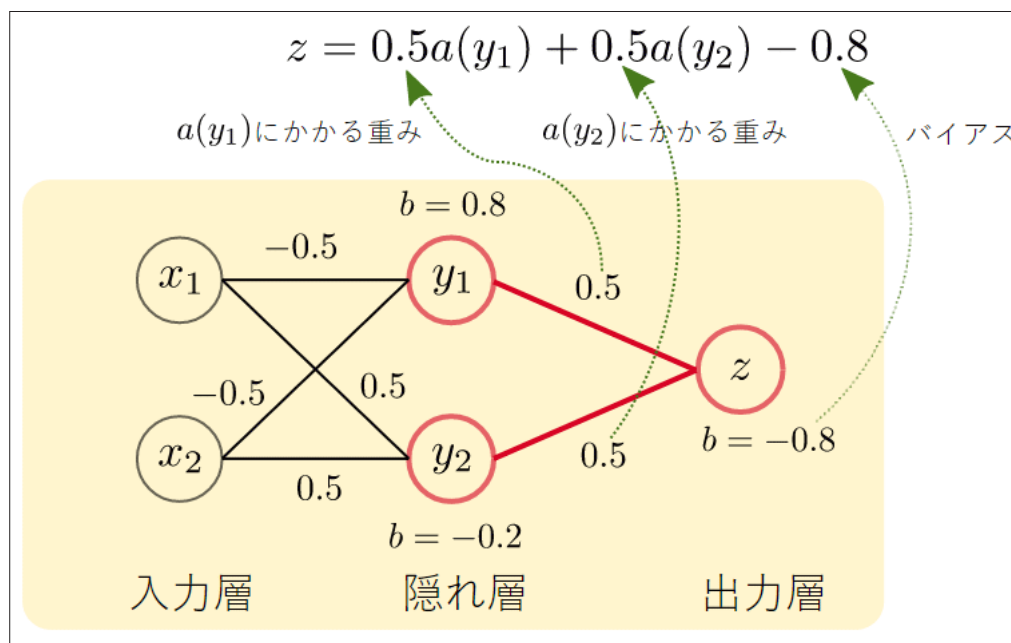


図 6 XOR を求めるニューラルネットワークの例

やはり、図 6 の赤い太線で示した部分が以下の式に対応しています。

$$z = 0.5a(y_1) + 0.5a(y_2) - 0.8 \quad \dots \boxed{B}$$

[B] …合成関数の例 3 の (7) 式を再掲したもの

[B] 式は、 $z$  に入力される値を表す式です。 $z$  からの出力は、活性化関数  $a$  を適用した結果です。つまり、 $a(z)$  が結果として出力されます。 $x_1$  や  $x_2$  に **0** や **1** を入力すると、 $a(z)$  がそれらの XOR を出力してくれるというわけです **\*3**。



**\*3** ここでは、重みやバイアスを **0.5** や **-0.5**、**-0.8** などのようにあらかじめ決めていましたが、ニューラルネットワークの学習とは、それらの重みを正解の値（教師データ）を基に調整していくことに他なりません。その際、合成関数を微分（偏微分）する必要が生じてきます。しかし、ここで見た関数  $a$  は **0** 以下の場合と **0** より大きい場合に分けて微分しても、微分係数は常に **0** にしかありません。そこで、一般には**シグモイド関数**や **ReLU 関数**と呼ばれる関数が活性化関数として使われます。

では、次に合成関数の微分法について見ていきましょう。最初に記したように、合成関数の微分はニューラルネットワークの学習に必須です。ただし、実際にニューラルネットワークの学習に使われる式を簡潔に表すためには線形代数の知識が必要になるので、ここでは、出発点のみということで、合成関数の微分の公式とその使い方に絞って説明します。

## 目標【その2】：合成関数の微分

合成関数の微分には、以下の公式が使えます。

$$\begin{aligned}\{g(f(x))\}' &= g'(f(x)) \cdot f'(x) \\ &= g'(y) \cdot f'(x)\end{aligned}\tag{1}$$

どの変数で微分するかを明確にするために、以下のように表すこともできます。

$$\frac{d}{dx}g(f(x)) = \frac{d}{dy}g(y) \cdot \frac{d}{dx}f(x)$$

さらに、 $y = f(x)$ 、 $z = g(y)$  とすると、この公式は、以下のように簡潔に表せます。

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}\tag{2}$$

合成関数は  $y$  で微分したものと  $x$  で微分したものの積になっています。これらの式は分数ではありませんが、あたかも分数であるかのように約分できるというわけです。(1) 式よりも意味が分かりやすいですね。

この式は、微分が次々とつながっているように見えるので「**連鎖律 (chain rule)**」と呼ばれることもあり、ニューラルネットワークで損失関数（正解と出力値の差を表す関数）の微分を各層に逆順にさかのぼりながら次々と伝えること（=**逆伝播**）を表すのに使われます。それによって、重みやバイアスの値を調整していくというわけです。

偏微分の場合も同様です。 $z$  が  $y$  の関数で、 $y$  が  $x_1, x_2, \dots$  の関数であるものとし、 $z$  を  $x_1$  で偏微分するなら、以下ようになります。

$$\frac{\partial z}{\partial x_1} = \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x_1}$$

## 解説【その2】：合成関数の微分（連鎖律）

合成関数は、1 ページ目で見ただけの方法で式を展開してから微分することもできますが、公式を使うと、ステップを分けて微分できます。ニューラルネットワーク（特にディープラーニング）の損失関数を逆伝播する場合のように、何段階にもわたる合成関数の場合、いちいち合成関数を全て展開してから微分するのは現実的ではありません。公式を使ってステップを分けると、途中の式が簡単になり、計算がかなりラクになります。

合成関数の微分の公式の意味を（1）式と（2）式とを例に見ておきましょう。

f と g の合成関数を x で微分する

$$\begin{aligned} \{g(f(x))\}' &= g'(f(x)) \cdot f'(x) \\ &= g'(y) \cdot f'(x) \end{aligned} \quad (1)$$

g(y) を y で微分したもの

f(x) を x で微分したもの

図7 合成関数の微分の公式（1）

（1）式では、 $g'(y)$  の部分が  $y$  での微分になることに注意してください。

合成関数 z を x で微分する

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx} \quad (2)$$

z を y で微分したもの

y を x で微分したもの

図8 合成関数の微分の公式（2）

（2）式の書き方だと、どの変数で微分するのがよく分かりますね。この式をよく見ると、以下の図9のように約分と同じような計算ができることが分かります。逆に言うと、何らかの変数を使って合成関数の微分を右辺のような何段階かの微分に変形できるということです。



$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx} \quad (2)$$

「約分」すれば  
左辺と等しくなる  
→右辺のように段階を分けて  
微分できる

図9 合成関数は段階を分けて微分できる

では、簡単な例を使って、合成関数の微分の公式の使い方に慣れておきましょう。なお、動画でも解説しているので、ぜひ参照してみてください。

### 合成関数の微分の例 1



2つの関数が以下のようなものであったとします。

$$f(x) = y = 3x^2 + 4$$

$$g(y) = y^2$$



このとき、合成関数  $g(f(x))$  を  $x$  で微分してみましょう。

$$\begin{aligned}
 \{g(f(x))\}' &= g'(f(x)) \cdot f'(x) \\
 &= g'(y) \cdot f'(x) \quad \dots [A] \\
 &= (y^2)' \cdot (3x^2 + 4)' \quad \dots [B] \\
 &= 2y \cdot 6x \quad \dots [C] \\
 &= 2(3x^2 + 4) \cdot 6x \quad \dots [D] \\
 &= (6x^2 + 8) \cdot 6x \\
 &= 36x^3 + 48x
 \end{aligned}$$

[A] …ここまでは合成関数の微分の公式

[B] … $g(y) = y^2$ 、 $f(x) = 3x^2 + 4$  を代入した

[C] …微分した。積の前の項は  $y$  で微分、後ろは  $x$  で微分していることに注意（後述）

[D] … $y = 3x^2 + 4$  を代入した

念のため、式を展開してから微分した結果と見比べておきましょう。

$$\begin{aligned}
 g(f(x)) &= g(y) \\
 &= y^2 \\
 &= (3x^2 + 4)^2 \quad \dots [E] \\
 &= (3x^2)^2 + 2(3x^2 \cdot 4) + 4^2 \quad \dots [F] \\
 &= 9x^4 + 24x^2 + 16 \\
 \therefore \{g(f(x))\}' &= (9x^4 + 24x^2 + 16)' \\
 &= 36x^3 + 48x \quad \dots [G]
 \end{aligned}$$

[E] … $y = 3x^2 + 4$  を代入した

[F] …二乗を展開した

[G] … $x$  で微分した

当然のことながら、ちゃんと答えが一致していますね。少し話を戻しますが「 $'$ 」を使った書き方だと、[B] 式を微分して [C] 式にするとところの方が分かりづらかったと思います。[B] 式を再掲しておきます。

$$(y^2)' \cdot (3x^2 + 4)' \quad \dots [B]$$

この場合、 $y^2$  は  $y$  で微分し、 $3x^2 + 4$  は  $x$  で微分する必要があるわけですが、それがはっきりと分かりませんね。しかし、以下の書き方で表すと、何で微分するかを迷うことはありません。「分母」にあたる部分に書かれた変数（例えば  $dz/dy$  なら  $y$ ）で微分するということが分かります。

このとき、 $y = 3x^2 + 4$ 、 $z = y^2$  なので、以下のようにになります。

$$\begin{aligned}\frac{dz}{dx} &= \frac{d}{dy} y^2 \cdot \frac{d}{dx} (3x^2 + 4) \quad \dots [H] \\ &= 2y \cdot 6x \quad \dots [I] \\ &= 2(3x^2 + 4) \cdot 6x \quad \dots [J] \\ &= (6x^2 + 8) \cdot 6x \\ &= 36x^3 + 48x\end{aligned}$$

[H] …右辺の  $z$  に  $y^2$  を代入し、 $y$  に  $3x^2 + 4$  を代入した

[I] …微分した

[J] … $y = 3x^2 + 4$  を代入した

## 合成関数の微分の例 2（偏微分の場合）

偏微分の場合も計算方法は同じです。これについても具体例で見ておきましょう。なお、こちらも動画での解説を用意してあります。ぜひ参照してみてください。



$$\begin{aligned}y &= 3x_1 + 2x_2 \\ z &= y^2\end{aligned}$$

のとき、 $z$  を  $x_1$  で偏微分すると以下のようにになります。

$$\begin{aligned}
\frac{\partial z}{\partial x_1} &= \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x_1} \\
&= \frac{\partial}{\partial y} y^2 \cdot \frac{\partial}{\partial x_1} (3x_1 + 2x_2) \quad \dots [A] \\
&= 2y \cdot 3 \quad \dots [B] \\
&= 2(3x_1 + 2x_2) \cdot 3 \quad \dots [C] \\
&= (6x_1 + 4x_2) \cdot 3 \\
&= 18x_1 + 12x_2
\end{aligned}$$

[A] …右辺の  $z$  に  $y^2$  を代入し、 $y$  に  $3x_1 + 2x_2$  を代入した

[B] …微分した

[C] … $y = 3x_1 + 2x_2$  を代入した

$$\begin{aligned}
z &= (3x_1 + 2x_2)^2 \\
&= (3x_1)^2 + 2 \cdot 3x_1 \cdot 2x_2 + (2x_2)^2 \\
&= 9x_1^2 + 12x_1 \cdot x_2 + 4x_2^2 \\
\therefore \frac{\partial z}{\partial x_1} &= (9x_1^2 + 12x_1 \cdot x_2 + 4x_2^2)_{x_1} \\
&= 18x_1 + 12x_2
\end{aligned}$$

念のため、合成関数の計算を行ってから偏微分して結果が一致することを確認しておきます。

はい、確かに一致していますね。合成関数の微分（連鎖律）も、微分の基本的な計算方法が分かっている場合は代入、四則演算、べき乗といった単純な計算の積み重ねだけでできることが分かったと思います。

ところで、今回の例として示した XOR の計算を行うニューラルネットワークでは、重みを表す変数が **6 個**、バイアスを表す変数が **3 個** あります。そのような単純な例でも、損失関数を重みやバイアスで偏微分し、逆伝播の計算を行う式を全て書くとあまりにも煩雑になってしまいます。とりあえず、同じ構造の例を図で表したものだけ掲載しておきますが、このまま計算することを考えると気が遠くなりそうですね。

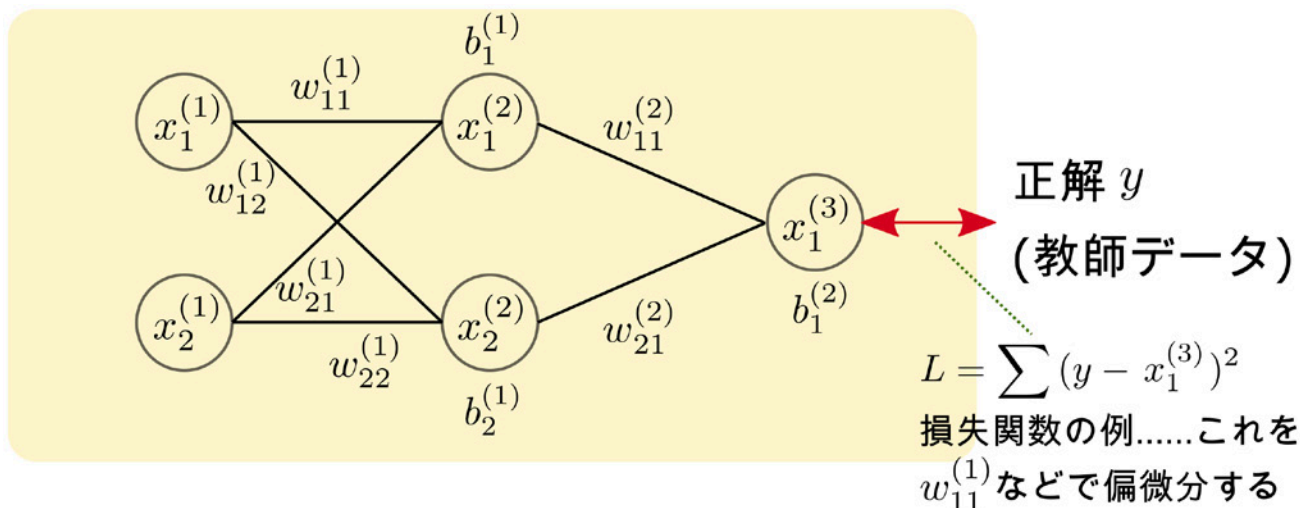


図 10 ニューラルネットワークと損失関数の例

XOR を求めるニューラルネットワークと構造は同じだが、重みなどを文字で表してある（変数名も少し変えてある）。変数名の右肩の (1)(2)(3) は何層目であるかを表すために付けてある。

ここでは、ノードを表す円の中に書かれた変数名はそのノードからの出力を表すものとする。例えば、中央の上のノードの  $x_1^{(2)}$  は第 2 層の 1 番目のノードからの出力を表す。バイアスの添え字については、何番目のノードへの入力にかかるバイアスであるかを表す。例えば、中央下の  $b_1^{(2)}$  は、第 1 層の 2 番目のノードにかかるバイアスを表す。重みの添え字は何番目の変数から何番目の変数に送られるときの重みであるかを表す。例えば、 $w_{21}^{(2)}$  であれば、2 番目の変数から 1 番目の変数に送られるときの重みであることを表す。

損失関数とは正解の値とニューラルネットワークの出力との差（誤差）を表すような関数です。例えば、誤差を二乗した値の総和（二乗和誤差）などが損失関数として使われます。損失関数を使って重みやバイアスを調整するには、損失関数を偏微分し、学習率  $\eta$ （イータと読みます）を掛けて値を更新します。例えば、損失関数を  $L$  とし、 $w_{11}^{(1)}$  を更新するのであれば、以下のような更新式が考えられます（第 7 回の記事で簡単に説明した勾配降下法です）。

$$w_{11}^{(1)} - \eta \frac{\partial L}{\partial w_{11}^{(1)}}$$

ここでは、損失関数が合成関数になっていることを確認しておいていただくだけで十分です。例えば、

$$x_1^{(3)} = a(w_{11}^{(2)} \boxed{x_1^{(2)}} + w_{21}^{(2)} x_2^{(2)} + b_1^{(2)}) \quad (1)$$

となっており、さかのぼると、

$$\boxed{x_1^{(2)}} = a(w_{11}^{(1)} x_1^{(1)} + w_{21}^{(1)} x_2^{(1)} + b_1^{(1)}) \quad (2)$$

となっています。枠で囲んだ部分を見れば合成関数であることが分かります（XOR を求めるニューラルネットワークの例で具体的に見たことも思い出してください）。これらの更新式を全ての重みとバイアスについて計算する必要があるわけです。

これらの式をもっと簡単に表す方法があれば、重みやバイアスを一気に求めることができるのですが、現時点ではまだ道具が足りていません。そういうわけで、今回は「はじめの一步」ということで合成関数の偏微分の方法を確認するところまでにとどめておくことにします。……で、実は、その道具というのがベクトルや行列、つまり線形代数です。次回からは線形代数を学んでいき、少しずつモヤモヤを晴らしていきたいと思います。

更新履歴（2025 年 2 月 19 日）

図 10 の損失関数の定義および  $w$  の添え字に一部誤りがあったため、図 10 を差し替え、補足説明を追記し、「ここでは、損失関数が合成関数になっている……」で始まる 1 文を修正しました。誤りがありましたことをお詫びし、訂正いたします。

