コメンタリー ～ 機械学習のエッセンス

加藤公一

Ver 20210119

1　はじめに

書籍「機械学習のエッセンス」（以下「文献 [1]」として引用します））について、補足的な説明をします。特に数学的な説明部分について、分かりづらいという指摘があったものや、誤植があるが増刷版での対応が難しいものなどについて説明します。

以下の解説では、適宜対応する文献 [1] のページを記しますが、書籍とは独立して、短編エッセイ集としても読めるように工夫するつもりです。

本ドキュメントは筆者の気まぐれにより更新されます。また、特にここを解説してほしいという要望があれば、追加で解説することがあるかもしれないですので、筆者にリクエストしてみてください。もちろん、すべてのリクエストに応えることは約束できませんが。

1.1　ライセンスについて

1.2　引用について

この文書を引用するときは GitHub のページへのリンク

`https://github.com/hamukazu/commentary_mlessence`

を示してください。

2　記号について

2.1　不等号

まず、$\leq$ という記号を初めてみておどろく人もいるようです。日本の中学・高校の教科書では $\leqq$ という記号をつかいますから。$\leq$ と $\leqq$、$\geq$ と $\geqq$ はそれぞれ同じ意味です。

2.2　部分集合

$a < b$ という表記は $a = b$ のときを含みませんが、集合包含関係を表す $A \subset B$ という表記は $A = B$ である場合を含むので気をつけてください。とくに「$A \subset B$ かつ $A \neq B$」を表す記号として $\subsetneq$ や $\subsetneqq$ という記号があります。形が雰囲気的に似ている（？）ので混同するかもしれませんが、$<$ はイコールの場合を含まず、$\subset$ はイコールの場合を含むので気をつけてください。また、日本語で「A は B の部分集合である」といった場合も同様に、$A = B$ である可能性を含んでいます。

2.3 対数

機械学習関連の書籍では、自然対数（ネイピア数 e を底とする対数）は ln で表すことが多いようですが、文献 [1] ではあえて log で表しました。これについてはいろいろな流儀があるのですが、機械学習界隈の多数決でいうと（きちんと統計をとったわけではないので、筆者の印象ですが）ln を使うことが多いようです。なぜ文献 [1] では log を使ったかというと、日本の高校の検定教科書ではその表記になっているからというのと、筆者自身が普段 log で表記しているからです。単純に「好みの問題」と言えるかもしません。本によっては log で常用対数（10 を底とする対数）を表すことがあるので気をつけてください。余談ですが、シンガポールで高校教育を受けた人に聞いたところ、自然対数は ln、常用対数は log で習ったと言っていました。

文脈によっては、そもそも対数の底がなにであるかを気にしないでよいことも多いです。例えば、対数尤度の最適化というのはよく機械学習で出てきますが、「$\log_a L$ を最小化する」ことと「$\log_b L$ を最小化する」ことは $a \neq b$ であっても同じことです。$\log_a L = \frac{\log_b L}{\log_b a}$ なので、定数倍の違いしかありませんから。

3 無限極限について

「n を無限に大きくしていったときに $1/n$ は限りなく 0 に近づく」ということは、直感的には明らかだと思うので、文献 [1] では（高校の検定教科書がそうしているように）「n を無限に大きくするとはどういうことか」の説明を省きました。多くの場合そのような「ざっくりとした理解」で間に合うのですが、一方で単純な計算でも無限というものを雑に扱ったために様々な矛盾（と思われるような結果）に直面してしまうこともあります。

例えば、計算式として $\frac{1}{\infty}$ というような書き方は間違いなのでやめましょう。これだと ∞ を数として扱っているように見えますが、∞ は数ではないので単純に演算はできません。

一般に数列 a_n について、

$$\lim_{n \to \infty} a_n = \alpha$$

という表記は、「n が無限に大きくなるとき a_n は α に近づく」と解釈されますが、このことの定義は次で与えられます。

> 任意の $\epsilon \in \mathbb{R}(\epsilon > 0)$ に対してある N が存在して $n > N$ ならば
> $|a_n - \alpha| < \epsilon$ となる

これだと分かりづらいかもしれないので、もう少し感情を（？）入れて表現すると次のようになります。

> どんなに小さな正の実数 ϵ に対しても、十分に大きな N を用意すれば、$n > N$ であるような n についてはすべて $|a_n - \alpha| < \epsilon$ とすることができる

この定義において ∞ という記号は出てきません。$n \to \infty$ という表現は、∞ という「数」があるということを言っているわけではないのです。

高校生向けの参考書を見ると、極限計算の分類として「これは $\frac{1}{\infty}$ 型ですね」というような説明を見ることがありますが、それは問題を分類して「$\frac{1}{\infty}$ 型」と名前をつけているだけなので間違いとは言えないです。だからといって計算式の途中で ∞ を数のように扱って $\frac{1}{\infty}$ のように書くのは間違いなので気をつけましょう。

4　二次形式の偏微分

二次関数の勾配とヘッセ行列の計算の部分（文献 [1]168〜169 ページ）について、途中の式展開が分かりづらいという指摘をいただいているので、ここに詳細を解説します。

二次形式とは n 次対称行列 $\boldsymbol{A}$ に対して

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$$

で定義されるものです。これで $\boldsymbol{A}$ を固定し、$\boldsymbol{x}$ を動かしたときにこの f がいつ最小になるかを考えます。

この式は、$\boldsymbol{A} = (a_{ij})$ とすると

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

と表されます。

f の勾配を求めたいので x_k で偏微分してみます。

$$\begin{aligned}
\frac{\partial f}{\partial x_k}(\boldsymbol{x}) &= \frac{\partial}{\partial x_k} \Big[a_{kk} x_k^2 + \sum_{j \neq k} a_{kj} x_k x_j + \sum_{i \neq k} a_{ik} x_i x_k \Big] \\
&= 2 a_{kk} x_k + \sum_{j \neq k} a_{kj} x_j + \sum_{i \neq k} a_{ik} x_i \\
&= 2 a_{kk} x_k + 2 \sum_{j \neq k} a_{ki} x_j \\
&= 2 \sum_{j=1}^{n} a_{kj} x_j \qquad (1)
\end{aligned}$$

ここが分かりづらいという指摘があったので、話をわかりやすくするため、とくに $n = 3$ として計算してみます。まず元の f の式は次のようになります。

$$\begin{aligned}
f(\boldsymbol{x}) &= \sum_{i=1}^{3} \sum_{j=1}^{3} a_{ij} x_i x_j \\
&= a_{11} x_1 x_1 + a_{12} x_1 x_2 + a_{13} x_1 x_3 \\
&\quad + a_{21} x_2 x_1 + a_{22} x_2 x_2 + a_{23} x_2 x_3 \\
&\quad + a_{31} x_3 x_1 + a_{32} x_3 x_2 + a_{33} x_3 x_3
\end{aligned}$$

これを例えば x_2 で偏微分してみます。x_2 を含まない項は 0 になるので、x_2 を含む項だけを考えます。x_2 を含む項でも、特に x_2 を 2 つ含む項（$a_{22} x_2 x_2 = a_{22} x_2^2$、$x_2$）が先に出てくる項（$a_{21} x_2 x_1$, $a_{23} x_2 x_3$）、x_2 が後に出てくる項（$a_{12} x_1 x_2$, $a_{32} x_3 x_2$）、と便宜上 3 つにわけて計算してみます。

$$
\begin{aligned}
\frac{\partial f}{\partial x_2}(\boldsymbol{x}) &= \frac{\partial}{\partial x_2}(a_{22}x_2^2) + \frac{\partial}{\partial x_2}(a_{21}x_2x_1 + a_{23}x_2x_3) + \frac{\partial}{\partial x_2}(a_{12}x_1x_2 + a_{32}x_3x_2) \\
&= 2a_{22}x_2 + (a_{21}x_1 + a_{23}x_3) + (a_{12}x_1 + a_{32}x_3) \\
&= 2a_{22}x_2 + (a_{21}x_1 + a_{23}x_3) + (a_{21}x_1 + a_{23}x_3) \\
&\qquad \text{（ここで } A \text{ が対称行列なので } a_{ij} = a_{ji} \text{ を使った）} \\
&= 2a_{22}x_2 + 2a_{21}x_1 + 2a_{23}x_3 \\
&= 2\sum_{j=1}^{3} a_{2j}x_j
\end{aligned}
$$

この式は式変形 (1) の具体的な例になっていることに注意してください。実際 (1) で $n = 3$, $k = 2$ とするとこの式になります。

ここでまだ $n = 3$ として話を進めていきましょう。次にヘッセ行列を求めたいのですが、ここでの計算を踏まえて x_1, x_3 による偏微分も同様に計算すると f の勾配が次のように計算できます。

$$
\nabla f = \begin{pmatrix} 2(a_{11}x_1 + a_{12}x_2 + a_{13}x_3) \\ 2(a_{21}x_1 + a_{22}x_2 + a_{23}x_3) \\ 2(a_{31}x_1 + a_{32}x_2 + a_{33}x_3) \end{pmatrix} = 2\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 2\boldsymbol{A}\boldsymbol{x}
$$

さらにこの勾配をとって $\nabla^2 f$ を計算したいのですが、例えば x_2 での偏微分を計算してみましょう。

$$
\begin{aligned}
\frac{\partial}{\partial x_2}(\nabla f) &= \frac{\partial}{\partial x_2}\begin{pmatrix} 2(a_{11}x_1 + a_{12}x_2 + a_{13}x_3) \\ 2(a_{21}x_1 + a_{22}x_2 + a_{23}x_3) \\ 2(a_{31}x_1 + a_{32}x_2 + a_{33}x_3) \end{pmatrix} \\
&= \begin{pmatrix} 2a_{12} \\ 2a_{22} \\ 2a_{32} \end{pmatrix} \\
&\qquad \text{（ここで } x_2 \text{ を含む項以外は } 0 \text{ になることに注意）}
\end{aligned}
$$

x_1, x_3 による偏微分も同様に計算できます。$\left[\frac{\partial}{\partial x_1}(\nabla f)\right]^T$, $\left[\frac{\partial}{\partial x_2}(\nabla f)\right]^T$, $\left[\frac{\partial}{\partial x_3}(\nabla f)\right]^T$ を縦に並べたものがヘッセ行列になるので、以下のように計算できます。

$$
\begin{aligned}
\nabla^2 f &= \begin{pmatrix} \left[\frac{\partial}{\partial x_1}(\nabla f)\right]^T \\ \left[\frac{\partial}{\partial x_2}(\nabla f)\right]^T \\ \left[\frac{\partial}{\partial x_3}(\nabla f)\right]^T \end{pmatrix} \\
&= \begin{pmatrix} 2a_{11} & 2a_{21} & 2a_{31} \\ 2a_{12} & 2a_{22} & 2a_{32} \\ 2a_{13} & 2a_{23} & 2a_{33} \end{pmatrix} \\
&= \begin{pmatrix} 2a_{11} & 2a_{12} & 2a_{13} \\ 2a_{21} & 2a_{22} & 2a_{23} \\ 2a_{31} & 2a_{32} & 2a_{33} \end{pmatrix} \quad \text{（ここで } \boldsymbol{A} \text{ が対称行列であることを使った）} \\
&= 2\boldsymbol{A}
\end{aligned}
$$

ここまでの計算を（$n=3$ とは限らない）一般の n についての説明に言い換えると以下のようになります。ヘッセ行列の (k,l) 成分を h_{kl} とすると、h_{kl} は ∇f の第 l 成分 $(\nabla f)_l$ を x_k で偏微分したものになります。$(\nabla f)_l$ は式 (1) で表されるので、次のような計算ができます。

$$\begin{aligned} h_{kl} &= \frac{\partial}{\partial x_k}(\nabla f)_l \\ &= \frac{\partial}{\partial x_k}\left(2\sum_{j=1}^{n} a_{lj}x_j\right) \\ &= 2a_{lk} \\ &= 2a_{kl} \quad \text{ここで } \boldsymbol{A} \text{ が対称行列であることを使った）} \end{aligned}$$

このことより次が示されました。

$$\nabla^2 f = 2\boldsymbol{A}$$

さて、一般になめらかな多変数関数 f の極大・極小について、次のような定理が知られています（文献 [1]168 ページ参照）

> f($\boldsymbol{x}$) が $\boldsymbol{x}_0$ で極大になるのは $\nabla f(\boldsymbol{x}_0)=\boldsymbol{0}$ かつ $\nabla^2 f(\boldsymbol{x}_0)$ が負定値のときであり、
> 極小になるのは $\nabla f(\boldsymbol{x}_0)=\boldsymbol{0}$ かつ $\nabla^2 f(\boldsymbol{x}_0)$ が正定値のときである。

このことを上記二次形式の計算で確かめてみましょう。

$$\nabla f(\boldsymbol{0}) = 2\boldsymbol{A}\boldsymbol{0} = \boldsymbol{0}$$

となります。なので、上記定理によれば、もし $\nabla^2 f(\boldsymbol{0})$ が正定値ならば f は $\boldsymbol{x}=\boldsymbol{0}$ で極小ということになりますが、もともと正定値の定義はなんであったか思い出しましょう。$\boldsymbol{A}$ が正定値であるとは、次で定義されます。

> $\boldsymbol{x}\neq\boldsymbol{0}$ であるベクトルに対して $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}>0$ である。

これはつまり $f(\boldsymbol{x})$ が $\boldsymbol{x}=\boldsymbol{0}$ で最小値（したがって極小値）をとることを意味しており、つまり定理が言っていることは二次形式 f については確認できました。極大値についても同様です。

5　二次方程式の数値解について

文献 [1] の 177〜178 ページで、二次方程式の数値計算をするときの式が間違っています。$b=0$ のときだけ正しい値にならないので、多くの場合には問題ないのですが、そのことが逆に厄介なバグかもしれません。

二次方程式

$$ax^2+bx+c=0$$

の解

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

について、分子の引き算が小さな数になるのを避けないといけないです（詳細は文献 [1]）。そのため、$b\geq 0$ のときは

$$x=\frac{-b-\sqrt{b^2-4ac}}{2a}$$

を計算し、$b < 0$ のときは

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

を計算します。2 つの解を $x = \alpha, \beta$ とすると、$\alpha\beta = c/a$ なので、もう片方の解はこれで計算します。

では一つ目の解の計算は。次のような関数 $s(\xi)$ を定義すると統一的に計算できます。

$$s(\xi) = \begin{cases} -1 & (\xi < 0) \\ 1 & (\xi \geq 0) \end{cases}$$

このとき一つ目の解は次の式で求められます。

$$x = \frac{-b - s(b)\sqrt{b^2 - 4ac}}{2a}$$

文献 [1] の方では sign 関数を使っていますが、それは $b = 0$ のときだけ上記の式と結果がことなります。

6 バイアスとバリアンス

文献 [1] の 303〜308 ページについてですが、本文の説明とコードの内容が合っていないのではという指摘がありました。その点について説明します。

グラフの描画では `fill_between` を使ってバイアスとバリアンスを可視化しています（文献 [1]307〜308 ページ参照）が、バイアスを b とバリアンス v とすると、バイアスとバリアンスの積み上げを表現するには、`fill_between` を 2 回呼び出すとして、1 回目は 0 と b、2 回目は b と $b + v$ を引数にして呼び出す必要があります。文献 [1] にあるように、データ D を使って予測した値を $\hat{f}_D(x)$ で表すと、データの集合 $\mathcal{D}$ が与えられたときに 2 乗誤差の平均は次のようになります。

$$E_{\mathcal{D}}\left[\left(f(x) - \hat{f}_D(x)\right)^2\right] = \left(f(x) - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]\right)^2 + E_{\mathcal{D}}\left[\left(\hat{f}_D(x) - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]\right)^2\right] \quad (2)$$

ここで $E_{\mathcal{D}}$ は考えうるデータ $D \in \mathcal{D}$ すべてについての平均ということです。つまり D が動くときの平均を意味します。

この右辺の第 1 項がバイアスであり、第 2 項がバリアンスなので、バイアス + バリアンスは左辺つまり、2 乗誤差の平均になります。これの式を利用して文献 [1]307〜308 ページではグラフの描画をしています。

ここではさらに文献 [1] には書いていない式 (2) の導出もやってみます。

$$\begin{aligned}
&E_{\mathcal{D}}\left[\left(f(x) - \hat{f}_D(x)\right)^2\right] \\
&= E_{\mathcal{D}}\left[f(x)^2 - 2f(x)\hat{f}_D(x) + \hat{f}_D(x)^2\right] \\
&= f(x)^2 - 2f(x)E_{\mathcal{D}}\left[\hat{f}_D(x)\right] + E_{\mathcal{D}}\left[\hat{f}_D(x)^2\right] \\
&= f(x)^2 - 2f(x)E_{\mathcal{D}}\left[\hat{f}_D(x)\right] + E_{\mathcal{D}}\left[\hat{f}_D(x)\right]^2 - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]^2 + E_{\mathcal{D}}\left[\hat{f}_D(x)^2\right] \\
&= \left(f(x) - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]\right)^2 + \left(E_{\mathcal{D}}\left[\hat{f}_D(x)^2\right] - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]^2\right)
\end{aligned}$$

この第 2 項（2 つめのカッコのなか）がバリアンスに一致すればいいのですが、

$$
\begin{aligned}
&E_{\mathcal{D}}\left[\left(\hat{f}_D(x) - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]\right)^2\right] \\
&= E_{\mathcal{D}}\left[\hat{f}_D(x)^2 - 2\hat{f}_D(x)E_{\mathcal{D}}\left[\hat{f}_D(x)\right]E_{\mathcal{D}}\left[\hat{f}_D(x)\right]^2\right] \\
&= E_{\mathcal{D}}\left[\hat{f}_D(x)^2\right] - 2E_{\mathcal{D}}\left[\hat{f}_D(x)\right]\cdot E_{\mathcal{D}}\left[\hat{f}_D(x)\right] + E_{\mathcal{D}}\left[\hat{f}_D(x)\right]^2 \\
&= E_{\mathcal{D}}\left[\hat{f}_D(x)^2\right] - E_{\mathcal{D}}\left[\hat{f}_D(x)\right]^2
\end{aligned}
$$

となるので、式 (2) が示せました。

参考文献

[1] 加藤公一「機械学習のエッセンス」SB クリエイティブ, 2018 年