研究員の眼

絶対値をとるか、二乗するか

機械学習での評価方法は誰が決める?

保険研究部 主席研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

最近、AI に関するニュースが各種メディアで連日のように報じられている。例を挙げると、
― 生成 AI で「フィッシングサイト」識別　警察庁が 2025 年度までに導入へ（毎日新聞 2024.3.21）
― 公的機関の AI 活用例、世界で共有　G7 デジタル相会合（朝日新聞デジタル 2024.3.16）
― 世界初の AI 規制法、EU 議会で可決 制裁金最大 56 億円（Forbes Japan 2024.3.14）
― 米新興企業の AI 投資 3.7 兆円に ロボや医療に裾野拡大（日本経済新聞 2024.3.12）
といった感じだ。いま、世の中は「AI 全盛時代」を迎えつつあるのかもしれない。

AI と言えば機械学習だ。AI は、まず、データをもとに機械学習をする。そして、与えられたデータを分類したり、与えられたデータをもとに予測をしたりする。通常、多くのデータで機械学習をしていけば、分類や予測の精度は高まっていく。2022 年に登場した生成 AI は、文章や画像などのコンテンツを作り出す。まさに、人工の“知能”と呼ぶのにふさわしい発展を続けている。

機械学習のうち予測に関するものは、以前から「回帰分析」として行われてきたものの、大幅な拡張と見ることができる。予測の機械学習では、どれくらい予測があたったか、の評価が重要となる。

今回は、予測の機械学習における評価について、考えてみることとしたい。

◇ 予測と正解の誤差 ― 絶対値をとるか、二乗するか

通常、予測の機械学習は、正解のあるデータを使う「教師あり学習」として行われる。モデルの予測値と正解の誤差を計算して、それが小さくなるようにモデルを改良していく。

ただし、誤差を単純な引き算として計算すると、一般に、複数の予測のうち、ある箇所のプラスの誤差と、別のある箇所のマイナスの誤差が相殺し合って誤差の合計が小さくなってしまう。

そこで、単純な引き算ではなく、引き算した結果のマイナスの値をプラスに変換するような計算が必要となる。そこで考えつくのが、(1) 誤差の絶対値をとる方法と、(2) 誤差を二乗する方法だ。

複数の予測で、誤差の絶対値をとって、その平均を計算したものは「平均絶対誤差（Mean Absolute Error, MAE)」と呼ばれる。一方、誤差を二乗して、その平均の平方根を計算したものは「平均平方根二乗誤差（Root Mean Squared Error, RMSE)」と呼ばれる。

RMSE と MAE はマイナスの値にはならず、どちらの評価指標も 0 に近いほど誤差が小さい、つまり予測の精度が高いことを意味する。数学的には、RMSE ≧ MAE （等号は複数の予測の誤差の絶対値がすべて等しい場合)となることが示される。

実は、MAE と RMSE の間では、どちらが優れた評価指標か、という議論が長らく繰り広げられてきた。

◇ RMSE は微分可能で使いやすい

まず、議論の大きな論点として、微分可能かどうかという点が挙げられる。機械学習においては、予測の評価に応じて予測値を見直し、再び評価してまた予測値を見直して…という作業を繰り返しながら予測の精度を高めていく。これは、「最適化アルゴリズム」と呼ばれる。

このとき、評価の算式(関数)が微分可能だと、予測値を見直したときに着実に正解に近付くことができる。つまり、最適化できる。

RMSE は二乗の算式なので微分ができる。一方、MAE は絶対値をとる算式なので、誤差が 0 となる箇所で微分ができない。(誤差がプラスの方からこの箇所に近付くと傾きがプラス、誤差がマイナスの方から近付くと傾きがマイナスとなり、傾きが不連続となる。)

そのため、機械学習の最適化アルゴリズムの観点からは RMSE が使いやすいということになる。

◇ MAE は外れ値に強い

一方、議論のもう 1 つの論点として、外れ値の影響をどのくらい受けるか、という点がある。複数の予測のうち、ある予測だけ正解から大きく外れた場合、予測の評価にどう影響するかという点だ。

RMSE は誤差を二乗するので、外れ値の影響が大きく出ることとなる。たった 1 ヵ所でも予測を大きく外すと、他の予測は大体当たっていたとしても、評価は下がってしまう。

一方、MAE は、二乗の計算がないため、外れ値があっても外れたなりの評価の低下にとどまる。

この様子を、具体例をもとに見ていくこととしよう。A 氏～J 氏の 10 人の成人の体重予測を考える。この予測では、MAE は 1.7、RMSE は 2.0 となっている。

	正解体重(kg)	予測体重(kg)	(正解－予測) の絶対値	(正解－予測) の二乗
A 氏	65	63	2	4
B 氏	75	77	2	4
C 氏	48	47	1	1
D 氏	58	60	2	4
E 氏	62	63	1	1
F 氏	85	85	0	0
G 氏	90	92	2	4
H 氏	51	53	2	4
I 氏	44	40	4	16
J 氏	74	75	1	1
平均	—	—	1.7	3.9
平均の平方根	—	—	—	2.0

これに対して、G 氏の体重を 150kg と予測して正解から大きく外してしまった場合を考えてみる。この場合、MAE は 7.5、RMSE は 19.1 となっている。両者とも G 氏の予測を外したことの影響が数値の増大として表れているが、MAE に比べて、RMSE はより大きく増大していることがわかる。

	正解体重(kg)	予測体重(kg)	(正解－予測) の絶対値	(正解－予測) の二乗
A 氏	65	63	2	4
B 氏	75	77	2	4
C 氏	48	47	1	1
D 氏	58	60	2	4
E 氏	62	63	1	1
F 氏	85	85	0	0
G 氏	90	150	60	3600
H 氏	51	53	2	4
I 氏	44	40	4	16
J 氏	74	75	1	1
平均	—	—	7.5	363.5
平均の平方根	—	—	—	19.1

MAE は、RMSE に比べると外れ値の影響を受けにくい評価指標と言えるだろう。

◇ 率で評価する方法もあるが…

RMSE と MAE には、それぞれ長所がある。そして、それとは別に共通する短所もある。それは、単位に依存するという点だ。先ほどの体重の予測の例で言えば、RMSE も MAE も(kg)の単位を持つ。

例えば、体重とともに身長も予測して、体重と身長の予測精度を比較したい、ということを考えると困ったことになる。体重は(kg)、身長は(cm)などと単位が異なってしまい、比較ができないためだ。

そこで、正解と予測を引き算した結果ではなく、その引き算の結果を正解の値で割り算して「誤差率」の形にして、評価に用いることが考えられる。

MAE と RMSE に対応して、(3) 誤差率の絶対値をとる方法と、(4) 誤差率を二乗する方法がありうる。誤差率の絶対値をとって、その平均を計算したものは「平均絶対誤差率 (Mean Absolute Percentage Error, MAPE)」と呼ばれる。誤差率を二乗して、その平均の平方根を計算したものは「平均平方二乗誤差率 (Root Mean Square Percentage Error, RMSPE)」と呼ばれる。

RMSPE は、RMSE と同様に微分可能だ。一方、MAPE は MAE と同様に比較的外れ値の影響を受けにくい。しかも、両者とも単位を持たない、つまり単位異存性がない。こう見ていくと、誤差率ベースの RMSPE や MAPE はいいことずくめではないか、という気がしてくる。だが、そううまくはいかない。“率”ならではの問題点もあるためだ。

先ほどの体重の予測の例で、10 人目として、成人の J 氏の代わりに新生児の“K ちゃん”が入っていたとしよう。(ここで、「なんで 1 人だけ新生児が入っているのか？ そもそも成人と新生児の体重を同一のモデルで予測することは、ナンセンスではないか？」という読者諸氏のご指摘もあるだろう。ご指摘は誠にその通りであるが、ここは、あくまで架空の話として進めさせていただきたい。)

新生児の K ちゃんが入っていた場合、MAE は 1.7、RMSE は 2.0 となった。たまたまではあるが、J 氏の代わりに K ちゃんが入っていたとしても、MAE と RMSE は、先ほどの予測のまま変わらなかった。

	正解体重(kg)	予測体重(kg)	(正解－予測) の絶対値	(正解－予測) の二乗
A 氏	65	63	2	4
B 氏	75	77	2	4
C 氏	48	47	1	1
D 氏	58	60	2	4
E 氏	62	63	1	1
F 氏	85	85	0	0
G 氏	90	92	60	3600
H 氏	51	53	2	4
I 氏	44	40	4	16
K ちゃん	2	3	1	1
平均	—	—	1.7	3.9
平均の平方根	—	—	—	2.0

ところが、誤差率ベースの MAPE と RMSPE で見ると、様子が違ってくる。

次の表は、上が J 氏が入っていた元々の予測、下が J 氏の代わりに K ちゃんが入っていたとした場合の予測だ。

	正解体重(kg)	予測体重(kg)	(正解－予測)/正解の絶対値	(正解－予測)/正解の二乗
A氏	65	63	0.03	0.0009
B氏	75	77	0.03	0.0007
C氏	48	47	0.02	0.0004
D氏	58	60	0.03	0.0012
E氏	62	63	0.02	0.0003
F氏	85	85	0	0
G氏	90	92	0.02	0.0005
H氏	51	53	0.04	0.0015
I氏	44	40	0.09	0.0083
J氏	74	75	0.01	0.0002
平均	—	—	0.029	0.0014
平均の平方根	—	—	—	0.037

	正解体重(kg)	予測体重(kg)	(正解－予測)/正解の絶対値	(正解－予測)/正解の二乗
A氏	65	63	0.03	0.0009
B氏	75	77	0.03	0.0007
C氏	48	47	0.02	0.0004
D氏	58	60	0.03	0.0012
E氏	62	63	0.02	0.0003
F氏	85	85	0	0
G氏	90	92	0.02	0.0005
H氏	51	53	0.04	0.0015
I氏	44	40	0.09	0.0083
Kちゃん	2	3	0.5	0.25
平均	—	—	0.078	0.0264
平均の平方根	—	—	—	0.162

J氏がKちゃんに代わる前は、MAPEが0.029、RMSPEが0.037であったのに対し、代わった後は、MAPEが0.078、RMSPEが0.162と拡大している。これは、体重の軽いKちゃんについて、誤差率が跳ね上がってしまうためだ。

このように、正解がゼロに近付くと、誤差率は大きくなる。そしてゼロになると無限大に発散してしまう。MAPEやRMSPEには、ゼロに近い値の誤差の評価が過大になるという問題があるわけだ。

◇ AIがさらに進化したときに人間の役割は?

結局、どの評価方法にも長所や短所がある。機械学習において、どの方法で誤差を評価すべきか。これについては、しばらくは、データの特徴や外れ値などを見ながら人間が判断していくことになる。

ただし、今後さらにAIが進化していくと、こうした機械学習の方法に関する判断までもAI自身が行うようになるかもしれない。「私の知能の進化の仕組みは、私が決めます。」と言わんばかりに。

そのとき、人間の役割はどうなるだろうか?

「考えることや判断ごとは、すべてAIにおまかせ」というのでは、どこか味気ない。思考することをAIにとられてしまっては、「人間は考える葦(あし)である」という、思想家パスカルの言葉も成り立たなくなってしまう。

その代わりに、人間には、AIが行った判断の妥当性を判断する「スーパーバイザー」のような役回りが待っているのかもしれない。

2045年には訪れるだろうと予想されている“シンギュラリティ(人間の知性を上回るAIの誕生)”まで、あと約20年だ。AIに関する日々のニュース報道をみながら、そんなことに思いを巡らせるのもよいだろう。

(参考文献)

「生成AIで「フィッシングサイト」識別　警察庁が2025年度までに導入へ」(毎日新聞 2024.3.21)

「公的機関のAI活用例、世界で共有　G7デジタル相会合」(朝日新聞デジタル 2024.3.16)

「世界初のAI規制法、EU議会で可決 制裁金最大56億円」(Forbes Japan 2024.3.14)

「米新興企業のAI投資3.7兆円に ロボや医療に裾野拡大」(日本経済新聞 2024.3.12)

「なっとく！機械学習」Luis G.Serrano著，株式会社クイープ監訳(翔泳社，2022年)

