尤度にもとづく予測評価法
A Likelihood-based Method
for Forecasts Assessment

http://hdl.handle.net/10787/3886

2016.8.20
石黒真木夫@統計思考院@統計数理研究所

1

概要

このメモの目的は、各種予測に性能評価（数値化）を与える方法を説明する事である。

過去になされた予測の成績を事実に基づいて評価することが重要であることに間違いないが、より重要なのは過去の経験を「今」に生かすことである。「今」のまだ事実が確認されていないことに関する予測の性能を示すことの方がより重要である。これが可能であるのは、(1) 過去の予測が一定の方式に従って生み出されたものであり、(2) 過去の予測の評価がその「予測方式」の評価になっており、(3)「今」の予測が同じ予測方式に則ったものである、という3条件が満たされている場合のみである。

このメモで、様々な予測方式の優劣を2次元的に評価する表現する方法を示す。2次元指標の成分に適当な重みづけをすれば1次元指標が得られるが、いかなる重みづけが適当だあるか決める科学的に客観的な基準は存在しない。しかし、社会的に予測が必要とされている場面で、予測の目的に照らしてこの重みづけを合意することは可能であり、「社会的ご」客観的な1次元指標を構成することが可能であると思われる。

事実とこれに関する予測の記録から予測方式の性能評価は得られるが、偏りのある記録に基づく性能評価は信頼性を持たない。偏りのない記録を得る方法についても解説する。

2

事実と予測

言ったことが事実を言い当てているか、という形の問題は多い。

たとえば、「明日は晴れるだろう」という予測が当たるかどうか、などが典型的である。

しかし、同じ天気を晴と思う人とそう思わない人がいる場合がある。事実というものが単純に観察の結果として客観的に把握できない場合があるのである。そのような場合には「事実認定」という作業が必要になる。裁判における「冤罪」の存在は、判決当時の事実認定が後に覆されることがあるということを意味している。

事実認定も統計学が扱うべき範囲の問題であるが、このメモでは、簡単のため、事実はすべて認定済であるとし、その事実と予測の関係を調べる方法について説明する。

2種の予測を区別して扱う必要がある。「明日は晴れ」という形の予測と「明日の降水確率は30%」という形の予測の2種である。前者を「カテゴリカル予測」、後者を「確率予測」と呼ぶことにする。

3

カテゴリカル予測の場合

たとえば、「明日立川で雨が降るだろう」という予測が、カテゴリカル予測である。天気を「雨」、「晴れ」、「それ以外」のようないくつかのカテゴリーに分類して、明日の天気がどのカテゴリーに属するかを予測する場合である。

明日になれば、この予測が当たっているかいないか分かる。

4

記号

S_i: i 番目の「予測」
F_k: k 番目の「事実」

例えば、

S_{800} = 「1990年から2020年の間に関東地方で震度7以上の地震が観測される」
F_{1001} = 「2011年3月11日に三陸沖でマグニチュード9の地震が起きた」

すべての「予測」と「事実」に「時刻」 が与えられているものとする。ただし、事実の時刻はその事実が起きた日付と時刻とし、予測の時刻はその予測が公表された時刻とする。記号としては事実F の時刻を$T(F)$, 予測 S の時刻を$T(S)$で表す。

「予測」と「事実」が与えられたとき、次の式で定義される関数 R を「当否関数」と呼ぶことにする。

$$R(F,S)=\begin{cases}1 & \text{予測 } S \text{ が事実 } F \text{ を「言い当て」ている場合}\\ 0 & \text{それ以外}\end{cases}$$

5

当否表

ある年の7月初旬の雨に関する予測、F_1、F_2、F_3 と記録、S_1、S_2、…、S_5 で決まる当否関数の値を表の形にまとめると、たとえば、次のようなものが得られる。

	S_1:7/1 に立川で雨	S_2:7/2 に立川で雨	S_3:7/7 に立川で雨	予測の有無
F_1:7/1 に立川で雨	1	0	0	1
F_2:7/3 に立川で雨	0	0	0	0
F_3:7/5 に立川で雨	0	0	0	0
F_4:7/7 に立川で雨	0	0	1	1
F_5:7/10 に立川で雨	0	0	0	0
予測の当否	1	0	1	

この表の右端の欄は、その「行」の事実が、予測されえいたときに1、それ以外0 となっている。 最下段は、その「列」の予測が当たっていたときに1, それ以外で 0 である。

この表から、3つの予測のうち2つが当たっていること、5回の雨のうち2回が予測されていたことが分かる。

6

的中率と捕捉率

予測の性能を測る指標として

$$\text{的中率} = \frac{\text{当たった予測の数}}{\text{予測の数}}$$

$$\text{捕捉率} = \frac{\text{予測が当たっていた事実の数}}{\text{予測したかった事実の数}}$$

が定義される。的中率、捕捉率ともに 1 となるのが理想であるが、これを達成するのは難しい。前スライドの例では

的中率 2/3=0.67　捕捉率　2/5=0.4 である。

的中率と捕捉率がそれぞれ、どの程度であれば予測の性能として十分であるか、を客観的に定めることはできない。予測方式Aの的中率が予測方式Bの的中率より高く、捕捉も予測方式Aの方が高ければ、予測方式Aの方がBより優れていると言えるが、Aの的中率は高いが、捕捉率が低いというような場合には、AとBのどちらをよしとするかを客観的に決めることはできない。

7

事実予測対照表

当否表は、事実と予測の関係を直感的に表すよい方法であるが、扱う事実や予測の数が多くなると、扱いにくい。同じ内容を次のような2列の表で表現することができる。 この形で、当たっている予測、当たってない予測、予測された事実、などが分かり、この表があれば予測の的中率、捕捉率が求められる。

事実	予測
F_1:7/1 に立川で雨	S_1:7/1 に立川で雨
	S_2:7/2 に立川で雨
F_2:7/3 に立川で雨	
F_3:7/5 に立川で雨	
F_4:7/7 に立川で雨	S_3:7/7 に立川で雨
F_5:7/10 に立川で雨	

8

客観的事実予測対照表

予測方式の客観的評価を得るためには対照表が客観的であることが必要である。たとえば、次のような対照表を示して、予測方式が優れたものであることを示そうとする場合、 以下の質問に対する答えを用意しておく必要がある。

1.「予測」といっているが、実は後出しではないですか？
2. 外れた予測を削除していませんか？
3 予測されていなかった事実を削除していませんか？

たとえば、前スライドの表から第2，3，4，6行を削除するとつぎのような表が得られる。

事実	予測
F_1:7/1 に立川で雨	S_1:7/1 に立川で雨
F_4:7/7 に立川で雨	S_3:7/7 に立川で雨

積極的に虚偽のデータを作っているのではないが、「不都合な真実」を隠すことによって、予測の性能がよいように見せていることになる。

要するに、同じ測定を行っただれても同じ対照表が作れるという再現性を保証することが必要ということである。具体的方法については後述する。

9

捕捉率と的中率の関係

	S_1:7/1に雨	S_2:7/2に雨	S_3:7/3に雨	S_4:7/4に雨	S_5:7/5に雨	S_6:7/6に雨	S_7:7/7に雨	予測の有無
F_1:7/1に雨	1	0	0	0	0	0	0	1
F_2:7/3に雨	0	0	1	0	0	0	0	1
F_3:7/5に雨	0	0	0	0	1	0	0	1
F_4:7/7に雨	0	0	0	0	0	0	1	1
F_5:7/10に雨	0	0	0	0	0	0	0	0
予測の当否	1	0	1	0	1	0	1	

たとえば、上のようこ、全ての陽で雨が降るという予測では、捕捉率は 1.0 となるが、的中率は、4/7=0.57 に下がる。両方を上げるのが難しいことが多い。、捕捉率 1.0、的中率 1.0 はあり得るが、これを実現できない場合が多い。たとえば、コイン投げのような確率的な現象では補足率、1.0、的中率 1.0 という予測が不可能であることは自明である。

逆に言えば、原因と結果が決定論的に結び付けられている現象で、原因が把握できる場合に限って補足率 1.0、的中率 1.0 の予測が可能となる。

捕捉率、的中率がともに1に近い予測を「科学的」と感じる人が多いように思われるが、捕捉率、予測率の両者が正確に見積もられている予測を科学的予測というべきである。

10

確率予測の場合

たとえば、「来週立川で少なくとも一回雨が降る確率は30%」というような予測が、確率予測である。

「来週」を過ぎても、この予測が当たっているかいないかは分からない。

しかし、毎週こう言い続けていて、100週過ぎてから、そのうちの30週で雨が観測されていたら、この確率予測方式が「当たっている」と言っていいような気がする。1000週過ぎて300週だったらもっと安心して当たっていると言えるだろう。では10週中の3週だったらどうだろう？予測の当否を判定していいのはどれくらいデータが溜まった時点なのだろう？

また、ずっと「確率30%」と言い続けるということはないはずである。関東だったら梅雨時に70%にあげたり、冬に5%に下げたりしたい。そんなときにどうしたらいいのだろう？

確率予測を評価する場合にはこれらのことを考慮する必要がある。

11

確率予測の評価

確率予測は、地域と時間の範囲を限ってその範囲で雨が降る確率という形で述べられるものだけを考えるものとする。例えば、「立川市で7月10日までに雨が降る確率が70%」というような予測、あるいは「東京都で7月10日までに雨が降る確率が90%」のようなものである。

この形の予測の評価にあたっては雨の有無が観測された各時点における単位面積あたり一日あたりの確率の対数尤度によるのが妥当である。たとえば「7/10までに東京都で雨が降る確率がP」という予測の妥当性を対数尤度を使って評価する場合、観測は一日ごと・地点ごとに得られるため、尤度を計算するには、予測確率Pを一日あたり・単位面積あたりの予測確率 p に変換する必要がある。そのためには、7/10までの日数 n と、東京都の面積 a を使って、以下の方法で p を計算するのが妥当である。p が求められたら、雨が降った場合には $\log p$ 、雨が降らなかった場合には $\log(1-p)$ という「得点」を与え、この得点が高い予測が良い予測とする。

1. 1日目は $p = P/(n \times a)$
2. 1日目に雨が降らなかった場合の2日目は $p = P/((n-1) \times a)$
3. 2日目までに雨が降らなかった場合の3日目は $p = P/((n-2) \times a)$
4. 等々
5. n 日目までに雨が降った翌日以降は、別個の予測をたてない限り予測なしの状態になるものとする。

12

対数尤度関数

確率 p で起きると予測されたことが、起きたとき、起きなかったとき、この予測を評価する値として、対数尤度関数が知られている。

問題とされていることが起きたときに 値1 をとり、起きなかったときに値 0 を取る変数を x で表すことにすると、対数尤度関数は p と x の関数として

$$f(p,x) = x\log p + (1-x)\log(1-p)$$

という形で定義される。$x = 1$ の場合、 この関数は p が 1.0 の時、0.0 という値をとり、1.0 から離れて0.0 に近づくに従って限りなくマイナス無限大に近づいていく。逆に、$x = 0$ の場合、 この関数は p が 0.0 の時、0.0 という値をとり、0.0 から離れて1.0 に近づくに従って限りなくマイナス無限大に近づいていく。つまり、この関数は、p が 1.0 に近いにも関わらす $x = 0$ ということが起きたとき、p が 0.0 に近いにも関わらす $x = 1$ ということが起きたときにペナルティを課すという働きを持っているので、p の「性能」を評価する量として使えるのである。、

ある1回かぎりの事象に関する確率予測の評価は対数尤度関数を使えば十分である.

13

「予測方式」の評価

一連の事象に対して、各事象それぞれに確率予測を対応させるなんらかの「予測方式」がある場合、この予測方式を全体として評価することが必要になる場合がある。

n 個の事象、$x_1, x_2, \ldots, x_n$ のそれぞれに対して、予測方式 A が、確率予測 $p_1, p_2, \ldots, p_n$ を対応させ、予測方式 B が、確率予測 $q_1, q_2, \ldots, q_n$ を対応させるような場合に対数尤度が問題を起こす可能性があるのである。直観的には、各事象ごとの予測に対する「ペナルティ」の総和を方式に対するペナルティと考えればいいように思われるがこれはうまくいかない。

たとえば、4 つの事象に関して、予測方式 A が確率 (0.2, 0.1, 0.8, 1.0) を割り当て。予測方式 B が確率 (0.2, 0.9, 0.8, 1.0) を割り当てるものとする。2番目の事象の予想確率がちがうのである。このときの「事実」が (0,1,1,0) であったとすると、2番目の事象に関しては 予測方式の方がより「当たっていた」というべきであり、予測方式 A より B の方が優れているとみられるが、4つの確率の対数尤度の和は、いずれの方式でもマイナス無限大となってしまい、大小比較ができない。確率予測に、確率 1.0 や、0.0 が含まれていると、それが「外れ」たときのペナルティが無限大になってしまうのである。

$p = 0.0$ あるいは 1.0 におけるの値がマイナス無限大にならないように関数f の定義を変えられればこのような問題は生じないが、そのような関数は知られていない。

14

確率予測方式の比較と選択

確率予測方式を固定するとサイズ n のデータをつぎの2つのサブセットに分類できる

1. 有限値の対数尤度が求められるデータ
2. 対数尤度がマイナス無限大になるデータ

それぞれのサブセットを、S_1 、S_2 と表すことにすると、$L(=S_1$から求められる対数尤度)と, N $(=S_2$ に含まれるデータの数) の組 (L,N) が予測方式の性能を集約的に表現する指標となる。

適当な重み w を使えば、

$$L+w\times N$$

が1次元指標となり、この値の大小で方式選択ができるが、w を客観的に定める方法は存在しない。複数の予測方式の比較選択は、予測の目的や S_2 の内容を吟味した上での主観的判断に委ねられることになる。

15

数値例 1

港区面積	20.37㎢
立川市面積	24.36㎢
東京都面積	2188㎢

日付	事実 立川	事実 港区	確率予測 A	単位面積あたり1日分予測確率 立川	単位面積あたり1日分予測確率 港区
7月1日	0	0	7月10日までに立川に雨が降る確率90%	3.69E-03	0.0
7月2日	0	0		4.11E-03	0.0
7月3日	0	1		4.62E-03	0.0
7月4日	0	0		5.28E-03	0.0
7月5日	0	0		6.16E-03	0.0
7月6日	1	1		7.39E-03	0.0
7月7日	0	0	7月16日までに東京のどこかで雨が降る確率100%	4.57E-05	4.57E-05
7月8日	0	0		5.08E-05	5.08E-05
7月9日	0	0		5.71E-05	5.71E-05
7月10日	0	0		6.53E-05	6.53E-05
7月11日	0	0		7.62E-05	7.62E-05
7月12日	0	1		9.14E-05	9.14E-05
7月13日	1	0	7月22日までに港区で雨が降る確率70%	0.0	3.44E-03
7月14日	0	0		0.0	3.82E-03
7月15日	0	0		0.0	4.30E-03
7月16日	0	0		0.0	4.91E-03
7月17日	0	0		0.0	5.73E-03
7月18日	0	0		0.0	6.87E-03
7月19日	1	0		0.0	8.59E-03
7月20日	0	0		0.0	1.15E-02
7月21日	0	1		0.0	1.72E-02
7月22日	0	0	7月31日までに港区で雨が降る確率0%	0.0	0.00E+00
7月23日	1	0		0.0	0.00E+00

L=	−18.35
N=	5

(注)確率予測の場合の事実予測対照表はこのような形になる。
地域を限った予測は、地域外での生起確率をゼロと言うに等しい。
地域と期間を限定したピンポイントの予測の方が価値が高い
確率予測の「有効期限」は、事象の有無が確定するまでである。

16

数値例2:ピンボケ予測

事実			予測方式 B	単位面積あたり1日分予測確率	
日付	立川	港区		立川	港区
7月1日	0	0	一様予測	7.95E-05	7.95E-05
7月2日	0	0		7.95E-05	7.95E-05
7月3日	0	1		7.95E-05	7.95E-05
7月4日	0	0		7.95E-05	7.95E-05
7月5日	0	0		7.95E-05	7.95E-05
7月6日	1	1		7.95E-05	7.95E-05
7月7日	0	0		7.95E-05	7.95E-05
7月8日	0	0		7.95E-05	7.95E-05
7月9日	0	0		7.95E-05	7.95E-05
7月10日	0	0		7.95E-05	7.95E-05
7月11日	0	0		7.95E-05	7.95E-05
7月12日	0	1		7.95E-05	7.95E-05
7月13日	1	0		7.95E-05	7.95E-05
7月14日	0	0		7.95E-05	7.95E-05
7月15日	0	0		7.95E-05	7.95E-05
7月16日	0	0		7.95E-05	7.95E-05
7月17日	0	0		7.95E-05	7.95E-05
7月18日	0	0		7.95E-05	7.95E-05
7月19日	1	0		7.95E-05	7.95E-05
7月20日	0	0		7.95E-05	7.95E-05
7月21日	0	1		7.95E-05	7.95E-05
7月22日	0	0		7.95E-05	7.95E-05
7月23日	1	0		7.95E-05	7.95E-05

L=	−75.52
N=	0

17

数値例3:神様予測

事実			予測方式 C	単位面積あたり1日分予測確率	
日付	立川	港区		立川	港区
7月1日	0	0	完全予測	0	0
7月2日	0	0		0	0
7月3日	0	1		0	1
7月4日	0	0		0	0
7月5日	0	0		0	0
7月6日	1	1		1	1
7月7日	0	0		0	0
7月8日	0	0		0	0
7月9日	0	0		0	0
7月10日	0	0		0	0
7月11日	0	0		0	0
7月12日	0	1		0	1
7月13日	1	0		1	0
7月14日	0	0		0	0
7月15日	0	0		0	0
7月16日	0	0		0	0
7月17日	0	0		0	0
7月18日	0	0		0	0
7月19日	1	0		1	0
7月20日	0	0		0	0
7月21日	0	1		0	1
7月22日	0	0		0	0
7月23日	1	0		1	0

L=	0.00
N=	0

18

カテゴリカル予測との関係

日付	事実 立川	事実 港区	予測方式 D	単位面積あたり1日分予測確率 立川	単位面積あたり1日分予測確率 港区
7月1日	0	0		0	0
7月2日	0	0		1	0
7月3日	0	1		0	1
7月4日	0	0		0	0
7月5日	0	0		0	0
7月6日	1	1		0	1
7月7日	0	0		0	0
7月8日	0	0		0	1
7月9日	0	0		1	0
7月10日	0	0		0	0
7月11日	0	0		0	0
7月12日	0	1	カテゴリカル予測	0	1
7月13日	1	0		1	0
7月14日	0	0		0	0
7月15日	0	0		0	0
7月16日	0	0		0	0
7月17日	0	0		0	0
7月18日	0	0		0	1
7月19日	1	0		0	0
7月20日	0	0		0	0
7月21日	0	1		0	1
7月22日	0	0		0	0
7月23日	1	0		1	0

L=	0.00
N=	6

		事実	
		0	1
予測	0	34	2
	1	4	6

的中率 ＝ 0.6
補足率 ＝ 0.75

カテゴリカル予測は確率予測のひとつであるが、逆は成り立たない。カテゴリカル予測についても (L, N) 指標を計算できる。

19

対数尤度大小の判定基準

$$L = \sum_{i \in S_1} (x_i \log p_i + (1 - x_i) \log(1 - p_i))$$

は、$x_i = 1$ となる確率が p_i で与えられる($i = 1,2, \dots, n$) ときにデータ S_1 が得られる確率の対数なので、e^L は確率として大小を判断できるはずであるが、通常非常に小さい値となるので「常識」では得られた値が大きいのかどうか判断できない。

基準となる予測方式を定めて、その基準予測方式の対数尤度との比較という形に持ち込むのがおそらく、唯一可能な方法であると思われる。

数値例1の $L = -18.35$ は数値例2の $L = -75.52$ と比べてずっとよい値であると言える。ただし、数値例2では $N = 0$ であるのに対して、数値例1では $N = 5$ であり、予測方式 A は当たるときはいいが、外すときは大外しをする予測、予測方式 B は、シャープな結果を出さないが、大間違いはしないという予測ということになる。どちらを好むか人によるだろう

20

客観的事実予測対照表構成法

予測方式の正確な評価指標を得るためには、事実予測対照表を追試可能な形で用意することが重要である。事実予測対照表に関して答えを用意しておくべき質問として

Q1:都合のよい事実だけ拾い出したのではありませんか？
Q2:予測が「後出し」なのではありませんか？

の2つがある。

Q1 に関しては、事実選択の基準を明確にして事実の選択に恣意性がないことを示めすことで答えるのが best である。

Q2 に関しては、「予測」が予測時点以前の情報だけに基づくものであることを示す形で答える必要がある。つまり、「予測」が客観的なデータに基く決定論的、必然的に構成されるものである必要がある。

21

最後に. . .

言ったことが事実を言い当ててているか、という形の問題は予測の場合に限られない。予測でない場合の例として、たとえば、「邪馬台国は九州にあったにちがいない」というがあげられる。この場合にも「当たっているか否か」が問題になる。

このメモは「予測が事実を言い当てているか否かを統計的に検証する方法」についてのメモであるが、「予測」ということばをすべて「命題」と読み替えれば「命題が事実を言い当てているか否かを統計的に検証する方法」についてのメモとなる。

22