

CTR 予測モデルの評価に AUC や log-loss は適切か?

Do the AUC and log-loss evaluate CTR prediction models properly?

片桐 智志 *¹

Satoshi Katagiri

*¹株式会社ファンコミュニケーションズ

F@N Communications, Inc.

Click-through rate (CTR) prediction is one of the most important task for web advertising platform companies. However, CTR prediction is a non-standard machine learning task, so conventional metrics, for example, area under the Receiver Operating Characteristic curve (AUC), and log-loss, a.k.a. cross-entropy, and so on, can be improper. Our target is develop a new metrics for CTR prediction. In this article, we state the drawbacks of such conventional metrics and perspective of a metric based on the calibration plot approach.

1. はじめに

広告のリアルタイム入札システム (RTB) は、消費者が広告枠のあるウェブページなどを閲覧するたびに、広告のリクエストがなされ、どの広告主がページの広告枠に出稿する権利についてオークションを自動で行い、出稿する広告を決定するシステムである。ここでのオークションは多くの場合、2 番目に高い額を提示した入札者 (広告主) が落札するという二位価格オークション (second-price auction) を採用している。理論上、二位価格オークションでは入札者の私的価値 (private value) と一致する価格で入札する “truth-telling 戦略” が支配戦略である [Krishna 10]。クリックに対して課金される料金体系の場合、広告オークションにおける私的価値とは、1 回の広告表示 (インプレッション) に対して消費者がどれくらいの確率でクリックするかである [田頭 13]。そのため、この確率を正確に見積もることは、RTB が顧客にとって有益であることに直結する。RTB プラットフォームを持つ多くの企業では、適切に入札のプライシングができるように、広告表示に対する click-through 率 (CTR) を機械学習によって予測する方法を研究または導入しており、その先行研究だけでも枚挙に暇がない。

CTR 予測を機械学習の問題として見ると 2 値分類問題とみなせるため、多くの研究では予測モデルの評価に area under the Receiving Operator Characteristic curve (AUC) や、対数損失 (交差エントロピー) が用いられている。しかし、標準的な機械学習の問題とは異なり、求められているのは予測値が正解ラベルにどれだけの的中しているかというよりも、広告リクエスト単位の予測確率がどれだけ適切であるか、という点である。第 2 節で詳細に述べる先行研究により、従来モデルの評価によく用いられてきた AUC や対数損失だけでは適切に評価できないことがわかっている。本研究では、これらを踏まえ、カリブレーションの指標として従来から提案されている [DeGroot 83] のカリブレーションプロットや [Caruana 04] の CAL とその問題点についても考察する。

2. 先行研究のサーベイ

Microsoft の研究チームによれば [Yi 13], AUC や対数損失などと比較して NE の性質について言及しており *¹, AUC あるいは対数損失 (または正規化エントロピー, 以下 NE) だけでは評価指標として完全ではないとしつつも, CTR 予測精度の評価問題に適した方法についての結論を述べていない。[He 14] では, AUC と NE を利用しているが, NE だけではデータ全体でみたクリック率と予測値のクリック率が必ずしも近似できていないとして, NE に加えてデータ全体のクリック率と予測値から計算できる期待クリック率の一致, という指標も重視している。

[Gail 05, Cook 07] では, 分類モデルの出力する予測確率を将来の病気の発病リスクとみなした場合について言及がある。疫学分野では, ラベルに分類される確率を正しく予測すること, 正例に予測される場合のモデルの条件分布とそうでない場合の条件分布の差別化, の 3 種類が要求される場合のいずれもありえるため, [Gail 05] ではそれぞれ, *accuracy*, *calibration* (カリブレーション), *discrimination*, と定義している。医療を例にすると, 現時点で病気が疑われる患者を診断することは, 予測確率よりも陽性と陰性をどれだけはっきり区別できるかが重要な *discrimination* のタスクであり, 一方でまだ発症していない人が将来発症する可能性や, 予後の死亡率などを知りたい (いわゆる *prognostic studies*) 場合は *calibration* のタスクとなる。従来使われている AUC は *discrimination* に対応し, 対数損失や NE は *accuracy* に対応する。しかし, CTR 予測について重要となるのは, カリブレーションである。

2.1 AUC の問題点

Microsoft の研究チームによれば [Yi 13], AUC は予測確率の大きさそのものを見ないことが問題であるとしている。AUC は予測確率の絶対値ではなく, 大きさをソートした際の順序を評価していることが問題であり, 実際にデータ全体のクリック割合とクリックの予測頻度が一致しないようなモデルであっても AUC が大きくなることもある。よって, 予測確率が異なる値でも AUC は変化しない。単純な例として, $y \in \{0, 1\}$ のラベルに対して 3 種類の予測モデルが, それぞれ予測確率 $\hat{\pi}_A$, $\hat{\pi}_B$, $\hat{\pi}_C$ を表 1 のように出力しているとする。このとき, $\hat{\pi}_B$ は, $\hat{\pi}_A$

連絡先: 片桐智志, 株式会社ファンコミュニケーションズ サービス開発部情報科学技術研究所, s_katagiri@fancs.com

*¹ 正確には, 相対情報ゲイン (RIG) についての議論だが, $RIG = 1 - NE$ という関係が成り立つため議論内容は NE に容易に転用できる。

y	0	0	0	1	1	1
$\hat{\pi}_A$	0.1	0.2	0.5	0.5	0.6	0.8
$\hat{\pi}_B$	0.2	0.3	0.6	0.6	0.7	0.9
$\hat{\pi}_C$	1	2	5	5	6	8

表 1 AUC が変わらない例

の各値に 0.1 を足したもので、 $\hat{\pi}_C$ は $\hat{\pi}_A$ を 10 倍にしたものだが、それぞれの AUC は全く同じになる。

AUC はそもそも discrimination を評価する指標であり、疾病リスクモデルの変数選択を例に、尤度比統計量やカイ二乗統計量と AUC とで反応の大きさの違いが指摘されている [Cook 07]*2。加えて、データの分布しだいで、AUC の事実上の最大値が変化するという問題も、具体例を示して指摘されている [Diamond 92, Gail 05]。

2.2 対数損失の問題点

対数損失 (交差エントロピー) も広く使われている指標であるが、カリブレーションを評価するには問題がある。

たとえば、真の確率 π とラベル y のペアについて、 $\pi > 0.5$ ならば $y = 1$ 、そうでなければ $y = 0$ となる場合を考える。このとき、

$$(\pi_1, y_1) = (0.4, 0),$$

$$(\pi_2, y_2) = (0.6, 1)$$

という 2 点だけのデータあるとする。このとき、カリブレーションの観点からすれば、予測モデルは真の確率に近い値を出力するのが望ましいため、 $\hat{\pi}_1 = 0.4, \hat{\pi}_2 = 0.6$ を出力するようなモデルが最も望ましい。このとき、対数損失は約 0.51 となる。一方で、 $\hat{\pi}_1 \rightarrow 0, \hat{\pi}_2 \rightarrow 1$ のときに対数損失は明らかにこれより小さくなる。例えば $\hat{\pi}_1 = 0.1, \hat{\pi}_2 = 0.9$ のときに約 0.11 となる。よって、対数損失の小さなモデルほどカリブレーションも良いとは限らない。

正規化エントロピー (NE) は、対数損失をデータの正例割合に基づく対数損失で割った指標であるため、CTR 予測のような不均衡データに対して対数損失よりも優れているとされる [Yi 13, He 14]。しかしながら、対数損失を正規化しただけの指標であるため、カリブレーションを評価できないという問題は NE に対してもそのまま当てはまる。

[Brier 50] による、回帰問題で使われる平均二乗誤差 (MSE) を分類問題にそのまま適用した Brier スコアもまた、同様の問題がある。

2.3 カリブレーションプロットと CAL

一方で、疫学分野では、カリブレーションの確認方法として、Hosmer-Lemeshow 検定が提案されている [Hosmer 89, Hosmer 80]。これはデータをいくつかのグループに分割し、それぞれでカイ二乗統計量を計算した和でカリブレーションがなされているかを検定する方法である。今回我々が求めているのは、AUC や対数損失に変わる相対的な指標であるので採用できないが、[Caruana 04] では、よく似たアイデアとして、カリブレーションプロット [DeGroot 83] の結果に対して平均絶対誤差 (MAE) を計算する CAL を紹介している。

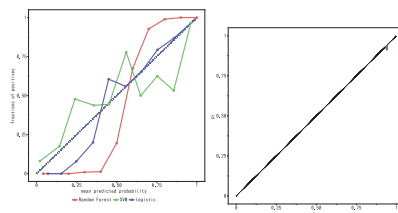


図 1 典型的なカリブレーション・プロットの例 (左) と一律同じ出力をする予測モデルのカリブレーションプロットの例 (右)

3. 適切な指標はどうあるべきか

AUC や NE の問題点に対して [He 14] は、カリブレーションの要件を満たすように、データのクリック頻度と予測確率に基づく期待値とが一致しているかについても確認する方法を提案している。しかしこれは、膨大なデータ全体でのクリック数と予測数の一致を見ているだけであり、1 つ 1 つの広告リクエストに対する予測 CTR の精度を保証するものではない。本研究では、良いカリブレーションの定義として、[Gail 05] で述べられているものを採用する。予測モデルが完全にカリブレーションされている (perfectly calibrated) とは、特徴量 x に対して予測確率を出力する予測関数 $\hat{f}(x)$ が、任意の x に対して π の条件付き期待値に等しい、つまり以下を満たすことを言う。

$$\hat{f}(x) := E[\pi | x] = \int \pi dG(\pi | x) \quad (1)$$

ここで、 $G(\pi | x)$ は π の条件確率密度関数である。モデルが良くカリブレーションされているかは、実際の確率とモデルの出力する予測確率の誤差がどれだけ小さいかで判断する。しかし、例えば仮に平均二乗誤差 (MSE) $N^{-1} \sum_{i=1}^N (\hat{f}(x_i) - \pi_i)^2$ で評価すると、真の確率 π_i は観測できない。ここで π_i を観測可能なラベル y_i に置き換えると、先述の対数損失や Brier スコアの問題が発生する。この点、カリブレーションプロットや CAL は、この問題に対して、データを分割したサブグループ内での頻度を真の確率の近似として使用していると解釈できる。

4. 考察

しかしながら、CAL には次のように少なくとも 2 点の問題が考えられる。(1) 観測点数がサブグループごとに異なるため、相対的に点数の多いサブグループの当てはまりが過小評価される傾向にある、(2) 出力されるすべての予測確率が同じ値である場合、分位数によるサブグループによる分割ができない。

(1), (2) いずれも、分位数ではなく、観測点数が同等になるように等分割するという方法が考えられる。しかし、(2) の場合は正例の多い不均衡データにおいて一律で大きな予測確率を出力するモデルに対しては図 1 のようなカリブレーションプロットを描き、CAL が良い値を示す可能性があり、NE で解消された問題が再び浮上する。

5. 結論と課題

本研究では、予測確率の精度を求める CTR 予測において、AUC や 対数損失、NE だけでは評価に不十分であるということを示した。続いてカリブレーションプロットに基づく CAL について考察し、カリブレーションの指標としては問題点が残ること示した。CAL をより適切な指標へと改善することは今後の課題であるが、サブグループの分割方法が重要になると予想

*2 [Cook 07] では AUC を c-統計量と呼んでいる。C は concordance の略である。

できる。我々は CAL をに対して取り組んでいるが、最終的な目的は適切な評価指標の考案だけでなく、カリブレーションの良いモデルの改良方法を開発することである。

参考文献

- [Brier 50] Brier, G. W.: VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY, *Monthly Weather Review*, Vol. 78, No. 1, pp. 1–3 (1950)
- [Caruana 04] Caruana, R. and Niculescu-Mizil, A.: Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria, in *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*, pp. 69–78, Seattle, WA, USA (2004), ACM Press
- [Cook 07] Cook, N. R.: Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction, *Circulation*, Vol. 115, No. 7, pp. 928–935 (2007)
- [DeGroot 83] DeGroot, M. H. and Fienberg, S. E.: The Comparison and Evaluation of Forecasters, *The Statistician*, Vol. 32, No. 1/2, pp. 12–22 (1983)
- [Diamond 92] Diamond, G. A.: What Price Perfection? Calibration and Discrimination of Clinical Prediction Models, *Journal of Clinical Epidemiology*, Vol. 45, No. 1, pp. 85–89 (1992)
- [Gail 05] Gail, M. H. and Pfeiffer, R. M.: On Criteria for Evaluating Models of Absolute Risk, *Biostatistics*, Vol. 6, No. 2, pp. 227–239 (2005)
- [He 14] He, X., Bowers, S., Candela, J. Q. n., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., and Herbrich, R.: Practical Lessons from Predicting Clicks on Ads at Facebook, in *Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining - ADKDD'14*, pp. 1–9, New York, NY, USA (2014), ACM Press
- [Hosmer 80] Hosmer, D. W. and Lemeshow, S.: Goodness of Fit Tests for the Multiple Logistic Regression Model, *Communications in Statistics - Theory and Methods*, Vol. 9, No. 10, pp. 1043–1069 (1980)
- [Hosmer 89] Hosmer, D. W. and Lemeshow, S.: *Applied Logistic Regression*, Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics, Wiley, New York (1989), OCLC: 19514573
- [Krishna 10] Krishna, V.: *Auction Theory*, Elsevier, Academic Press, Amsterdam, 2. ed edition (2010), OCLC: 845563467
- [Yi 13] Yi, J., Chen, Y., Li, J., Sett, S., and Yan, T. W.: Predictive Model Performance: Offline and Online Evaluations, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, p. 1294, New York, New York, USA (2013), ACM Press
- [田頭 13] 田頭 幸浩, 山本 浩司, 小野 真吾, 塚本 浩司, 田島 玲: オンライン広告における CTR 予測モデルの素性評価, 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), 郡山市, 福島県 (2013)

補遺: 各指標の定義

AUC

ROC 曲線の下側の面積である。

平均自乗誤差 (MSE) ・ Brier スコア:

平均自乗誤差は、予測値と真値の差の 2 乗平均で、(2) のように定義される。2 値分類に限定すれば、MSE と [Brier 50] による Brier スコアが同一のものであるのは明らかである。

$$\text{Brier} := \text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{\pi}_i)^2 \quad (2)$$

対数損失 (交差エントロピー) ・ 正規化エントロピー:

対数損失は、(3) で定義される。

$$\text{logloss} := - \frac{1}{N} \sum_{i=1}^N [y_i \ln \hat{\pi}_i + (1 - y_i) \ln(1 - \hat{\pi}_i)] \quad (3)$$

正規化エントロピー (NE) は、対数損失を、データ全体の正例の割合に対する対数損失で除したものであり、(4) のように定義される。ラベル数が極端に不均衡である場合、簡単に低い対数損失を算出できる問題があるが、NE はデータの割合で調整することでこの問題を解消している [He 14]。

$$\begin{aligned} \text{NE} &:= \frac{\text{logloss}}{- (\bar{p} \ln \bar{p} + (1 - \bar{p}) \ln(1 - \bar{p}))}, \\ \bar{p} &:= \frac{1}{N} \sum_{i=1}^N y_i \end{aligned} \quad (4)$$

CAL:

予測値を十分位数を区切りに B 個に分割する。 k 番目のビンに属する集合が b_k , $\#b_k$ はその要素数で、 $\bar{p}_k$ はそこに属する正例ラベルの頻度とすると、CAL は、(5) のように定義される [Caruana 04]。

$$\text{CAL} := \frac{1}{N} \sum_{k=1}^B \left| \bar{p}_k - \frac{1}{\#b_k} \sum_{i \in b_k} \hat{\pi}_i \right| \quad (5)$$