

オンライン広告におけるスパース性と遅れコンバージョンを考慮した予測モデル

Conversion Prediction in Online Advertising Based on Sparsity and Delayed Feedback

今井 優作

Yusaku Imai

株式会社サイバーエージェント AI Lab

AI Lab, CyberAgent, Inc.

In this paper, we propose a model for predicting the conversion rate (CVR) in online advertising. Some of the issues encountered in CVR prediction include 1) the sparsity of categorical features in the collected data and 2) the lag between when the user clicks the ad and takes any following actions. Our proposed model addresses these problems by modeling factorized interactions and conversion delay. We demonstrate the effectiveness of our proposed model with experiments using real traffic logs.

1. はじめに

近年、インターネットやスマートフォンの普及に伴い、オンライン広告費が年々拡大を続けており、2017年時点でテレビ広告費に次ぐ規模となっている。オンライン広告における効果測定のための重要な指標のひとつにコンバージョン率 (CVR; conversion rate) がある。ここで、コンバージョンとは、商品購入や資料請求など、Web サイト上から獲得できる最終成果のことを指す。CVR 予測の精度を高めることで、限られた予算の中で広告効果を最大化するよう、適切な広告の配信が可能となる。

CVR 予測の課題として、1) 素性として質的データを多く含むためスパースなデータが生じる、2) 広告をクリックしてからコンバージョンするまでに長い時間を要する、ことが挙げられる。本稿では、上記の問題を解決するための予測モデルを提案し、従来手法よりも高い精度で CVR 予測が可能であることを示す。

2. 関連研究

CVR 予測に用いられる代表的なモデルのひとつにロジスティック回帰 (LR; logistic regression) がある [Lee 12, Chapelle 14b]。

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-\phi_{\text{LR}}(\mathbf{w}, \mathbf{x}))}$$

$$\text{with } \phi_{\text{LR}}(\mathbf{w}, \mathbf{x}) = w_0 + \sum_{j=1}^n w_j x_j \quad (1)$$

ここで、 n は素性数、 $\mathbf{w} = (w_0, w_1, \dots, w_n)$ は重みベクトル、 $\mathbf{x} = (1, x_1, \dots, x_n)$ は素性ベクトルを表す。

ロジスティック回帰は一般化線形モデルの一種であり、素性数に対してスケールが容易、モデルの学習を高速に行える、というような特徴を持つ。一方で、ロジスティック回帰は説明変数間の交互作用を考慮できない問題がある。

Rendle は交互作用のモデリング方法として *Factorization Machines* (FM) を提案している [Rendle 10]。次元数 $d = 2$ の FM は

連絡先: 今井優作, 株式会社サイバーエージェント,
imai_yusaku_xa@cyberagent.co.jp

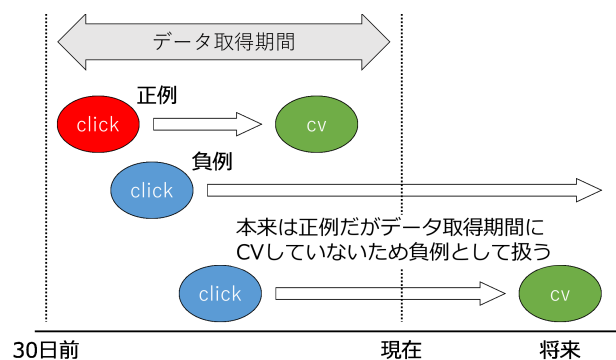


図 1: 遅れコンバージョンの概要図

$$\phi_{\text{FM}}(\mathbf{w}, \mathbf{x}) = w_0 + \sum_{j=1}^n w_j x_j + \sum_{j_1=1}^n \sum_{j_2=j_1+1}^n \langle \mathbf{v}_{j_1}, \mathbf{v}_{j_2} \rangle x_{j_1} x_{j_2} \quad (2)$$

と定義される。ここで、 $\langle \cdot, \cdot \rangle$ は 2 つのベクトルの内積を表す。

$$\langle \mathbf{v}_{j_1}, \mathbf{v}_{j_2} \rangle := \sum_{f=1}^k v_{j_1, f} \cdot v_{j_2, f} \quad (3)$$

また、 \mathbf{v}_j は素性 x_j に対する k 次元の重みベクトルを表す。

FM では交互作用項を低次元の重みベクトルの内積に分解して表現するため、スパースなデータに対しても適切なパラメータ推定が可能となる。

3. Delayed Feedback Model

オンライン広告において、広告をクリックしてからコンバージョンするまでに長い時間を要する問題がある。本来ならば、コンバージョンをすれば正例、しなければ負例として予測モデルの学習に用いられるが、遅れコンバージョンがデータ取得期間よりもあとになった場合は負例として扱われてしまう。遅れコンバージョンの概要を図 1 に示す。

Chapelle は遅れコンバージョンを考慮した CVR 予測モデルとして *Delayed Feedback Model* (DFM) を提案している [Chapelle 14a] . DFM はコンバージョン確率 $P(C|X)$ とコンバージョンの遅れ分布 $P(D|X, C = 1)$ の 2 つのパラメトリックな確率モデルで構成される .

$$\begin{aligned} P(C = 1, D = d|X = \mathbf{x}) \\ = P(C = 1|X = \mathbf{x}) \cdot P(D = d|X = \mathbf{x}, C = 1) \end{aligned} \quad (4)$$

ここで, X は素性ベクトル, D はクリックからコンバージョンまでの遅延時間, $C \in \{0, 1\}$ は広告をクリックしたユーザがコンバージョンするかどうかを表す変数である . (4) 式の第一項は

$$\begin{aligned} P(C = 1|X = \mathbf{x}) = p(\mathbf{x}) \\ \text{with } p(\mathbf{x}) = \frac{1}{1 + \exp(-\phi_{\text{LR}}(\mathbf{w}_c, \mathbf{x}))} \end{aligned} \quad (5)$$

のロジスティック回帰, 第二項は

$$\begin{aligned} P(D = d|X = \mathbf{x}, C = 1) = \lambda(\mathbf{x}) \exp(-\lambda(\mathbf{x})d) \\ \text{with } \lambda(\mathbf{x}) = \exp(\phi_{\text{LR}}(\mathbf{w}_d, \mathbf{x})) \end{aligned} \quad (6)$$

の指数分布であり, モデルのパラメータは 2 つの重みベクトル $\mathbf{w}_c, \mathbf{w}_d$ となる .

(4) 式において, クリックから時間 e が経過した時点でコンバージョンしている確率は以下のように導出できる .

$$P(Y = 1, D = d|X = \mathbf{x}, E = e) = \lambda(\mathbf{x}) \exp(-\lambda(\mathbf{x})d)p(\mathbf{x}) \quad (7)$$

ここで, E はクリックからの経過時間, $Y \in \{0, 1\}$ は経過時間 E の時点でコンバージョンしているかどうかを表す変数であり, $C = 1$ かつ $D \leq E$ のとき $Y = 1$ となる .

また, クリックから時間 e が経過した時点でコンバージョンしていない確率は以下の式で求めることができる .

$$\begin{aligned} P(Y = 0|X = \mathbf{x}, E = e) \\ = P(Y = 0, C = 0|X = \mathbf{x}, E = e) \\ + P(Y = 0, C = 1|X = \mathbf{x}, E = e) \end{aligned} \quad (8)$$

(8) 式の第一項は

$$P(Y = 0, C = 0|X = \mathbf{x}, E = e) = 1 - p(\mathbf{x}), \quad (9)$$

第二項は

$$P(Y = 0, C = 1|X = \mathbf{x}, E = e) = p(\mathbf{x}) \exp(-\lambda(\mathbf{x})e) \quad (10)$$

となる .

DFM ではクリックからコンバージョンするまでの遅れ度合いを統計的にモデリングするため, 正例か負例かというラベル情報がまだ与えられていないデータを学習に用いることが可能となる . オンライン広告では, キャンペーンの追加や停止が短期間で発生することからデータの鮮度がモデルの精度に大きな影響を与えるため, 最新のデータを学習に用いられることは CVR 予測において非常に優位となる .

4. 提案法

4.1 Delayed Feedback Factorization Machines

CVR 予測の課題として, 1) 素性として質的データを多く含むためスパースなデータが生じる, 2) 広告をクリックしてからコンバージョンするまでに長い時間を要する, ことが挙げられる . 本稿では, 上記の問題を解決するための予測モデルとして, 3. で示した CVR 予測モデルに対して 2. で示した相互作用項を組み合わせた非線形な予測モデル *Delayed Feedback Factorization Machines* (DFFM) を提案する .

DFFM ではコンバージョン確率 $P(C|X)$ とコンバージョンの遅れ分布 $P(D|X, C = 1)$ を (2) 式を基に

$$\begin{aligned} P(C = 1|X = \mathbf{x}) = p(\mathbf{x}) \\ \text{with } p(\mathbf{x}) = \frac{1}{1 + \exp(-\phi_{\text{FM}}(\mathbf{w}_c, \mathbf{x}))}, \end{aligned} \quad (11)$$

$$\begin{aligned} P(D = d|X = \mathbf{x}, C = 1) = \lambda(\mathbf{x}) \exp(-\lambda(\mathbf{x})d) \\ \text{with } \lambda(\mathbf{x}) = \exp(\phi_{\text{FM}}(\mathbf{w}_d, \mathbf{x})) \end{aligned} \quad (12)$$

と表す .

4.2 モデルの学習

m 個のデータセット $(\mathbf{x}_i, y_i, e_i, d_i), i = 1, \dots, m$ が与えられたとき, 以下の最適化問題を解いてモデルのパラメータ $\mathbf{w}_c, \mathbf{w}_d$ を求める .

$$\arg \min_{\mathbf{w}_c, \mathbf{w}_d} L(\mathbf{w}_c, \mathbf{w}_d) + \frac{\mu}{2} (\|\mathbf{w}_c\|_2^2 + \|\mathbf{w}_d\|_2^2) \quad (13)$$

ここで, μ は正則化パラメータ, $\|\mathbf{w}\|_2$ は L2 ノルムである . また, L は negative log-likelihood を表し,

$$\begin{aligned} L(\mathbf{w}_c, \mathbf{w}_d) = & - \sum_{i, y_i=1}^m [\log p(\mathbf{x}_i) + \log \lambda(\mathbf{x}_i) - \lambda(\mathbf{x}_i)d_i] \\ & - \sum_{i, y_i=0}^m \log [1 - p(\mathbf{x}_i) + p(\mathbf{x}_i) \exp(-\lambda(\mathbf{x}_i)e_i)] \end{aligned} \quad (14)$$

となる .

パラメータ \mathbf{w}^* の推定に確率的勾配降下法 (SGD; stochastic gradient descent) を用いる . (14) 式における $\mathbf{w}_c, \mathbf{w}_d$ の勾配は以下のように導出できる .

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}_c} = & \sum_{i, y_i=1}^m -\frac{1}{p(\mathbf{x}_i)} \frac{\partial p(\mathbf{x}_i)}{\partial \mathbf{w}_c} \\ & + \sum_{i, y_i=0}^m \frac{1 - \exp(-\lambda(\mathbf{x}_i)e_i)}{1 - p(\mathbf{x}_i) + p(\mathbf{x}_i) \exp(-\lambda(\mathbf{x}_i)e_i)} \frac{\partial p(\mathbf{x}_i)}{\partial \mathbf{w}_c}, \end{aligned} \quad (15)$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}_d} = & \sum_{i, y_i=1}^m \left(d_i - \frac{1}{\lambda(\mathbf{x}_i)} \right) \frac{\partial \lambda(\mathbf{x}_i)}{\partial \mathbf{w}_d} \\ & + \sum_{i, y_i=0}^m \frac{p(\mathbf{x}_i)e_i \exp(-\lambda(\mathbf{x}_i)e_i)}{1 - p(\mathbf{x}_i) + p(\mathbf{x}_i) \exp(-\lambda(\mathbf{x}_i)e_i)} \frac{\partial \lambda(\mathbf{x}_i)}{\partial \mathbf{w}_d} \end{aligned} \quad (16)$$

ここで, $\frac{\partial p(\mathbf{x})}{\partial \mathbf{w}_c}$ と $\frac{\partial \lambda(\mathbf{x})}{\partial \mathbf{w}_d}$ は連鎖律により導出された項であり,

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{w}_c} = p(\mathbf{x})(1-p(\mathbf{x})) \cdot \frac{\partial}{\partial \mathbf{w}_c} \phi_{\text{FM}}(\mathbf{w}_c, \mathbf{x}), \quad (17)$$

$$\frac{\partial \lambda(\mathbf{x})}{\partial \mathbf{w}_d} = \lambda(\mathbf{x}) \cdot \frac{\partial}{\partial \mathbf{w}_d} \phi_{\text{FM}}(\mathbf{w}_d, \mathbf{x}) \quad (18)$$

となる. $\frac{\partial}{\partial \mathbf{w}} \phi_{\text{FM}}(\mathbf{w}, \mathbf{x})$ は以下のように導出できる.

$$\frac{\partial}{\partial \theta} \phi_{\text{FM}}(\mathbf{w}, \mathbf{x}) = \begin{cases} 1 & \text{if } \theta \text{ is } w_0 \\ x_j & \text{if } \theta \text{ is } w_j \\ x_j \sum_{j'=1}^n v_{j',f} x_{j'} - v_{j,f} x_j^2 & \text{if } \theta \text{ is } v_{j,f} \end{cases} \quad (19)$$

パラメータ \mathbf{w} の更新に AdaGrad を用いる [Duchi 11].

$$r \leftarrow r + g\mathbf{w}, \quad (20)$$

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\eta}{\sqrt{r}} g\mathbf{w} \quad (21)$$

ここで, η は学習率, $g\mathbf{w}$ は損失関数の勾配, r は勾配の二乗の累積和を表す.

上記のパラメータ更新を繰り返すことで, 提案モデルの学習が行われる. 素性ベクトル \mathbf{x} に対する CVR 予測は以下の式で求めることができる.

$$P(Y = 1 | X = \mathbf{x}, E = \ell) = p(\mathbf{x})(1 - \exp(-\lambda(\mathbf{x})\ell)) \quad (22)$$

ここで, ℓ はクリックに紐づくコンバージョンの計測期間であり, 広告主が任意に設定することができる. 計測期間を超過したコンバージョンは無効となるため, $D \leq \ell$ となる.

5. 実験

提案法の有効性を評価するための実験を行った. 実験に用いるデータセットおよび実験条件は [Chapelle 14a] と同一とした: テスト期間として 7 日間が存在し, 各テスト日に対して前 3 週間分のデータをモデルの学習に用いる. なお, 本データセットではコンバージョン計測期間 ℓ は 30 日間に固定されている.

遅れコンバージョンの扱い方として以下を定義する.

- NAIVE, データ取得期間よりもあとのコンバージョンは負例として扱う.
- DFM, データ取得期間よりもあとにコンバージョンする確率を (10) 式で計算する.
- ORACLE, データ取得期間に関わらずコンバージョンは正例として扱う, upper bound に該当する.

これらの機構に対して, (1) 式のロジスティック回帰 (LR) と (2) 式の Factorization Machines (FM) を適用し, 一標本 t 検定による有意差検定を行う. また, 評価尺度として average negative log-likelihood (NLL) と average area under the ROC curve (AUC) を用いる.

モデルの学習には SGD と AdaGrad を用い, 10-fold cross-validation による early-stopping を用いて収束判定を行った. また, ハイパーパラメータは, validation データに対する予測誤差が最小となった $k = 16$, $\mu = 2 \times 10^{-5}$, $\eta = 0.02$ とした. 実験結果を表 1, 2 に示す.

表 1: NLL の比較

	LR	FM	Diff
NAIVE	0.40637	0.40155	0.00482*
DFM	0.39722	0.39074	0.00648*
ORACLE	0.38843	0.38655	0.00188*

* $p < .01$, t 分布の自由度は 6

表 2: AUC の比較

	LR	FM	Diff
NAIVE	0.83958	0.84155	0.00197*
DFM	0.84392	0.84642	0.00250**
ORACLE	0.84788	0.85016	0.00228*

* $p < .01$, ** $p < .05$, t 分布の自由度は 6

表より, いずれの評価指標においても FM を用いることで, LR に比べ, モデルの精度が向上していることがわかる. また, 提案モデルの改善幅が最も大きく, これは説明変数間の交互作用を考慮することで遅れコンバージョンをより高い精度でモデル化できているためと考えられる.

6. おわりに

本稿では, オンライン広告におけるスパース性と遅れコンバージョンを考慮した予測モデルを提案した. 評価実験により, 説明変数間の交互作用を考慮することで遅れコンバージョンをより高い精度でモデル化でき, 従来手法よりも高い精度で CVR 予測が可能であることを確認した.

今後は, 実際の広告配信で A/B テストを行うなど, 提案法の有効性に対するさらなる検証を進めたい.

参考文献

- [Chapelle 14a] Chapelle, O.: Modeling Delayed Feedback in Display Advertising, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1097–1105 (2014)
- [Chapelle 14b] Chapelle, O., Manavoglu, E., and Rosales, R.: Simple and Scalable Response Prediction for Display Advertising, *ACM Trans. Intell. Syst. Technol.*, Vol. 5, No. 4, pp. 61:1–61:34 (2014)
- [Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *J. Mach. Learn. Res.*, Vol. 12, pp. 2121–2159 (2011)
- [Lee 12] Lee, K.-c., Orten, B., Dasdan, A., and Li, W.: Estimating Conversion Rate in Display Advertising from Past Performance Data, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 768–776 (2012)
- [Rendle 10] Rendle, S.: Factorization Machines, in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pp. 995–1000, Washington, DC, USA (2010), IEEE Computer Society