

ВИКОРИСТАННЯ UPLIFT-МОДЕЛЮВАННЯ ДЛЯ ВИЯВЛЕННЯ ГРУПИ ВПЛИВУ

М. О. Бондар^{1, а}, А. М. Родіонов¹

¹ Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
Фізико-технічний інститут

Анотація

Uplift-моделювання розглядається у сучасних задачах машинного навчання, які полягають у передбаченні ефекту впливу на вибірку. Цей клас задач широко застосовується у медицині та прямому маркетингу. У даній роботі висвітлені існуючі підходи uplift-моделювання, а також у якості прикладної задачі, основаної на моделюванні підйомів, запропоновано розрахунок максимального прибутку інформаційної кампанії при обмеженні на її вартість.

Ключові слова: uplift, модель, treatment, control, class transformation, two models

Вступ

Для традиційних підходів вирішення задачі класифікації, завдання полягає в передбаченні на основі ряду атрибутів вектора умовну ймовірність його приналежності до кожного з класів. Але існують задачі машинного навчання, у яких питання постає інакше: потрібно передбачити ефект деякого впливу на вибірку і обрати підгрупу, яка найбільше йому піддається.

Такий тип задач вирішується за допомогою Uplift-моделювання, яке спрямоване на знаходження не самого класу, до якого слід віднести об'єкт, а різниці між поведінкою об'єкта у двох групах, а саме лікування (treatment) та контроль (control). Об'єкти в лікувальній або цільовій групі (treatment group) зазнавали певних дій, тоді як об'єкти контрольної групи (control group) – ні.

1. Попередні роботи

У роботі [1] висвітлено основну задачу, яку вирішує uplift-моделювання, а саме: якщо ми хочемо, щоб кожна одиниця маркетингових витрат досягла найбільшого впливу на поведінку споживача, тоді нам потрібно точно змоделювати зміну поведінки, що є результатом цих дій. У роботах [1, 2] описано метод двох моделей, який є універсальним з точки зору використання методів класифікації і широко використовуваним. Але він має недолік, який полягає у тому, що участь у прогнозуванні підйому беруть дві різні моделі, тому їх похибки сумуються. У роботі [3] запропоновано метод трансформації класів, який усуває цей недолік. Цей метод був використаний для визначення групи пацієнтів, які потребують вживання сильнодіючих препаратів для боротьби з хворобою і відокремлення їх від тих, які можуть

одужати власними силами, або ж тих, для кого це лікування було б безнадійне.

2. Позначення

Введемо основні позначення та сформулюємо задачу моделювання uplift:

$\mathbf{X} \in R^m$ — вектор атрибутів об'єкту.

$Y \in \{0, 1\}$ — змінна класу, до якої будемо відносити об'єкти, де $Y = 1$ — позитивна реакція об'єкту, $Y = 0$ — негативна реакція об'єкту.

$W \in \{1, 0\}$ — змінна впливу на об'єкт, де $W = 1$ — показник того, що об'єкт потрапив в цільову (treatment) групу, де була комунікація, $W = 0$ — показник того, що об'єкт потрапив в контрольну (control) групу, де комунікації не було.

Тоді, Y^1 — потенційна реакція об'єкту, якщо з ним була комунікація, Y^0 — потенційна реакція об'єкту, якщо комунікації не було.

Тоді завдання моделювання uplift полягає в побудові моделі, яка буде обчислювати:

$$Uplift = P(Y = 1 | x_1 \dots x_m, W = 1) - \\ - P(Y = 1 | x_1 \dots x_m, W = 0)$$

— uplift для об'єкту $X = (x_1 \dots x_m)$.

Основна проблема даної задачі полягає в тому, що uplift напряду порахувати неможливо, адже неможливо одночасно вплинути і не вплинути на об'єкт. Тому вона потребує спеціальних методів вирішення. Основні з них – це метод двох моделей та його модифікації, який ґрунтується на побудові двох окремих класифікаторів для оцінки ймовірності виконання цільової дії для treatment і control груп, а також метод трансформації класів, який за допомогою функції перетворення змінних класу та впливу дозволяє застосувати лише одну модель.

^аbondar.m.2000@gmail.com

3. Огляд основних підходів uplift-моделювання

3.1. Метод двох моделей (Two models approach)

3.1.1. Дві незалежні моделі (Two independent models)

Далі наведено послідовність кроків для побудови моделі прогнозу uplift:

- 1) Будується перша модель класифікації, що оцінює ймовірність виконання цільової дії серед елементів вибірки, з якими відбулася взаємодія.

$$P_T = P(Y = 1 \mid x_1 \dots x_m, W = 1)$$

- 2) Будується друга модель класифікації, яка оцінює ту ж ймовірність, але серед елементів, з якими не проводилася взаємодія.

$$P_C = P(Y = 1 \mid x_1 \dots x_m, W = 0)$$

- 3) Потім для кожного нового елементу розраховується різниця оцінок ймовірностей двох моделей.

$$uplift = P_T(x_1 \dots x_m) - P_C(x_1 \dots x_m)$$

Тобто:

$$\begin{aligned} treat_model : (x_1 \dots x_m) &\rightarrow \\ \rightarrow P(Y = 1 \mid x_1 \dots x_m, W = 1) \end{aligned}$$

$$\begin{aligned} control_model : (x_1 \dots x_m) &\rightarrow \\ \rightarrow P(Y = 1 \mid x_1 \dots x_m, W = 0) \end{aligned}$$

$$\begin{aligned} uplift_model : (x_1 \dots x_m) &\rightarrow \\ treat_model(x_1 \dots x_m) - control_model(x_1 \dots x_m) \end{aligned}$$

3.1.2. Дві залежні моделі (Dependent feature representation)

- 1) Навчаємо класифікатор за контрольними даними:

$$P_C = P(Y = 1 \mid x_1 \dots x_m, W = 0)$$

- 2) Потім використовуємо передбачення моделі в якості нової ознаки для навчання другого класифікатора на цільовій частині вибірки, тим самим вводячи залежність між двома наборами даних:

$$P_T = P(Y = 1 \mid x_1 \dots x_m, P_C(x_1 \dots x_m), W = 1)$$

- 3) Щоб отримати uplift для кожного спостереження, обчислимо різницю:

$$\begin{aligned} uplift(x_1 \dots x_m) &= P_T(x_1 \dots x_m, P_C(x_1 \dots x_m)) - \\ &- P_C(x_1 \dots x_m) \end{aligned}$$

Тобто:

$$\begin{aligned} control_model : (x_1 \dots x_m) &\rightarrow \\ \rightarrow P(Y = 1 \mid x_1 \dots x_m, W = 0) \end{aligned}$$

$$\begin{aligned} treat_model : (x_1 \dots x_m, control_model(x_1 \dots x_m)) &\rightarrow \\ \rightarrow P(Y = 1 \mid x_1 \dots x_m, P_C(x_1 \dots x_m), W = 1) \end{aligned}$$

$$\begin{aligned} uplift_model : (x_1 \dots x_m) &\rightarrow \\ treat_model(x_1 \dots x_m, control_model(x_1 \dots x_m)) &- \\ - control_model(x_1 \dots x_m) \end{aligned}$$

3.2. Метод трансформації класів (Class transformation)

Даний підхід був запропонований Rzepakowski, P. і Jaroszewicz, S. Він полягає у тому, щоб створити одну модель, яка безпосередньо моделює підйом, замість окремого моделювання ймовірностей treatment та control.

Даний метод дозволяє використання різноманітних класифікаторів для безпосереднього моделювання підняття. Також похибка даного методу значно менша, порівняно з попереднім методом, адже враховується похибка лише однієї моделі. Але в нього є недолік: він добре працює лише з вибірками, збалансованими у поділі на цільову і контрольну групи.

Визначимо змінну $Z \in \{0, 1\}$ таку, що

$$Z_i = Y_i W_i - (1 - Y_i)(1 - W_i)$$

Тобто

$$Z = \begin{cases} 1, & \text{якщо } W = 1 \text{ і } Y = 1, \\ 1, & \text{якщо } W = 0 \text{ і } Y = 0, \\ 0, & \text{інакше.} \end{cases}$$

Тоді ймовірність події $Z = 1$:

$$\begin{aligned} P(Z = 1 \mid x_1 \dots x_m) &= P(Y = 1, W = 1 \mid x_1 \dots x_m) + \\ &+ P(Y = 0, W = 0 \mid x_1 \dots x_m) = \\ &= P(Y = 1 \mid x_1 \dots x_m, W = 1)P(W = 1 \mid x_1 \dots x_m) + \\ &+ P(Y = 0 \mid x_1 \dots x_m, W = 0)P(W = 0 \mid x_1 \dots x_m) \end{aligned}$$

Далі зваживши на те, що поділ на цільову і контрольну групу відбувається випадково, не в залежності від значення якоїсь ознаки, та припустивши, що поділ вибірки на цільову і контрольну групу – збалансований, тобто:

$$P(W = 1) = P(W = 0) = \frac{1}{2}$$

$$\begin{aligned} 2P(Z = 1 \mid x_1 \dots x_m) &= P^T(Y = 1 \mid x_1 \dots x_m) + \\ &+ P^C(Y = 0 \mid x_1 \dots x_m) = P^T(Y = 1 \mid x_1 \dots x_m) + \\ &+ 1 - P^C(Y = 1 \mid x_1 \dots x_m) \end{aligned}$$

$$uplift = 2P(Z = 1 \mid x_1 \dots x_m) - 1$$

Тобто величина $2P(Z = 1 \mid x_1 \dots x_m) - 1$ і буде визначати підйом для об'єкта $x_1 \dots x_n$.

4. Метод розрахунку максимального прибутку підприємства

У цьому розділі представлений метод, у якому на основі вибору цільової аудиторії методами uplift обирається найбільш лояльна група користувачів.

Далі на основі вхідних фінансових обмежень компанії, таких як сума витрат на комунікацію з клієнтом та максимальний бажаний дохід на одного клієнта, розраховується максимально можливий прибуток підприємства від проведеної кампанії.

Введемо умовні позначення та опишемо кроки даного методу:

- N — кількість обраних клієнтів для комунікації
- up_i — значення uplift для i -го клієнта
- C — загальна допустима сума витрат
- α — сума витрат, яка може бути витрачена на комунікацію з одним клієнтом
- β — максимальний дохід, який може бути одержаний від одного клієнта

Задача полягає у максимізації прибутку компанії, який описаний системою (1) за кількістю користувачів, з якими проводитиметься комунікація:

$$\begin{cases} \sum_{i=1}^N (\beta \cdot up_i - \alpha) \rightarrow \max_N \\ \sum_{i=1}^N \alpha \leq C \end{cases} \quad (1)$$

- 1) За допомогою вже підготовленої на основі частини вибірки моделі передбачити uplift для осіб, з якими ще не було проведено комунікацію.

$$uplift_model(\mathbf{X}_1 \dots \mathbf{X}_n)^T \rightarrow (\text{user_id}, \text{uplift})$$

- 2) Обрати клієнтів, для яких модель спрогнозувала додатний uplift, тобто саме тих, з якими слід проводити комунікацію.
- 3) Відсортувати вектори отриманої матриці у порядку спадання по uplift. Таким чином на перших місцях опиняться клієнти, які є найбільш лояльними до даного способу впливу.

Далі вирішити систему (1) — це і буде передбачення очікуваного прибутку від кампанії — R . Тоді $N = \text{argmax}(R)$ — кількість клієнтів, з якими слід провести комунікацію.

5. Оцінювання якості моделей

Розглянемо один з методів оцінки ефективності моделей підйому, яка найкраще підійде до нашої задачі. Наша мета — оцінити точність кожного підходу та обрати підгрупу для подальшого впливу.

У задачі uplift-моделювання використовувати класичні метрики для оцінки якості моделі не вийде, так як для елементів вибірки не існує заздалегідь відомих значень uplift. Ця проблема виникає через особливості даної задачі: для кожного тестового випадку ми знаємо лише одну з його реакцій, або після того, як було здійснено вплив, або коли не було вжито жодної дії. Таким чином, методи оцінки класифікаторів підняття базуються на припущенні, що випадки, які оцінюються моделлю, дійсно поведуться однаково. Всі метрики розраховуються, так або інакше групуючи елементи вибірки. Наприклад, розглядається 10% вибірки, 20% і так далі.

У рамках нашої задачі порівнювати моделі найкраще за допомогою метрики uplift by decile [4], так як заздалегідь невідомо, яка частина аудиторії повинна бути обрана в якості цільової. Для цього потрібно відсортувати вибірку за спаданням величини uplift та порахувати різницю середніх значень Y у лікувальній та контрольній групі для кожного k -персентелю вибірки:

$$Uplift_k = \langle Y_k | W = 1 \rangle - \langle Y_k | W = 0 \rangle,$$

де Y_k — значення змінних класу на $k\%$ вибірки.

Висновки

Розглянуті основні широко використовувані методи uplift-моделювання, визначені їх переваги та недоліки. Підхід трансформації класів використовує лише одну модель класифікації, що зменшує ймовірність помилки, але потребує збалансованості поділу вибірки на лікувальну та контрольну групи. В той час метод двох моделей не вимагає цього, але за рахунок використання двох моделей класифікації, має більшу ймовірність помилки.

Також запропонований метод для прогнозу загального максимального прибутку від впливу маркетингової кампанії. Він здатний підлаштовуватися під обрану вибірку клієнтів, оскільки використані у ньому методи uplift-моделювання дають змогу виділити саме тих осіб, які найбільш схильні до позитивної реакції впливу на них.

У майбутньому планується дослідити метод uplift-моделювання за допомогою дерев рішень, більш точно описати задачу оптимізації, а також визначити для конкретних наборів даних, які саме атрибути векторів вибірки найбільше корелюють зі значеннями uplift задля більш точного прогнозу прибутку.

Перелік використаних джерел

1. Radcliffe N. Using control groups to target on predicted lift: Building and assessing uplift model // Direct Marketing Analytics Journal. — 2007. — P. 14–21.
2. Uplift Modeling with ROC: An SRL Case Study / Nassif H., Kuusisto F., Burnside E. S., and Shavlik J. W. // Late Breaking Papers of the 23rd International Conference on Inductive Logic Programming. — 2013. — Vol. 1187 of CEUR Workshop Proceedings. — P. 40–45. — Access mode: <http://ceur-ws.org/Vol-1187/paper-07.pdf>.
3. Maciej Jaskowski, Szymon Jaroszewicz. Uplift modeling for clinical trial data // ICML Workshop on Clinical Data Analysis. — 2012.
4. P. Gutierrez, J.-Y. G  rardy. Causal Inference and Uplift Modeling: A review of the literature. // JMLR: Workshop and Conference Proceedings. — 2016. — Vol. 67. — P. 1–13. — Access mode: <http://proceedings.mlr.press/v67/gutierrez17a/gutierrez17a.pdf>.