Recenzja rozprawy doktorskiej

pt. „Linear regression for Uplift Modeling"

Krzysztofa Rudasia

Zagadnienie rozpatrywane w rozprawie

Rozprawa dotyczy zagadnienia, które można traktować jako przykład statystycznego modelowania związków przyczynowych (coś, co jest nazywane *causal modeling* lub *causal inference*). Używane w rozprawie narzędzia matematyczne są typowe dla najbardziej klasycznego działu statystyki matematycznej: teorii modeli liniowych. Rozpatruje się wpływ pewnego „działania" (*treatment*) na jednostki populacji. Zakłada się (w najprostszej, zasadniczej wersji modelu), że zmienna odpowiedzi y jest postaci

$$y = \begin{cases} x'\beta^T + \varepsilon^T & \text{jeśli jednostka jest poddana „działaniu";} \\ x'\beta^C + \varepsilon^C & \text{jeśli jednostka nie jest poddana „działaniu",} \end{cases}$$

gdzie x oznacza wektor zmiennych objaśniających (*covariates*). Celem jest estymacja „efektu" $\beta^U = \beta^T - \beta^C$ na podstawie próbki losowej $(X, y) = \{(x_i, y_i), i = 1, \ldots n\}$ podzielonej na grupę „badaną" (X^T, y^T) i grupę „kontrolną" (X^C, y^C). Podziału dokonuje się losowo, poprzez randomizację, aby zidentyfikować efekt działania / braku działania na tę samą jednostkę: $x'\beta^U$. Typowo traktuje się zmienne objaśniające x jako losowe. Pozwala to sformułować zadanie predykcji zmiennej losowej $x'_{test}\beta^U$ dla nowej jednostki wylosowanej z populacji. Takie postawienie zadania, typowe dla teorii uczenia maszynowego, umożliwia obiektywną empiryczną ocenę proponowanych algorytmów. To podejście, reprezentowane dość konsekwentnie w rozprawie Rudasia, jest moim przekonaniu słuszne.

Mimo, że klasyczna teoria liniowej statystyki pozwala estymować β^U w standardowy sposób, rozpatrywane są również rozwiązania wykorzystujące specyfikę zagadnienia. Są to w szczególności „*uplift estimator*" i „*corrected uplift estimator*" badane w rozprawie Rudasia (Rudaś jest współautorem tego drugiego estymatora, wraz z promotorem rozprawy). Podstawowe pytanie dotyczy porównania różnych metod estymacji β^U. Praca Rudasia udziela częściowej odpowiedzi na to pytanie.

Omówienie zawartości rozprawy

Rozprawa składa się (oprócz Streszczenia, spisu treści, bibliografii itp.) z 7 rozdziałów. Rozdział 1 przedstawia zagadnienie i ma charakter wstępny. Rozdział 2 zawiera wyniki asymptotyczne dotyczące zachowania 3 estymatorów: klasycznego „podwójnego" (*double*), znanego wcześniej wyspecjalizowanego estymatora zwanego *uplift* i jego ulepszonej wersji zwanej *corrected uplift*. Ten

rozdział jest według mnie najważniejszą częścią pracy. W Rozdziale 3 Autor rozważa sytuację źle wyspecyfikowanego modelu liniowego. Zakłada się, że regresja zawiera człon nieliniowy taki sam dla obu grup jednostek. Wyniki mają postać podobną do tych z Rozdziału 1. Rozdział 4 jest poświęcony innemu niż w poprzednich rozważniach schematowi randomizacji, który Autor nazywa „randomizacją prostą" (*simple randomization*). Rozdział 5 omawia regularyzację estymatorów parametru β^U przez dodanie składnika „kary" typu ℓ^2. Otrzymuje się zatem odpowiednie wersje regresji grzbietowej. Rozdział 6 omawia estymatory „ściągające" typu Jamesa-Steina. Tutaj także stosuje się znane wcześniej pomysły (m.in. estymatory SMSE) do specjalnego zadania estymacji parametru β^U i porównuje się własności otrzymanych rozwiązań. Rozdział 7 jest poświęcony wynikom symulacyjnym, które ilustrują i uzupełniają zawarte w rozprawie wyniki teoretyczne.

Ocena wyników rozprawy

Uwagi ogólne.

Zagadnienie rozpatrywane przez Rudasia jest dobrze postawione i ważne dla zastosowań. To jest dużą zaletą rozprawy. Z drugiej strony, teoria modeli liniowych, na których rozprawa się koncentruje, jest bardzo rozwinięta. Narzędzia używane przez Autora rozprawy są klasyczne, proponowane metody mają charakter modyfikacji metod klasycznych. W rezultacie wyniki Rudasia mają na ogół charakter drobnych ulepszeń, czasami osiągniętych dużym nakładem sił, poprzez skomplikowane rachunki. Przytoczę przykłady. (W moich uwagach koncentruję się na wynikach asymptotycznych, które opisują najważniejsze własności estymatorów. W epoce dużych baz danych, zachowanie estymatorów dla małych próbek jest mniej istotne.)

- Porównanie Tw. 2 z Tw. 4 pokazuje, że wyspecjalizowany „uplift estimator" ma *gorsze* własności asymptotyczne niż standardowy „double estimator". Tu już nie chodzi o drobne ulepszenie, tylko zasadnicze pogorszenie! (Rudaś nie jest za to odpowiedzialny, bo nie jest autorem „uplift estimator", ale ten fenomen pokazuje tło pracy).

- W podrozdziale 2.3 Rudaś poprawia „uplift estimator". Pomysł stojący za proponowanym tu „corrected uplift estimator" jest elegancki i zasługuje na uznanie. Niemniej, osiąga się tylko tyle, że nowy estymator ma *taką samą* asymptotykę, jak klasyczny „double estimator".

- Porównanie Tw. 6, 7 i 8 w Rozdziale 3 jest trochę bardziej skomplikowane. W obecności członu nieliniowego, „uplift estimator" jest czasami gorszy, czasami lepszy asymptotycznie od „double estimator". „Corrected uplift estimator" autorstwa Rudasia ma *taką samą* asymptotykę, jak klasyczny „double estimator".

- Jeśli chodzi o regularyzację rozważaną w Rozdziale 5, to udowodnione tu Tw. 11 jest „niesprawiedliwym porównaniem": porównuje się regularyzowany „uplift" z *nie*regularyzowanym „double", co prowadzi do łatwego do zgadnięcia wyniku.

- W Rozdziale 6 Rudaś rozwija ładny pomysł SMSE („estymator ściągający minimalizujący błąd średniokwadratowy" zaproponowany w 1975 roku) i adaptuje ten pomysł do „uplift estimator". Dowód Tw. 13 jest bardzo rachunkowy i skomplikowany. Sam jednak wynik ma charakter „niesprawiedliwego porównania", w dodatku otrzymany jest przy bardzo ograniczających założeniach.

- W Rozdziale 7, doświadczenia symulacyjne, ogólnie mówiąc, świadczą o niewielkich różnicach MSE dla dużych liczności próbek. Nawiasem mówiąc, nie jest jasno powiedziane, jak jest obliczany „predictive MSE" (duża próbka testująca?).

Powyższe ogólne uwagi można podsumować stwierdzeniem, że rozprawa Rudasia nie zawiera przełomowych wyników. Zawiera jednak szereg wartościowych uzupełnień i spostrzeżeń. Zasadnicza część wyników jest poprawna. Praca świadczy o opanowaniu przez Autora warsztatu i dobrej orientacji w badanym kręgu zagadnień. Zawarte w rozprawie rachunki wymagały sporej biegłości i miejscami pomysłowości. Jest to, w moim przekonaniu, dostateczny wkład, który pozwala rozprawę ocenić pozytywnie.

Uwagi krytyczne.

- W całej pracy powtarza się następujące założenie o zmiennych objaśniających: $\mathrm{E}x'_{i\cdot} = 0$, $\mathrm{Var}x'_{i\cdot} = \Sigma$ i rozpatruje się regresję bez wyrazu wolnego. Dla klasycznego estymatora „double" można po prostu zastąpić $\mathrm{Var}x'_{i\cdot}$ przez $\mathrm{E}x'_{i\cdot}x_{i\cdot}$ i wtedy $x_{i\cdot}$ może mieć pierwszy wyraz 1, a wzór na rozkład asymptotyczny nie ulega zmianie. *Być może* podobnie jest dla innych estymatorów rozpatrywanych w rozprawie, ale nie jestem pewny. W każdym razie, użyteczne w praktyce twierdzenia powinny być sformułowane bez nierealistycznego założenia $\mathrm{E}x'_{i\cdot} = 0$ i obejmować regresję z wyrazem wolnym.

- Kwestia randomizacji (przydziału jednostek do grup „T" i „C") zajmuje w rozprawie sporo miejsca. Według mnie, Autor w swoich rozważaniach na ten temat komplikuje proste sprawy i walczy z trudnościami, które sam stwarza. Randomizacja zwana „complete" polega na tym, że dla n jednostek w próbce ustalamy deterministyczne liczności grup $n^T = q^T n$ i $n^C = q^C n$. Następnie (jak słusznie zauważa Autor na str. 17) możemy losowo posortować n jednostek i początkowe n^T z nich zaliczyć do grupy „treatment". Jeśli rozważamy losowe zmienne x, to takie postępowanie

jest równoważne pobraniu dwóch niezależnych próbek o liczności n^T i n^C z jednorodnej populacji i koniec. Jednostki *są* niezależne, jeśli je ponumerujemy zgodnie z posortowaniem. W terminologii używanej przez Autora jednostki są *warunkowo niezależne* przy ustalonych zmiennych g_i. To jest formalnie poprawne, bardzo mądrze wygląda, ale jest całkowicie niepotrzebne i utrudnia redakcję dowodów. Z drugiej strony mamy randomizację zwaną „simple". Dla każdej jednostki losuje się niezależnie przydział do grupy ‚T" z prawdopodobieństwem q^T (lub do grupy „C" z prawdopodobieństwem q^C). W rezultacie jednostki mogą być traktowane jako niezależne bez przenumerowania. Tylko co to nam daje w porównaniu z „complete" randomization? Daje to losowe liczności grup i w konsekwencji pogorszenie własności estymatorów. W modelu z losowymi zmiennymi x rozważanie randomizacji „simple" jest zbędne. Oczywiście, randomizacja „simple" upraszcza analizę w przypadku *deterministycznego* planu, ale ten model nie przystaje do zasadniczej filozofii „uplift modeling".

- Zasadniczą wielkością w „uplift modeling" jest „predictive MSE". W rozprawie to pojęcie pojawia się wielokrotnie. Tym dziwniejszy jest fakt, że podstawowy, ogólny wzór (1.6) na „predictive MSE" jest błędny. Po pierwsze, coś się nie zgadza z wymiarami macierzy. Ale to drobiazg, wystarczy rozważyć pojedynczy (losowy) wektor x_{test} (niech to będzie wektor kolumnowy, $p \times 1$). Można zdefiniować predictive MSE jako $\mathrm{E}(x'_{\text{test}}\hat{\beta} - x'_{\text{test}}\beta)^2$. Korzystając z niezależności $\hat{\beta}$ i x_{test} otrzymujemy

$$\begin{aligned}\text{Predictive MSE} &= \mathrm{E}\, x'_{\text{test}}(\hat{\beta}-\beta)(\hat{\beta}-\beta)'x_{\text{test}}\\ &\quad \mathrm{E}\,\mathrm{tr}(x'_{\text{test}}(\hat{\beta}-\beta)(\hat{\beta}-\beta)'x_{\text{test}})\\ &\quad \mathrm{E}\,\mathrm{tr}((\hat{\beta}-\beta)(\hat{\beta}-\beta)'x_{\text{test}}x'_{\text{test}})\\ &\quad \mathrm{tr}(\mathrm{E}\,(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\mathrm{E}\,x_{\text{test}}x'_{\text{test}})\\ &\quad \mathrm{tr}(V\Sigma),\end{aligned}$$

gdzie $\Sigma = \mathrm{E}x_{\text{test}}x'_{\text{test}}$ i $V = \mathrm{E}(\hat{\beta}-\beta)(\hat{\beta}-\beta)'$. Jeśli – tak jest dla estymatora „double" – macierz V jest (asymptotycznie) postaci $c(\sigma^2/n)\Sigma^{-1}$ to

$$\text{Predictive MSE} \approx \mathrm{tr}\left(c\frac{\sigma^2}{n}I_p\right) = c\frac{\sigma^2 p}{n},$$

dla pewnego skalara c. Jeśli jednak model jest źle wyspecyfikowany, to V nie jest, nawet w przybliżeniu, proporcjonalna do Σ^{-1}.

Praca z pewnością by zyskała, gdyby Autor konsekwentnie (i poprawnie) rozważał własności predykcji $x'_{\text{test}}\hat{\beta}^U$ a nie tylko estymatora $\hat{\beta}^U$. Większość wyników jest jednak sformułowana w terminach $\hat{\beta}^U$. Z błędnego wzoru (1.6) Autor (na szczęście) na ogół nie korzysta.

Uwagi o redakcji pracy

Praca jest napisana starannie i ma dość przejrzysty układ. Niektóre oznaczenia mogą się nie podobać. Na przykład $X^{T'}$ oznacza transpozycję macierzy X^T (dlaczego nie X'_T?). Traktowanie $x_{i\cdot}$ jako wektorów wierszowych nie ułatwia lektury ($x'_{i\cdot}x_{i\cdot}$ jest macierzą $p \times p$). Ale Autor ma prawo do wybrania dziwnych oznaczeń.

Podsumowanie i wnioski

Pomimo pewnych zastrzeżeń i usterek, rozprawa spełnia warunki stawiane pracom doktorskim z matematyki. Wnoszę o dopuszczenie Pana Krzysztofa Rudasia do dalszych etapów przewodu doktorskiego.

Warszawa, 22 listopada 2021

WNiemiro

Wojciech Niemiro