ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
Seria III: MATEMATYKA STOSOWANA XXXV (1992)

Recenzja

Mirosław Krzyśko
Analiza dyskryminacyjna,
Wydawnictwa Naukowo-Techniczne, Warszawa 1990, str 155.

Jest to pierwsza monografia w języku polskim poświęcona — tak ważnym w praktyce statystycznej — problemom dyskryminacji. Zakres poruszanych zagadnień jest bardzo szeroki. *Książka składa się z pięciu rozdziałów omawiających kolejno teoriodecyzyjne modele klasyfikacji, metody estymacji w analizie dyskryminacyjnej, zmienne dyskryminacyjne będące uogólnieniem funkcji dyskryminacyjnej Fishera, metody sekwencyjne analizy dyskryminacyjnej oraz analizę dyskryminacyjną szeregów czasowych* (Przedmowa Autora). Przekonania i upodobania Autora znajdują, oczywiście, odbicie w rozkładzie akcentów i selekcji materiału. Najwięcej miejsca poświęca się modelom dyskryminacji z normalnymi rozkładami wewnątrz populacji. Odpowiada to, jak sądzę, pozycji zajmowanej przez te modele w zastosowaniach. Szkoda trochę, że zabrakło miejsca dla modeli dyskretnych (klasyfikacji na podstawie cech jakościowych). Dwa ostatnie rozdziały są wprowadzeniem do bardziej specjalnych działów teorii.

Wielka ilość faktów i twierdzeń, zgromadzona w książce, została opatrzona pełnymi i szczegółowymi dowodami. Czytelnik, zgłębiający tajniki analizy dyskryminacyjnej (student matematyki lub pracownik naukowy) znajdzie tu samowystarczalny wykład zawierający wiele nowych wyników i bogatą bibliografię. Wykład jest ożywiony i umotywowany przykładami zastosowań w naukach przyrodniczych i doświadczeń symulacyjnych, jasno przedstawionych i dobrze udokumentowanych. W partiach książki dotyczących zastosowań (np. w świetnie napisanym p. 2.10) poruszane są ważne zagadnienia, których Autor, niesłusznie chyba, nie włączył w główny nurt wykładu. Mam na myśli testowanie hipotezy o równości macierzy kowariancji, ilustrację zasady sprawdzania rozwiązania na *nowym* zbiorze danych, dyskusję na temat ilości cech, które warto uwzględnić w klasyfikacji.

Zalety książki są niewątpliwe. Czy jednak Autorowi udało się *przedstawienie teorii i wybranych metod analizy dyskryminacyjnej w logicznie upo-*

rządkowany sposób i formie dogodnej w praktyce (taki cel deklaruje Przedmowa)? Obawiam się, że nie w pełni. Spróbuję to uzasadnić, koncentrując się na kilku sprawach.

Wiele jest w książce zawiłych i trudnych rachunków. Zabrakło miejsca na dyskusję pewnych logicznych subtelności. Najjaskrawszym przykładem jest sposób wprowadzenia (p.2.1) pojęcia statystycznej reguły klasyfikacyjnej (tj. reguły opartej na próbie uczącej). Autor nie widzi potrzeby zdefiniowania tego, fundamentalnego dla całej teorii, pojęcia i poprzestaje na lakonicznym stwierdzeniu (31_{10}): *Niech d_n, gdzie n jest liczebnością próby, oznacza regułę klasyfikacyjną opartą na danych z próby.* Ani słowa o tym, z jakiej to populacji próba, jak się ona ma do obserwacji poddawanych klasyfikacji. Dwie linijki niżej pojawia się sformułowanie: [...] *ryzyko bayesowskie $r(d_n, q)$ jest zbieżne według prawdopodobieństwa* [...]. Niedawno obserwowałem, ile czasu zajęło grupie studentów IV roku zrozumienie — o jakim prawdopodobieństwie tutaj mowa! Inna rzecz, że byli to studenci matematyki, dociekliwi i krytyczni. Praktycznie zorientowany czytelnik przełknie, zapewne, ten fragment bez zmrużenia oka. Ale logicznemu uporządkowaniu teorii to nie posłuży.

Specyfika analizy dyskryminacyjnej jest taka, że w monografii tego przedmiotu muszą się znaleźć i twierdzenia matematyczne, i metody heurystyczne. Można mieć pretensję do Autora tylko o to, że nie ułatwia czytelnikowi odróżnienia jednego od drugiego. W rozdziale 4 (metody sekwencyjne) pewne procedury klasyfikacyjne opisuje się używając, wprowadzonych w rozdziale 1, pojęć statystycznej teorii decyzji. Niestety, brak precyzyjnego, teoriodecyzyjnego sformułowania zadań, które te procedury mają rozwiązywać. Znamienne, że w całym rozdziale 4 nie pojawia się ani jedno twierdzenie. W p. 4.2 Autor definiuje sekwencyjną procedurę, która wymaga rozpatrzenia dwóch (lub więcej) cech, podczas gdy nie jest wykluczone, że do osiągnięcia żądanych nierówności $P(i|j) \leq \alpha_{ij}$ wystarczy jedna cecha (ale inna procedura). Jest dla mnie zresztą zagadką, dlaczego Autor tutaj nagle troszczy się o całą macierz błędów $(P(i|j))$, skoro w prostszej, niesekwencyjnej sytuacji w rozdziale 1 zadowolił się minimalizacją skalarnej funkcji $\sum_{ij} q_j S(i|j) P(i|j)$ — ryzyka bayesowskiego? Zadanie sekwencyjnej klasyfikacji dla takiego skalarnego kryterium ma optymalne rozwiązanie, oparte na równaniu Bellmana (por. p.2.8 w monografii Devijvera i Kittlera). W p. 4.5 terminy teoriodecyzyjne są używane w sposób tak dowolny, że trudno zrozumieć, o co chodzi. Oto czytamy (110_{14}): *Zakładamy, że znajomość cechy y pozwala na poprawne zaklasyfikowanie rozpatrywanego obiektu.* Dalej (111_{13}): *Przez $S(j|i)$ oznaczamy stratę poniesioną po zaklasyfikowaniu badanego obiektu do populacji π_j, podczas gdy w rzeczywistości jest on reprezentantem populacji π_i.* Niby w porządku. Tymczasem ze wzoru (4.31) jasno widać, że 1) znajomość cechy y nie zawsze pozwala na poprawne zaklasyfikowanie (gdyby tak było, to

$\int_{A_i} f_{im}(y|x)dy$ byłaby po prostu równa jedności); 2) jeżeli *błędnie* zaklasyfikujemy obiekt pochodzący z π_i, dla którego zdarzy się, że $Y_i \notin A_i$, to nie ponosimy żadnej straty!

Kontrowersyjna wydaje mi się myśl przewodnia rozdziału 2. Piszę to z pewnym wahaniem, ponieważ rozdział ten, zawierający oryginalny i ciekawy materiał, starannie zredagowany, uważam za najlepszą i najważniejszą część książki. Zastrzeżenia, ktore poniżej sformułuję są, przyznaję, mocno subiektywne i w niczym tej oceny nie podważają. We Wprowadzeniu Autor wyraźnie sugeruje, że *wszystkie* statystyczne reguły klasyfikacyjne powstają poprzez wstawienie estymatorów gęstości w gotowe wzory, zawierające nieznane gęstości. To nie jest prawda. Dalej czytamy (32^{10}): *Zagadnienie jakości klasyfikacji przy nieznanych rozkładach prawdopodobieństwa sprowadza się, jak widać, do zagadnienia jak najlepszej estymacji funkcji gęstości.* Ależ czytelnik, który potraktowałby tę deklarację poważnie, rzuciłby książkę w kąt i pobiegłby do biblioteki po monografię o estymacji gęstości! Dalsze rozważania w omawianym rozdziale są konsekwentną realizacją przyjętego punktu widzenia. Autor formułuje rozmaite kryteria optymalności (minimalizacja wariancji przy żądaniu nieobciążoności, zasada bayesowska) i estymuje optymalnie $\log f_i(x)$, gdzie f_i jest gęstością w i-tej populacji. Ciekawe, jakie wyniki dałaby optymalna estymacja, powiedzmy, $\operatorname{arctg} f_i(x)$, funkcji równoważnych? Z tym wiąże się zabawna kwestia, pozornie dotycząca jedynie terminologii. Otóż to, co Autor nazywa estymatorem bayesowskim funkcji dyskryminacyjnej, w pracach Geissera jest nazywane estymatorem quasi-bayesowskim. Odwrotnie, Autor określa jako quasi-bayesowski estymator, nazwany przez Geissera bayesowskim. Dlaczego? Wobec szczupłości miejsca muszę odesłać zainteresowanych do prac Geissera. W największym skrócie rzecz ma się tak: estymacja dla Geissera jest środkiem prowadzącym do klasyfikacji; dla Autora omawianej książki sama staje się celem.

Włączenie do książki p 2.9 zawierającego, z konieczności bardzo pobieżny, przegląd nieparametrycznych estymatorów gęstości, mija się chyba z celem. Ani czytelnika zainteresowanego teorią taki przegląd nie zadowoli, ani czytelnik praktycznie nastawiony nie znajdzie dostatecznych informacji o sposobie korzystania z estymatorów. Chodzi nie tylko o wybór parametru wygładzającego. Chodzi też o odpowiedź na pytanie: kiedy, w konkretnym praktycznym zadaniu, trzeba uciec się do nieparametrycznych metod kasyfikacji?

Mało uwagi poświęca Autor sprawie oceny prawdopodobieństw poprawnej decyzji. Co prawda, odsyła czytelnika do znanej bibliografii Toussainta. Zdaje się jednak sugerować, że w praktyce wystarczy posługiwać się prymitywnym oszacowaniem opartym na nierówności Bonferroniego, wstawiając w razie potrzeby estymatory w miejsce nieznanych parametrów (40^{11} i 40^{16}).

Jest to zachęta do niezbyt odpowiedzialnego postępowania. Nb. przykry błąd, odwrócenie znaku nierówności w oszacowaniu $P(i|i)$, powtarza się w książce trzykrotnie (str. 17, 21, 137). Niedostrzeżenie tego podczas korekty świadczy chyba o małym znaczeniu, jakie przypisuje Autor temu, czy statystyk powinien - przekazując rozwiązanie praktykowi — oszacować jakość tego rozwiązania z góry, czy z dołu!

Zagadnienie poszukiwania optymalnego rozwiązania w ograniczonej klasie reguł (np. liniowych) jest potraktowane pobieżnie. Zamiast ogólnego omówienia tej ważnej koncepcji, Autor w p. 1.6 zagłębia się nagle w szczegółową analizę bardzo specjalnego wyniku Andersona i Bahadura, łamiąc spójność wykładu. Zapomina przy tym sprecyzować, że rozważania dotyczą rozkładów normalnych.

Chciałbym jeszcze dodać kilka drobnych uwag.

W tw. 1.7 założenie $\alpha \geq 0$, $\beta \geq 0$ jest zbędne. To widać z dowodu.

Dowód tw. 1.8, przepisany dosłownie z podręcznika Zubrzyckiego, był całkowicie jasny w swoim oryginalnym kontekście. W nowym miejscu winien był zostać nieco inaczej zredagowany.

W lemacie 2.1 błąd drukarski. Poniżej (34_8) oznacza się tą samą literą i wskaźnik sumowania oraz $\sqrt{-1}$.

Autor nie ustrzegł się pewnych niestaranności językowych. Kilka przykładów: [zmienne losowe] *mają rozkłady niezależne* (46_{10}); *rozkłady brzegowe* (powinno być: rozkład brzegowy jednej zmiennej oraz warunkowy — drugiej, 46_2). Aby rozkład warunkowy jednej zmiennej losowej przy danej wartości drugiej był normalny, nie wystarcza założenie normalności rozkładów brzegowych (114).

Drobne uzupełnienie historyczne. Definicję zgodnej reguły klasyfikacyjnej Autor podaje (29_9) za Van Ryzinem (1966). Bardzo podobna definicja została sformułowana już w 1951 r. przez Fix i Hodgesa w pracy, wydobytej niedawno z zapomnienia.

Szczerze polecam ciekawą i pożyteczną książkę Krzyśki wszystkim, którzy chcą pogłębić i rozszerzyć swoją wiedzę z analizy dyskryminacyjnej. Tym czytelnikom, którzy nie są statystykami i chcą znaleźć metodę dyskryminacji dopasowaną do konkretnego problemu, doradziłbym ostrożność. Studentom, którzy stykają się z przedmiotem po raz pierwszy, doradziłbym inną książkę.

WOJCIECH NIEMIRO