Ze świata USOS. Część 5 – Oceanarium, czyli o nurkowaniu w otchłani danych

Przemysław BIECEK[*]

Co wspólnego ma ocean i system USOS? Okazuje się, że znacznie więcej niż tylko literę „o” występującą w obydwu słowach.

Eksploracja oceanu to bardzo wciągający temat. Co znajduje się w otchłani? Źródła minerałów, piękne widoki, dziwaczne stwory? Każdy może bawić się na brzegu w szukanie muszelek, ale aby zejść głębiej i zobaczyć coś, czego nikt inny jeszcze nie widział, potrzebny jest i trening, i specjalistyczny sprzęt.

A co z systemem USOS? Bazy danych tego systemu zawierają wiele informacji. Począwszy od danych, takich jak oceny wystawione studentom, poprzez wyniki ankiet wystawione przez studentów, do wyników z rekrutacji, preferencji w zapisach na przedmioty itp. Początki systemu USOS sięgają roku 1999, przez te kilkanaście lat na niektórych wydziałach zebrał się prawdziwy ocean danych.

Analizując te dane, możemy dowiedzieć się czegoś ciekawego o życiu na uczelni. Oczywiście, co innego będzie ciekawe dla dziekana, co innego dla prowadzącego zajęcia, a jeszcze co innego dla studenta. Zdecydowana większość użytkowników systemu USOS to studenci, dlatego poniżej spojrzymy na ten ocean danych z perspektywy studenta.

Postawimy trzy pytania oraz pokażemy, jak wygląda proces znajdowania odpowiedzi na każde z nich. Każde kolejne pytanie wymagać będzie coraz sprawniejszego aparatu matematycznego i będzie bardziej wymagające obliczeniowo.

Rozwiązanie zadania F 850.
Zaniedbajmy dla uproszczenia średnicę rury i oznaczmy przez d średnicę utworzonej przez nią obręczy. Wskutek obrotu Ziemi z częstością Ω różnica prędkości górnego i dolnego krańca rury względem układu inercjalnego związanego z osią obrotu Ziemi jest równa $\Delta v = \Omega d \cos\varphi$. O ile woda może swobodnie płynąć w rurze, obrót rury zmienia prędkość fragmentów rury w układzie inercjalnym, a wody – nie; prowadzi do powstania różnicy prędkości tych fragmentów rury i zawieranej przez nie wody równej właśnie Δv. Podstawiając $d = 1m$, otrzymujemy prędkość rzędu 0,01 mm/s; można ją próbować dostrzec pod mikroskopem, używając zawiesiny zamiast wody.

Zbieramy muszelki, czyli co znajdziemy na brzegu

Pierwsze pytanie, które często przychodzi na myśl studentom, to *Na jakie inne zajęcia zapisana jest ta brunetka, która chodzi ze mną na mikroekonomię?* Odpowiedź na takie pytanie jest stosunkowo prosta, o ile ma się bezpośredni dostęp do bazy systemu USOS. Dane są przechowywane w postaci tabel, które odpowiadają relacjom, np. takim jak *student X jest zapisany na zajęcia Y*.

Aby dowiedzieć się, na jakie inne zajęcia chodzi kolega/koleżanka z naszej grupy, musimy znać strukturę tych tabel. Najpierw wśród osób zapisanych z nami na mikroekonomię odnajdziemy interesującą studentkę y, a następnie sprawdzimy, na jakie inne zajęcia y jest zapisana. Wystarczy dostęp do bazy danych i znajomość języka zapytań SQL.

Snorkeling, czyli pływanie z maską i rurką

Drugie pytanie, które zadamy, dotyczyć będzie mapy „popularności” i „trudności” przedmiotów. Na wielu wydziałach studenci mogą samodzielnie wybierać przedmioty z puli przedmiotów obieralnych. Czyż nie byłoby wspaniale mieć mapę raf i mielizn? Mapę z zaznaczoną zdawalnością dla każdego przedmiotu oraz informacją, jak oceniali go studenci w poprzednich semestrach.

W przeciwieństwie do pierwszego pytania informacja o„popularności”, czy „trudności” przedmiotu nie jest bezpośrednio przechowywana w bazie danych. Te charakterystyki trzeba wyznaczyć na podstawie danych z tabel (nazywanych surowymi, nieprzetworzonymi danymi), tworząc agregaty. Jednym z takich agregatów może być średnia arytmetyczna opinii studentów

[*]Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego, Uniwersytet Warszawski

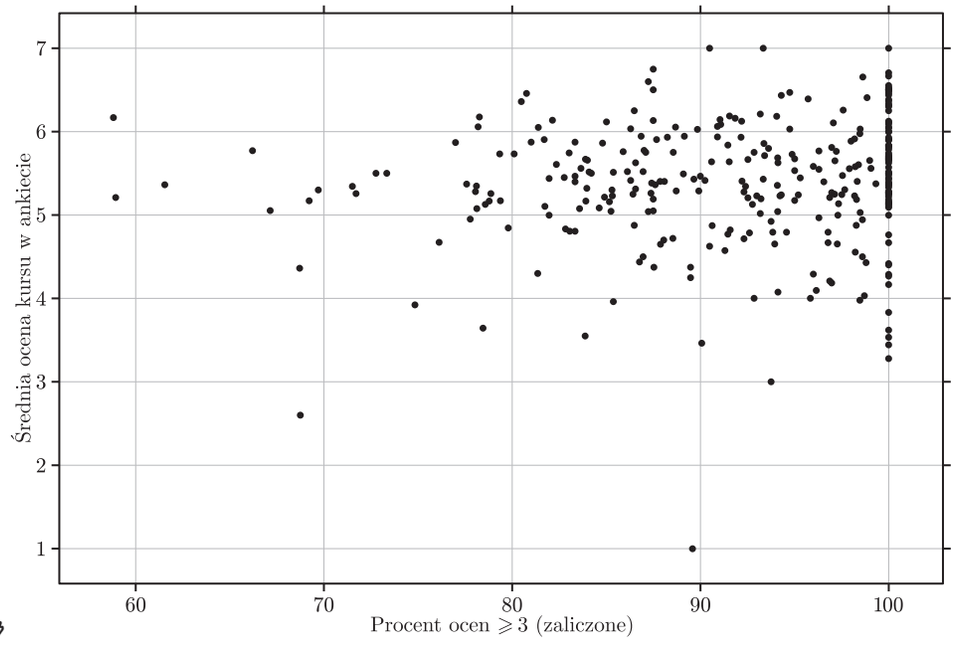
o przedmiocie. Na podstawie danych z wynikami ankiet możemy obliczyć

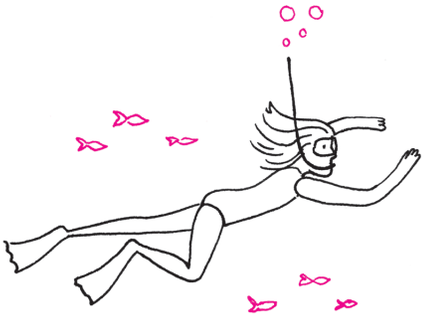
$$o(k) = \frac{\sum_{y \in Y} o(y,k)}{\#Y},$$

gdzie $o(k)$ to średnia opinia o przedmiocie k, $o(y,k)$ to opinia o przedmiocie k studenta y, Y to zbiór wszystkich studentów, a symbolem $\#Y$ oznaczamy liczność zbioru studentów. W podobny sposób możemy obliczyć procent studentów, którzy zaliczyli przedmiot k.

Obliczanie średniej arytmetycznej oznacza, że traktujemy ocenę 2 jak połowę oceny 4. To nie zawsze ma sens, przecież 2 to nie jest pół zaliczenia, ale brak zaliczenia.

Obliczywszy dla każdego przedmiotu „popularność" (średnia ocena z ankiet) i „trudność" (zdawalność), możemy te dwie charakterystyki przedstawić na wykresie punktowym. Gdy wybieralnych przedmiotów jest dużo, z wykresu odczytać można więcej niż z tabeli liczb. Poniższy wykres przedstawia sytuacje z Wydziału MIMUW. Każdy przedmiot jest przedstawiony za pomocą pojedynczej kropki. Nazwy przedmiotów usunięto, zamazałyby one cały wykres. Jak widzimy, w ofercie są przedmioty o bardzo wysokiej zdawalności i o niższej zdawalności. Co ciekawe, wydaje się, że zdawalność nie ma większego związku z oceną przedmiotu w ankietach.





Płetwonurkowanie, czyli schodzimy głębiej

Pierwsze z zadanych pytań dotyczyło chwili obecnej. *Na jakie przedmioty zapisana jest ta urocza brunetka w **tym** semestrze?* Drugie pytanie dotyczyło danych historycznych, ale wynik rodzi pokusę, by ekstrapolować go na przyszłość. Gdy **ja** będę wybierał przedmiot na przyszły semestr, jaka jest szansa, że ten przedmiot będzie **mnie** się podobał i jaka jest szansa, że **ja** będę miał trudności z zaliczeniem tego przedmiotu?

Chciałbym więc mieć bardziej spersonalizowaną mapę, opartą na danych studentów **takich jak ja**. Ale, oczywiście, nie ma studentów dokładnie takich jak ja. Studenci są różni i żadnych dwóch nie jest takich samych (prowadzę zajęcia dla studentów od wielu lat i jeszcze nie zdarzyło mi się nie móc rozróżnić dwóch osób). Przewidując więc, jak bardzo dany przedmiot będzie mi się podobał, będę szukał opinii studentów **podobnych** do mnie. To już głębsze rejony oceanu. Musimy jakoś określić miarę podobieństwa studentów.

W języku potocznym łatwiej operować terminem *podobieństwo*. Ale w opisie matematycznym znacznie łatwiej operować przeciwieństwem podobieństwa, czyli *niepodobieństwem*. Mając jedno, łatwo jest wyznaczyć drugie, dlatego poniżej będziemy używać raz jednego, raz drugiego terminu.

Interesuje nas określenie miary niepodobieństwa dwóch studentów, oznaczmy ich przez x i y. Ich niepodobieństwo będziemy oznaczać $d(x,y)$. Niektóre algorytmy analizy danych wymagają od funkcji $d(x,y)$ symetrii

Rozwiązanie zadania M 1412.
Bez utraty ogólności możemy założyć, że $\bar{a} = 0$. Wówczas

$$\max_{1\leqslant i\leqslant n} a_i \geqslant 0 \geqslant \min_{1\leqslant i\leqslant n} a_i.$$

Zauważmy, że dla $1 \leqslant k < l \leqslant n$ mamy

$$\begin{aligned}|a_1 - a_2| + |a_2 - a_3| + \ldots + |a_{n-1} - a_n| &\geqslant \\ \geqslant |a_k - a_{k+1}| + \ldots + |a_{l-1} - a_l| &\geqslant \\ \geqslant |a_k - a_l|.&\end{aligned}$$

Zatem w szczególności

$$\begin{aligned}&|a_1 - a_2| + |a_2 - a_3| + \ldots + |a_{n-1} - a_n| \geqslant \\ &\geqslant |\max a_i - \min a_i| = \max a_i - \min a_i \geqslant \\ &\geqslant \max\{\max a_i, -\min a_i\} = \\ &= \max_{1\leqslant i\leqslant n} |a_i| \geqslant \frac{1}{n}\sum_{i=1}^{n} |a_i|.\end{aligned}$$

$(d(x,y) = d(y,x))$, rozróżnialności $(d(x,y) = 0 \Leftrightarrow x = y)$ i spełniania warunku trójkąta. Nie zawsze jednak da się wszystkie te wymagania spełnić. Poniżej opiszemy metody wymagające jedynie symetrii.

Funkcja $d(x,y)$ opisuje, jak bardzo studenci x i y są niepodobni. Ale można być podobnym/niepodobnym na wiele sposobów. Na przykład, są podobni, bo chodzili do tej samej szkoły. Podobni, bo mają zbliżone wyniki z rekrutacji. Podobni, bo na pierwszym semestrze mieli podobne oceny.

Z tego powodu wygodnie jest definiować podobieństwo przez składowe, np. tak

$$d(x,y) = \sum_i w_i d_i(x,y),$$

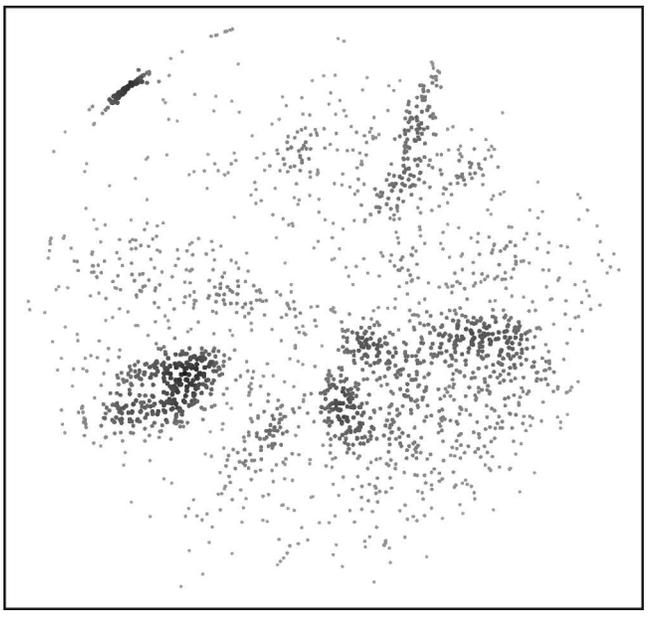
gdzie $d_i(x,y)$ to składowa i niepodobieństwa, a w_i to waga, z jaką ta składowa wpływa na końcową wartość niepodobieństwa. Wagi są istotne, ponieważ jeżeli będziemy chcieli przewidzieć, czy jakiś przedmiot nam się będzie podobał, to będziemy szukać studentów o podobnych gustach, będziemy więc większą wagę przykładać do składowych opisujących gusta (np. podobne wybory przedmiotów obieralnych). Jeżeli interesować nas będzie trudność przedmiotu, to będziemy większą wagę przykładać do składowych związanych z biegłością w danej dziedzinie.

Mając miarę podobieństwa/niepodobieństwa, możemy chcieć wiedzieć, czy nie da się ze zbioru wszystkich studentów wydzielić podzbiorów studentów „podobnych". Aby zilustrować to zagadnienie, rozważmy następującą funkcję niepodobieństwa

$$d_z(x,y) = 1 - \frac{\#(K_x \cap K_y)}{\#(K_x \cup K_y)},$$

gdzie K_x to zbiór przedmiotów, na które zapisany jest student x, K_y to zbiór przedmiotów, na które zapisany jest student y.

Zauważmy, że $d_z(x,y)$ może być równe 0, nawet gdy $x \neq y$. Wystarczy, by x i y byli zapisani na te same przedmioty. Trzeba z tym żyć albo zmienić funkcję niepodobieństwa.



W naszym przykładzie dla Wydziału MIMUW mamy dane dla trzech tysięcy studentów. Macierz niepodobieństwa dla każdej pary studentów ma więc trzy tysiące wierszy i trzy tysiące kolumn. Jak coś zobaczyć w takiej macierzy?

Z pomocą przyjdzie nam technika skalowania wielowymiarowego (*Multidimensional scaling*), która pozwala na znalezienie r-wymiarowej reprezentacji obiektów dobrze odwzorowującej niepodobieństwo między obiektami. Dla $r = 2$ otrzymamy dwuwymiarowy opis dla każdego studenta, który możliwie dobrze zachowa opisaną funkcję niepodobieństwa. Przypadek $r = 2$ jest interesujący, ponieważ dwa wymiary można przedstawić na wykresie punktowym.

Zamieszczony obok wykres przedstawia wynik skalowania dla wspomnianych trzech tysięcy studentów. Każdy punkt odpowiada jednemu studentowi. Punkty bliskie sobie powinny odpowiadać studentom podobnym według zadanej miary niepodobieństwa, a punkty dalekie od siebie odpowiadać powinny studentom niepodobnym.

Jak widzimy, punkty nie tworzą jednej chaotycznej chmury, ale odnaleźć można podgrupy studentów wybierających podobne przedmioty! Niektóre

z nich odpowiadają specjalnościom, widać jednak, że granice między grupami nie są ostre i wielu studentów jest „gdzieś pomiędzy".

Co jeszcze możemy zrobić z macierzą niepodobieństwa? Wykorzystajmy ją do oszacowania/odgadnięcia $o(x,k)$, czyli opinii studenta x o przedmiocie k.

Przyjmując, że opinia studenta x będzie podobna do opinii studentów podobnych do x, mogę np. uśrednić opinię 10 studentów najbardziej podobnych do x lub studentów o niepodobieństwie mniejszym od 0,1. W wielu przypadkach dobre wyniki daje uśrednianie opinii wielu studentów poprzez ważenie głosu studenta y w zależności od tego, jak bardzo jest on podobny do studenta x. Im bardziej podobny, tym ważniejsza będzie jego opinia.

Tak otrzymujemy wzór na ocenę opinii studenta x o przedmiocie k jako ważoną średnią opinii innych studentów

$$\hat{o}(x,k) = \frac{\sum_{y \in Y} o(y,k) \cdot h(d(x,y))}{\sum_{y \in Y} h(d(x,y))},$$

gdzie Y to zbiór wszystkich studentów, $h(d(x,y))$ to funkcja określająca, jak niepodobieństwo studentów x i y przekłada się na wagę opinii studenta y w przewidywaniu gustów studenta x. Jednym z częstych wyborów jest $h(t) = \exp(-t^2)$.

Daszek nad funkcją o oznacza szacowanie opinii. Aby poznać prawdziwą opinię $o(x,k)$, student x musiałby się na przedmiot k zapisać, a później musiałby go ocenić.

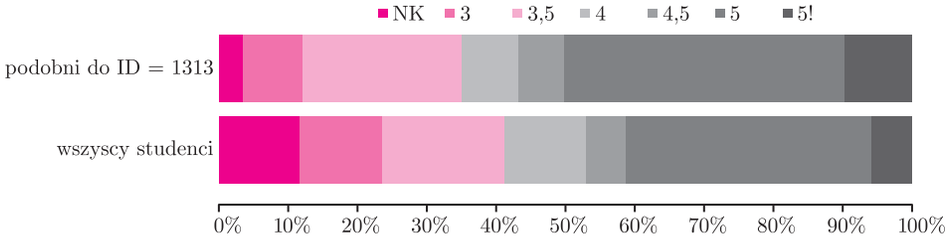
Możemy teraz oszacować opinię studenta x o każdym z przedmiotów. Podobnie możemy oszacować jego szanse zaliczenia przedmiotu i przedstawić mu bardziej spersonalizowaną mapę „trudność–popularność".

Jak przejść z powyżej opisywanej ważonej średniej do ważonego rozkładu? Pozostawiamy to jako łamigłówkę dla Czytelnika.

Przedstawiony wzór na „średnią" można dowolnie modyfikować. Na przykład, zamiast wyznaczania średniej zdawalności możemy wyznaczyć rozkład ocen wśród podobnych studentów ważony podobieństwem do wybranego studenta.

Zobaczmy, jak to wygląda na konkretnym przykładzie. Weźmy pod lupę przedmiot *Pakiety statystyczne*: *R i SAS*, który prowadzę. Następnie wybierzmy studenta o wdzięcznym identyfikatorze ID = 1313, który na ten przedmiot jeszcze nie uczęszczał, i zobaczmy, czy wśród studentów, którzy ten przedmiot realizowali, ci podobni do ID = 1313 otrzymywali lepsze czy gorsze oceny.

Na poniższym wykresie przedstawiono procent studentów, którzy otrzymali określoną ocenę z przedmiotu *Pakiety statystyczne*: *R i SAS*. Dolny pasek opisuje sytuacje wszystkich studentów realizujących ten przedmiot. Pasek na górze jest ważony podobieństwem do ID = 1313. Okazuje się, że „podobni" studenci mają częściej lepsze oceny. Również szanse niezaliczenia (oznaczone jako NK) są niższe niż w przypadku „losowego" studenta. Nic, tylko się zapisywać!



Na podobnej zasadzie działają systemy rekomendacyjne sugerujące zakup książki w księgarni internetowej czy proponujące film do obejrzenia. I podobnie jak w przypadku książek czy filmów, nie ma gwarancji, że to, co podobało się osobom o podobnych profilach do naszego, spodoba się również nam. Gwarancji nie ma, ale czasem nawet informacja „niepewna" może być użyteczna.