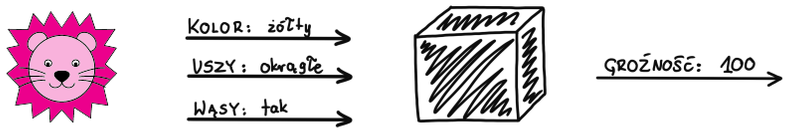
Czarna skrzynka

*Oskar SKIBSKI**

* Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

Wyobraźmy sobie, że przechadzając się po lesie, znaleźliśmy czarną skrzynkę. Nie pochodzi ona jednak z samolotu (zresztą samolotowe „czarne skrzynki” są zwykle pomarańczowe). Po oględzinach okazuje się, że skrzynka potrafi robić jedną rzecz: kiedy pokaże się jej jakieś zwierzę (lub jego zdjęcie), to rozpoznaje jego kolor, kształt uszu i to, czy ma wąsy, i na podstawie tych atrybutów wyświetla informację, jak groźne jest to zwierzę.

Czarną skrzynką nazywa się teoretyczny model, w którym nie wiadomo nic o budowie wewnętrznej, a obserwować możemy jedynie „wejście” i „wyjście”.



Pokazujemy skrzynce zaprzyjaźnionego lwa. Lew jest żółty, ma okrągłe uszy, ma wąsy. Skrzynka pokazuje wynik 100 – zgadza się! Wszyscy wiemy, że lwy są groźne, a 100 to całkiem spora liczba. Pokazujemy zdjęcie świni. Świnia jest różowa, ma spiczaste uszy i nie ma wąsów. Wynik: 30 – też się zgadza (świnia nie jest zbytnio groźna, chociaż łatwo mogłaby nas stratować).

Ale skąd ta skrzynka to wie? Jest to czarna skrzynka, zgodnie z zasadami nie możemy więc, niestety, zajrzeć do środka. Pewnie zaszyty jest tam jakiś algorytm uczenia maszynowego – na etapie produkcji pokazano mu wiele, wiele zwierząt i wprowadzono informację, jak bardzo są groźne, a on na tej podstawie wypracował w sobie przekonanie, że wie, po czym poznać groźność zwierzęcia. No i teraz, widząc żółte zwierzę z okrągłymi uszami i wąsami, swoją inteligencją (może trochę sztuczną, ale jednak) ocenia je na 100 punktów.

Nadzieję na to, że w pełni zrozumiemy działanie skrzynki, musimy niestety porzucić. Ale czy przynajmniej jest jakiś sposób, aby zrozumieć, dlaczego otrzymujemy taki a nie inny rezultat? Na przykład dowiedzieć się, która cecha lwa spowodowała tak wysoki wynik? Czy to przez kolor, czy uszy, czy wąsy?

Pierwszym pomysłem byłoby oberwanie lwu wąsów albo przemalowanie go na inne kolory, pokazanie skrzynce i porównanie wyników ze zwykłym lwem. Pomysł ten trzeba jednak uznać za ryzykowny. Ponadto opieralibyśmy się wówczas na sztucznych, a nie prawdziwych danych. Dla nich algorytm uczenia maszynowego w ogóle nie musi przecież dobrze działać. Moglibyśmy też spróbować kupić skrzynkę, która nie bierze konkretnej cechy pod uwagę. Ten pomysł jest kosztowny i dla nas niedostępny – jesteśmy przecież w lesie. Ograniczymy się zatem do pokazywania skrzynce różnych zwierząt i zapisywania wyników. Tylko czy da się z tych wyników wyciągnąć jakieś sensowne wnioski?

Lundberg, Scott M., and Su-In Lee, „A unified approach to interpreting model predictions”, *Advances in Neural Information Processing Systems* 30 (2017).

Na nasze szczęście w krzakach niedaleko czarnej skrzynki znajdujemy pracę autorstwa Scotta Lundberga i Su-In Lee z 2017 roku. Opisują oni w niej metodę SHAP, która mimo swojej nazwy, przypominającej raczej o obiedzie, szybko stała się standardowym narzędziem wykorzystywanym w analizie algorytmów uczenia maszynowego. Użyjemy jej także my.

Załóżmy, że pokazaliśmy maszynie 1000 różnych zwierząt. Średnia ocena groźności tych zwierząt wyszła 20. Okazuje się, że żółte zwierzęta miały średnią groźność 50, te z okrągłymi uszami 10, a te z wąsami 80. Czy to nam wystarczy? Coś nam to już mówi, ale cechy są często komplementarne lub substytucyjne, tzn. uzupełniają się albo są wymienne. Na przykład sama długość zębów nie mówi za wiele, jeżeli nie znamy długości całego zwierzęcia (trzycentymetrowe zęby u dwumetrowego kotowatego to nic ciekawego, jednak u zwierzęcia wielkości żaby zwróciłyby naszą uwagę).

Dlatego lepszym podejściem jest rozpatrzenie także grup cech. Żółte zwierzęta z okrągłymi uszami mają średnią groźność 40 (Kubuś Puchatek pewnie zaniża tu średnią). Żółte zwierzęta z wąsami – 90, a zwierzęta z wąsami i okrągłymi

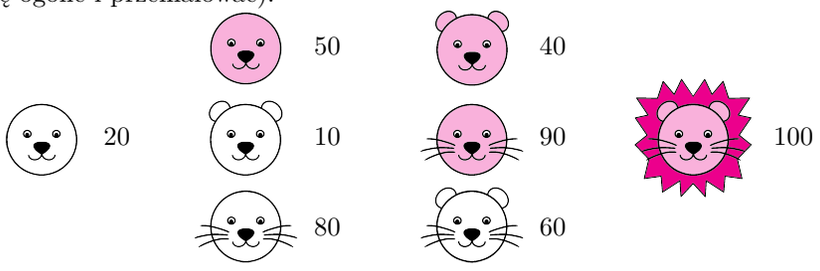
Rozwiązanie zadania M 1735.
Dla pewnej liczby całkowitej d mamy:

$$a_2 = a_1 + d,\ \ldots,\ a_n = a_1 + (n-1)d.$$

Dla $1 \leqslant i \leqslant n-1$ mamy $i \mid a_i = a_1 + (i-1)d$, zatem $i \mid a_1 - d$. Podobnie uzasadniamy $n \nmid a_1 - d$.

Gdyby $n = ab$ dla pewnych względnie pierwszych liczb całkowitych $1 < a, b < n$, to $a \mid a_1 - d$ oraz $b \mid a_1 - d$, skąd $n \mid a_1 - d$, a to jest sprzeczność. Zatem n jest potęgą liczby pierwszej.

uszami – 60. Wyniki te podsumowuje następujący diagram (lwa musieliśmy trochę ogolić i przemalować):



Oczytany Czytelnik już wie, na co patrzy – jest to przecież (prawie) gra koalicyjna! Graczami są nasze trzy cechy: żółty kolor, okrągłe uszy i wąsy, a grę opisuje funkcja, nazywana *funkcją charakterystyczną*, która każdej niepustej grupie graczy przypisuje pewną wartość. No właśnie, niepustej, a pusta koalicja powinna mieć wartość zero. Dlatego aby uzyskać grę koalicyjną, musielibyśmy od wszystkich wartości odjąć 20.

Gry koalicyjne pojawiały się już w *Delcie* w analizie sieci terrorystycznych, Δ^{11}_{16}, siły partii politycznych, Δ^{11}_{20}, i podziału lodów, Δ^{4}_{22}.

Do poznania istotności każdej z cech wystarczy, że użyjemy jednej z metod podziału w grach koalicyjnych. Dostaniemy wtedy informację, że na groźność lwa składa się w jakiejś części kolor, w jakiejś – uszy, a w jakiejś – wąsy. Metod jest wiele, ale jak przychodzi co do czego, stosowana jest jedna: wartość Shapleya. Tak jest też w metodzie SHAP, co tłumaczy jej nazwę.

Wprowadźmy trochę notacji do naszej leśnej historii. Mamy pewien zbiór m cech C i pewien zbiór m–krotek X. Naszym wejściem są właśnie owe m-krotki – kolejne ich pozycje odpowiadają kolejnym cechom, jak w przykładzie na marginesie. Nasza znaleziona czarna skrzynka to funkcja B, która dla każdej krotki zwraca liczbę rzeczywistą. Dla ustalonej krotki $x \in X$ definiujemy grę $f_x : 2^C \to \mathbb{R}$ następująco: dla każdej grupy cech $S \subseteq C$ mamy:

$$f_x(S) = \mathrm{avg}_{y \in X : x_S = y_S} B(y),$$

gdzie x_S to krotka x obcięta do elementów z S. Wartość $f_x(S)$ to zatem średnia wartość zwracana przez czarną skrzynkę dla wszystkich możliwych krotek z X, które mają te same wartości dla cech z S. W szczególności $f_x(C)$ to ocena groźności zwierzęcia stojącego za krotką x.

W naszej sytuacji mamy:
$C = \{(\mathrm{k})\mathrm{olor}, (\mathrm{u})\mathrm{szy}, (\mathrm{w})\mathrm{ąsy}\}$,
$X = \{(\text{żółty}, \text{okrągłe}, \text{tak}),$
$(\text{różowy}, \text{spiczaste}, \text{nie}), \dots\}$.

Czarna skrzynka B działa tak:
$B(\text{żółty}, \text{okrągłe}, \text{tak}) = 100$,
$B(\text{różowy}, \text{spiczaste}, \text{nie}) = 30, \dots$

Grę f_x definiujemy dla
$x = (\text{żółty}, \text{okrągłe}, \text{tak})$ jak na obrazku:
$f_x(\emptyset) = 20$, $f_x(k) = 50$, $f_x(u) = 10$,
$f_x(w) = 80$, $f_x(ku) = 40$, $f_x(kw) = 90$,
$f_x(uw) = 60$, $f_x(kuw) = 100$.

Powyżej i też dalej używamy skrótowego zapisu zbiorów, np. abc reprezentuje zbiór $\{a, b, c\}$ itp.

Naszym celem jest przedstawienie wyniku $B(x)$ w następującej postaci:

$$(*) \qquad B(x) = \phi_0 + \sum_{c \in C} \phi_c,$$

gdzie ϕ_c to wpływ cechy c, a ϕ_0 to wartość bazowa równa po prostu $f_x(\emptyset)$, czyli średnia groźność wszystkich zwierząt.

Naturalnym pomysłem przy ocenie danej cechy jest patrzenie na to, ile ona wnosi do wartości innych grup. Wkład marginalny cechy c do grupy $S \subseteq C \setminus \{c\}$ to po prostu różnica między wartością grupy z daną cechą i bez niej: $f_x(S \cup \{c\}) - f_x(S)$. Wartość Shapleya to właśnie ważona suma wkładów marginalnych:

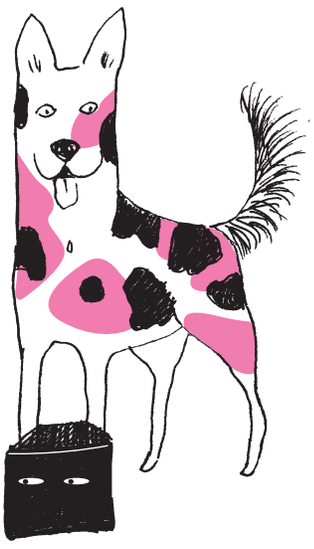
$$\phi_c(f_x) = \sum_{S \subseteq C \setminus \{c\}} \frac{|S|!(|C| - |S| - 1)!}{|C|!} \left(f_x(S \cup \{c\}) - f_x(S)\right).$$

Te dziwne wagi w wartości Shapleya biorą się z następującej interpretacji: Wyobraźmy sobie, że dodajemy cechy jedna po drugiej i każdą z nich oceniamy, patrząc na jej wkład marginalny do obecnego już zbioru cech. Np. dla kolejności kolor-uszy-wąsy mamy następujące oceny: $\phi_k = 50 - 20 = 30$, $\phi_u = 40 - 50 = -10$ i $\phi_w = 100 - 40 = 60$. Patrzenie na jedną kolejność nie byłoby jednak sprawiedliwe, dlatego rozpatrujemy średnią po wszystkich kolejnościach i wychodzą nam takie właśnie wagi: jest $|C|!$ różnych kolejności i dokładnie w $|S|!(|C| - |S| - 1)!$ cechę c dodajemy do cech S.

W standardowych grach koalicyjnych suma ocen wszystkich graczy, czyli u nas cech, jest równa $f_x(C)$. Skoro $f_x(\emptyset)$ niekoniecznie jest zerem, to tak zdefiniowane oceny nie zsumują się do $f_x(C)$, tylko do $f_x(C) - f_x(\emptyset)$, czyli tak, jak chcemy, biorąc pod uwagę wartość bazową.

A czemu mamy używać akurat wartości Shapleya do wyznaczenia ϕ_c? Powyższy wzór wygląda mało intuicyjnie. Jeżeli ma być dobrze, to inaczej się jednak nie da! Dowodzi tego szereg aksjomatycznych charakteryzacji wartości Shapleya, czyli wyników, które pokazują, że jest to jedyna metoda mająca wiele pożądanych własności.

W naszym przypadku szukamy wartości ϕ_c, które będą spełniać równanie $(*)$. Naturalna jest także symetria – mówiąca, że cechy, które są symetryczne



w grze f, powinny dostać taką samą ocenę (tzn. jeżeli $f(S\cup\{c\}) = f(S\cup\{c'\})$ dla każdego $S \subseteq C\setminus\{c,c'\}$, to $\phi_c = \phi_{c'}$). Autorzy metody SHAP powołali się także na monotoniczność zaproponowaną przez Peytona Younga, która mówi, że jeżeli w grze f cecha c wypada lepiej niż w f', to powinna mieć w tej grze wyższą ocenę. To, gdzie cecha wypada lepiej, mierzymy, porównując wkłady marginalne koalicja po koalicji; jeżeli do wszystkich koalicji w f wnosi co najmniej tyle co w f', a do którejś koalicji nawet więcej, to wtedy wypada lepiej. Formalnie, dla dowolnych gier f, f':

$$\big(\forall_{S\subseteq C\setminus\{c\}} f(S\cup\{c\}) - f(S) \geqslant f'(S\cup\{c\}) - f'(S)\big) \Rightarrow \phi_c(f) \geqslant \phi_c(f').$$

Okazuje się teraz, że wartość Shapleya jest jedyną metodą oceny, która zapewni nam te trzy własności. Dowód jest naprawdę elegancki, przedstawmy go zatem.

Twierdzenie 1. *Jeżeli wartości $(\phi_c)_{c\in C}$ spełniają równanie $(*)$, symetrię i monotoniczność, to muszą być to oceny zwracane przez wartość Shapleya.*

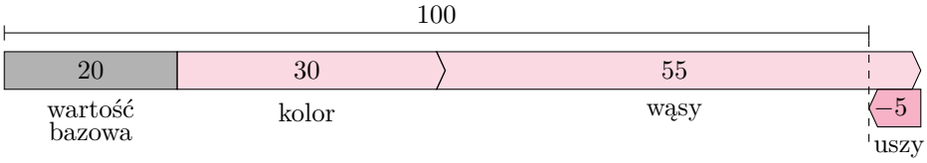
Dowód. Łatwo sprawdzić, że wartość Shapleya spełnia wszystkie trzy własności. Pokażemy, że jest tylko jedna metoda oceny, która je spełnia.

Dla dowolnej grupy $S \subseteq C$ zdefiniujmy prostą grę u_S tak: $u_S(T) = 1$, jeżeli $S \subseteq T$, $u_S(T) = 0$, wpp. Łatwo zauważyć, że cechy spoza S mają zerowe wkłady marginalne do dowolnej grupy T. Dowolną grę f możemy teraz jednoznacznie przedstawić jako $f = \sum_{S\subseteq C} \alpha_S u_S$ dla pewnych stałych $(\alpha_S)_{S\subseteq C}$. *Złożonością gry*, $Z(f)$, nazwiemy liczbę niezerowych stałych α_S. Przeprowadźmy indukcję po złożoności gry.

Jeżeli $Z(f) = 0$, to w grze wszystkie wartości są zerowe i z symetrii oraz równania $(*)$ mamy $\phi_c(f) = 0$ dla każdego $c \in C$. Załóżmy, że udało nam się udowodnić tezę dla gier o złożoności mniejszej niż k (dla pewnego $k > 0$). Weźmy grę o złożoności k i niech $S_1, \ldots, S_k$ będą grupami z niezerowymi wagami. Rozpatrzmy najpierw cechę c, która nie należy do wszystkich grup $S_1, \ldots, S_k$. W takiej sytuacji możemy, opierając się na grze f i jej wagach $(\alpha_S)_{S\subseteq C}$, zdefiniować grę $f' = \sum_{S\subseteq C, c\in S} \alpha_S u_S$. Gra f' powstała przez usunięcie z f gier u_S takich, do których cecha c ma zerowy wkład marginalny (bo $c \notin S$), więc wkłady marginalne cechy c są w niej takie same jak w grze f. Z monotoniczności ocena c jest więc taka sama, czyli znana, bo f' ma przecież niższą złożoność.

Pozostaje nam zatem rozpatrzyć cechy, które należą do wszystkich koalicji $S_1, \ldots, S_k$. One są jednak symetryczne w grze f i możemy wyznaczyć sumę ich ocen z równania $(*)$ oraz ocen cech, które nie pojawiają się we wszystkich zbiorach $S_1, \ldots, S_k$. Oceny tych cech też są zatem jednoznacznie wyznaczone, co kończy dowód. □

Nie mamy już więc wątpliwości, że to wartości Shapleya powinniśmy użyć, aby odpowiedzieć na pytanie, czemu lew otrzymał aż 100 punktów. Stosując równanie do naszej gry, otrzymujemy następujące wyniki:



A więc to głównie przez wąsy lew wydaje się taki groźny! Czegoś nowego się dzisiaj dowiedzieliśmy.

A co, jak zastosujemy naszą metodę do innego zwierzęcia? Wyniki mogą być kompletnie inne. Może się na przykład okazać, że wąsy w zestawieniu z innymi cechami dodają obliczu zwierzęcia łagodności (np. przypominają nam o naszym poczciwym dziadku). Może też być tak, że kluczową rolę odegrają akurat uszy. Wszystko też zależy od konkretnych danych: gdybyśmy pokazali skrzynce inne zwierzęta, moglibyśmy dostać zupełnie inne wyniki. Niewątpliwie nasza czarna skrzynka skrywa jeszcze wiele tajemnic, ale udało nam się chociaż trochę zajrzeć do jej środka, i to tylko patrząc na nią z zewnątrz.

Przykładowo:
$u_{ku}(T) = \begin{cases} 1 & \text{jeśli } T = ku, kuw, \\ 0 & \text{wpp.} \end{cases}$

Naszą oryginalną grę możemy teraz przedstawić tak:

$$f_x = 20u_\emptyset + 30u_k - 10u_u + \\ + 60u_w - 20u_{kw} - 10u_{uw} + 30u_{kuw}.$$

Zdefiniujmy na przykład grę $f' = 30u_k - 20u_{kw} + 30u_{kuw}$ powstałą przez wykasowanie z f_x gier u_S t.że $k \notin S$. Dostajemy następujące wkłady marginalne dla żółtego koloru:
$f'(k) - f'(\emptyset) = 30 - 0 = 30$,
$f'(ku) - f'(u) = 30 - 0 = 30$,
$f'(kw) - f'(w) = 10 - 0 = 10$,
$f'(kuw) - f'(uw) = 40 - 0 = 40$.
Łatwo sprawdzić, że w oryginalnej grze f_x wkłady marginalne są takie same ($50-20$, $40-10$, $90-80$ i $100-60$).

Przykładowo, dla żółtego koloru mamy:

$$\phi_k = \frac{1}{3}(50-20) + \frac{1}{6}(40-10) + \\ + \frac{1}{6}(90-80) + \frac{1}{3}(100-60) = 30.$$