Prostota i niepewność

Ta część zajmuje się kwestią, jak dokonywać przewidywań w nieznanym środowisku. Po krótkim opisie ważnych postaw filozoficznych dotyczących rozumowania indukcyjnego i wnioskowania, dokładniej opisujemy, co rozumiemy przez indukcję, i motywujemy, dlaczego możemy skupić się na zadaniach przewidywania sekwencji. Najważniejszą koncepcją jest zasada brzytwy Ockhama (prostota). Rzeczywiście, można wykazać, że najlepszym sposobem dokonywania przewidywań jest oparcie się na najkrótszym (najprostszym) opisie sekwencji danych, jaki do tej pory widziano. Najbardziej ogólne skuteczne opisy można uzyskać za pomocą ogólnych funkcji rekurencyjnych lub równoważnie, używając programów na maszynach Turinga, zwłaszcza na uniwersalnej maszynie Turinga. Długość najkrótszego programu opisującego dane nazywa się złożonością Kołmogorowa danych. Niestety złożoność Kołmogorowa nie jest skończenie obliczalna, co sprawia, że konieczne jest wprowadzenie kilku słabszych pojęć obliczeniowych. Teoria prawdopodobieństwa jest potrzebna do radzenia sobie z niepewnością. Środowisko może być procesem stochastycznym (np. domy gry lub fizyka kwantowa), który można opisać za pomocą „obiektywnych" prawdopodobieństw. Ale również niepewna wiedza o środowisku, która prowadzi do przekonań na jego temat, może być modelowana za pomocą „subiektywnych" prawdopodobieństw. Stare pytanie pozostawione otwarte przez subiektywistów, jak wybierać prawdopodobieństwa a priori, rozwiązuje uniwersalne prawdopodobieństwo a priori Solomonoffa, które jest luźno powiązane ze złożonością Kolmogorova. Głównym wynikiem Solomonoffa jest to, że uniwersalne (subiektywne) a posteriori zbiega się do prawdziwego (obiektywnego) środowiska (prawdopodobieństwa) fi. Jedynym założeniem dotyczącym /x jest to, że /i (które nie musi być znane!) jest obliczalne. Problem nieznanego środowiska /i jest zatem rozwiązany dla wszystkich problemów typu indukcyjnego, takich jak przewidywanie sekwencji i klasyfikacja. Na koniec pokazujemy (nie)istnienie uniwersalnych prawdopodobieństw a priori dla innych wprowadzonych pojęć obliczalności. Po wolniejsze i bardziej szczegółowe wprowadzenie do złożoności Kolmogorowa i indukcji Solomonowa oraz większości dowodów należy zapoznać się z doskonałą książką Li i Vitanyi.

Wprowadzenie

Jednym z bardzo ważnych i nietrywialnych aspektów inteligencji jest wnioskowanie indukcyjne. Po omówieniu kilku przykładów przedstawiamy podstawy filozoficzne, a następnie sekwencyjne ustawienie, które nas interesuje.

Przykłady problemów indukcyjnych

Jakie jest prawdopodobieństwo, że jutro wzejdzie słońce? Nasuwa się kilka odpowiedzi: Prawdopodobieństwo jest nieokreślone, ponieważ nigdy nie było eksperymentu, który sprawdzałby istnienie słońca jutro (problem klasy odniesienia). Prawdopodobieństwo wynosi 1, ponieważ we wszystkich eksperymentach w przeszłości słońce wzeszło. Prawdopodobieństwo wynosi 1 - ε, gdzie ε << 1 to proporcja gwiazd we wszechświecie, które eksplodują jako supernowa na dzień. Prawdopodobieństwo można wywnioskować z typu, wieku, rozmiaru i temperatury słońca, nawet jeśli nigdy nie obserwowaliśmy innej gwiazdy o dokładnie takich właściwościach. Prawdopodobieństwo wynosi d+1/d+2 , gdzie d to liczba minionych dni, w których wzeszło słońce (reguła Laplace'a). Innym przykładem są rozszerzające się sekwencje binarne, takie jak 100100100001111- 1101101010100... Sekwencja wygląda losowo, więc prawdopodobnie jest to również jej kontynuacja. Dokładniejsze przyjrzenie się ujawnia, że ciąg jest rozwinięciem binarnym TT, więc prawdopodobnie lepiej będzie przewidzieć jego kontynuację 010001.... Wolimy odpowiedź 010001..., ponieważ widzimy w ciągu więcej struktury niż tylko losowe cyfry.

Jako inny przykład rozważmy ciągi liczb $x_1,x_2,x_3,x_4,...$, takie jak 1,2,3,4,... testów IQ. Praktycznie każdy przewiduje $x_5 = 5$ jako następną liczbę, ponieważ xi = i dla i = 1...4, ale $x_5 = 29$ można również argumentować, ponieważ $x_i = i^4 - 10i^3 + 35i^2 - 49i + 24$. Wolimy odpowiedź 5, ponieważ relacja liniowa obejmuje mniej dowolnych parametrów niż wielomian 4-tego rzędu. Trudniejsze jest 2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59,?. Następną liczbą może być 61, ponieważ jest to następna liczba pierwsza, lub 60, ponieważ jest to rząd następnej grupy prostej. Większość odpowie 61, ponieważ liczby pierwsze są bardziej znanym pojęciem niż grupy proste. Powyższe przykłady pokazują, że znajdowanie reguł przewidywania dla każdego konkretnego (nowego) problemu jest uciążliwe i podatne na nieporozumienia lub sprzeczności. Potrzebujemy formalnej ogólnej teorii przewidywania.

Ockham, Epikur, Hume, Bayes, SolomonofF

Ogólnie rzecz biorąc, indukcja to proces przewidywania przyszłości na podstawie przeszłości, a dokładniej, to proces znajdowania reguł w (przeszłych) danych i wykorzystywania tych reguł do odgadywania przyszłych danych. Przewidywanie pogody, prognozowanie giełdowe lub ciągłe serie liczbowe w teście IQ to nietrywialne przykłady. Tworzenie dobrych prognoz odgrywa centralną rolę w naturalnej i sztucznej inteligencji w ogóle, a w uczeniu maszynowym w szczególności. Z jednej strony indukcja wydaje się mieć miejsce w życiu codziennym poprzez znajdowanie regularności w poprzednich obserwacjach i wykorzystywanie ich do przewidywania przyszłości. Z drugiej strony ta procedura wydaje się dodawać wiedzę o przyszłości z poprzednich obserwacji. Ale jak możemy wiedzieć coś o przyszłości? Ten dylemat i zasada indukcji w ogóle mają długą historię filozoficzną:

• Zasada Epikura wielorakich wyjaśnień (342?-270? p.n.e.) Jeśli więcej niż jedna teoria jest zgodna z obserwacjami, zachowaj wszystkie teorie.

• Zasada brzytwy Ockhama (prostota) (1290?-1349?) Bytów nie należy mnożyć ponad konieczność - lub - zachować najprostszą teorię zgodną z obserwacjami.

• Negacja indukcji Hume'a (1711-1776) Wiara w możliwość prawdziwej indukcji nie może być uzasadniona racjonalnie.

• Reguła Bayesa dla prawdopodobieństw warunkowych (1702-1761) . Mówi nam, jak aktualizować nasze przekonania/prawdopodobieństwa podczas pozyskiwania nowych danych. Solomonoff sprytnie zjednoczył zasady Epikura, Ockhama i Bayesa w jedną formalną uniwersalną teorię wnioskowania indukcyjnego. Spośród wszystkich możliwych schematów indukcji jest to optymalna metoda dokonywania przewidywań.

Konfiguracja problemu

Każdy problem indukcyjny można sformułować jako zadanie przewidywania sekwencji. Jest to najwyraźniej zilustrowane w domenie przewidywania szeregów czasowych. Po zaobserwowaniu danych x_t w czasach $t < n$, zadaniem jest przewidzenie symbolu n-tego x_n z sekwencji $x_1...x_{n-1}$.Klasyfikacja może być również postrzegana jako zadanie przewidywania sekwencji. Zadanie klasyfikowania nowej instancji z_n po zobaczeniu par (instancja, klasa) $(z_1,c_1),...,(z_{n-1},c_{n-1})$ można sformułować jako przewidywanie kontynuacji sekwencji $z_1c_1...z_{n-1}c_{n-1}z_n$. Uczenie maszynowe często zajmuje się znalezieniem prawdziwego, predykcyjnego lub przyczynowego modelu na podstawie zaobserwowanych danych. Ten krok jest ważny dla zrozumienia rozważanej domeny. Zrozumienie jest często celem samym w sobie, ale ostatecznie celem jest zastosowanie modelu do dokonywania przewidywań. W tym ujęciu uczenie się modelu jest tylko krokiem pośrednim. Bezpośrednie badanie przewidywań opartych na wcześniejszych obserwacjach bez omawiania modeli zostało nazwane

podejściem prekwencyjnym przez Dawida do przewidywań sekwencji, a wnioskowanie transdukcyjne przez Vapnika do klasyfikacji i regresji. Kilka trudnych kwestii jest unikanych przez porzucenie modeli. Obejmuje to pytania o spójność modelu, tj. czy prawdziwy model może być nauczony i jak oddzielić szum od użytecznych danych . Można nawet pójść o krok dalej i zapytać, dlaczego chcemy tworzyć przewidywania. Zazwyczaj celem przewidywania jest maksymalizacja zysku/wartości lub równoważnie minimalizacja straty. Rozważając tylko zyski lub straty, unika się pytań o to, czy algorytmy przewidywania zbiegają się do najlepszego możliwego algorytmu przewidywania (tj. czy są samostrojące). Algorytmy, dla których strata zbiega się do minimalnej możliwej straty, nazywane są samooptymalizującymi . Jest to słabsze zapotrzebowanie niż zdolność do samostrojenia, ale często jest to wszystko, na czym nam naprawdę zależy. Głównym celem jest badanie algorytmów, które minimalizują stratę. Zbieżność rozkładów prawdopodobieństwa a posteriori lub samych algorytmów lub modeli jest rozważana tylko wtedy, gdy jest to przydatne dla ostatecznego celu minimalizacji strat. Podsumowując nasze podejście:

• Każdy problem indukcyjny można sformułować jako zadanie przewidywania sekwencji.

• Oddzielenie szumu od danych nie jest konieczne w tym ustawieniu.

Po wyjaśnieniu podejścia musimy teraz zagłębić się w matematykę, zanim będziemy mogli przedstawić schemat indukcji SolomonofF.

Algorytmiczna teoria informacji

W tej sekcji podajemy bardzo krótkie wprowadzenie do złożoności Kolmogorowa.

Definicje i notacja

Piszemy IN={1,2,3,...} dla zbioru liczb naturalnych, IB^* dla zbioru skończonych ciągów binarnych i IB^∞ dla zbioru nieskończonych ciągów binarnych. Używamy liter i,k,n dla liczb naturalnych, x,y,z dla skończonych ciągów, ε dla pustego ciągu, 1^n ciągu n jedynek, $\ell(x)$ dla długości ciągu x, i ω dla nieskończonych ciągów. Piszemy xy dla konkatenacji ciągu x z y. Każdy przeliczalny zbiór można zidentyfikować za pomocą IN za pomocą bijekcji. Ciąg możemy interpretować jako binarną reprezentację liczby naturalnej. Niestety, naiwna identyfikacja nie będzie unikatowa, ponieważ na przykład ciągi 00101 i 101 reprezentują liczbę 5. Otrzymujemy bijekcję, jeśli mapujemy x na liczbę naturalną, która ma reprezentację binarną 1x (x z prefiksem 1). Odejmujemy 1 od tej liczby, ponieważ potrzebujemy bijekcji między IB^* i IN_0= {0,1,2,3,...} . Przy tej identyfikacji $\log_2(x+1)-1 < \ell(x) \leq \log_2(x+1)$. Ciąg x nazywany jest (właściwym) prefiksem y, jeśli istnieje z(≠ε) takie, że xz = y. Zbiór ciągów nazywany jest bezprefiksowym, jeśli żaden element nie jest właściwym prefiksem innego. Zbiór bezprefiksowy V nazywany jest również kodem prefiksowym. Kody prefiksowe mają ważną właściwość spełniania nierówności Krafta

$$\sum_{x\in\mathcal{P}} 2^{-\ell(x)} \leq 1. \qquad (2.1)$$

Można to wykazać, przypisując każdemu x ∈ P przedział $\Gamma_x := [0.x\,,0.x+\ 2^{-\ell(x)}) \subseteq [0,1)$, gdzie $0.x \equiv x\cdot 2^{-\ell(x)}$ jest liczbą rzeczywistą z rozwinięciem binarnym x po przecinku. Długość przedziału Γ_x wynosi $2^{-\ell(x)}$. Przedziały są rozłączne, ponieważ V jest bezprefiksowe, stąd $\sum_{x\in\mathcal{P}} 2^{-\ell(x)} = \sum_{x\in\mathcal{P}} \mathrm{length}(\Gamma_x) \leq \mathrm{length}([0,1]) = 1$. Można również pokazać odwrotność (2.1). Dla $\bar{x} := 1^{\ell(x)}0x$ zbiór $\{\bar{x} : x \in I\!B^*\}$ tworzy kod prefiksowy z $\ell(\bar{x}) = 2\ell(x)+1$. Dla

$x' := \overline{\ell(x)}x = 1^{\ell(\ell(x))}0\ell(x)x$ zbiór $\{x' : x \in \mathbb{B}^*\}$ tworzy asymptotycznie krótszy kod prefiksowy z $\ell(x') = \ell(x) + 2\ell(\ell(x)) + 1$. Parujemy ciągi x i y (oraz z) za pomocą <x,y> :=x'y (oraz <x,y,z> :=x'y'z), które są jednoznacznie dekodowalne, ponieważ x' i y' są prefiksami. Ponieważ ' służy jako separator, zapisujemy również f{x,y) zamiast f{x'y) dla funkcji f. Skracamy $\lim_{n\to\infty}[f(n) - g(n)] = 0$ przez $f(n) \stackrel{n\to\infty}{\longrightarrow} g(n)$ i mówimy, że f zbiega się do g, nie implikując, że samo $\lim_{n\to\infty} g(n)$ istnieje. Zapisujemy $f(n) \sim g(n)$ i mówimy, że f jest asymptotycznie proporcjonalne do g, jeśli ∃0< c <∞: $\lim_{n\to\infty} f(n)/g(n) = c$. Zapisujemy $a \lesssim b$, jeśli a nie jest dużo większe niż b, z precyzją niesprecyzowaną. Notacja Big-O f(x) = O(g{x)) oznacza, że istnieją stałe c i $x_0 > 0$ takie, że $|f(x)| \leq c|g(x)| \; \forall x > x_0$. Notacja małego o f{x) = o(g{x)) skraca się do $\lim_{x\to\infty} f(x)/g(x) = 0$. Piszemy $f(x) \stackrel{\times}{\leq} g(x)$ dla f(x) = 0(g{x)) i $f(x) \stackrel{+}{\leq} g(x)$ dla $f(x) \leq g(x) + O(1)$. Odpowiednie równości można zdefiniować w podobny sposób. Zachodzą one, jeśli odpowiadające im nierówności zachodzą w obu kierunkach.

Maszyny Turinga

Maszynę Turinga można uważać za zidealizowaną formę komputera. Składa się ona z taśm (pamięci), głowic odczytu/zapisu, tabeli reguł (programu) i stanu wewnętrznego (wskaźnika instrukcji). Formalną definicję można znaleźć w dowolnym podręczniku teorii obliczalności. Zbiór częściowo rekurencyjnych funkcji pokrywa się ze zbiorem funkcji obliczalnych za pomocą maszyny Turinga. Mówimy, że zbiór obiektów S = {o_1,o_2,o_3,...} można (efektywnie) wyliczyć, jeśli istnieje maszyna Turinga odwzorowująca i na <o_i>), gdzie <> jest pewnym domyślnym kodowaniem elementów w S. Znaczenie częściowo rekurencyjnych funkcji i maszyn Turinga wynika z następujących tez:

Teza 2.3 (Turing) Wszystko, co człowiek może rozsądnie uznać za możliwe do obliczenia przy użyciu ustalonej procedury, może być również obliczone przez maszynę Turinga

Teza 2.4 (Church) Klasa algorytmicznie obliczalnych funkcji numerycznych (w sensie intuicyjnym) pokrywa się z klasą funkcji częściowo rekurencyjnych

Musimy uzupełnić tezy Turinga i Churcha w następujący sposób:

Założenie 2.5 (Krótki kompilator) Jeśli mamy dwa naturalne, równoważne systemom Turinga systemy formalne F1 i F2, to zawsze istnieje jeden krótki program na systemie F2, który jest w stanie zinterpretować wszystkie programy F1

Oznacza to, że różnica rozmiaru najkrótszego opisu F1l i najkrótszego opisu F2 (czegoś) jest nie tylko ograniczona przez uniwersalną stałą, ale ta stała jest również stosunkowo mała dla naturalnych systemów formalnych. Łatwo jest formalnie przekształcić interpreter w kompilator, dołączając interpreter do programu, który ma być interpretowany i „sprzedając" wynik jako skompilowaną wersję. Rozszerza to tezy Churcha i Turinga w dwóch aspektach. Po pierwsze, mówi, że równoważność jest skuteczna, tj. że istnieje jeden program (interpreter/kompilator), który skutecznie przekształca programy F1 w programy F2. Tezy Churcha i Turinga stwierdzają tylko, że klasy funkcji obliczalnych pokrywają się, pozostawiając otwartą możliwość, że nie ma skutecznego sposobu transformacji. Po drugie, i co ważniejsze, rozszerzona teza stwierdza, że kompilator jest krótki, jeśli oba systemy formalne są naturalne. Powyższe tezy nie mogą być udowodnione jako prawdziwe lub fałszywe, ponieważ ludzkie, rozsądne, intuicyjne i naturalne nie zostały zdefiniowane rygorystycznie. Można zdefiniować intuicyjnie obliczalny jako obliczalny Turinga, a naturalny równoważny system Turinga jako

taki, który ma mały (powiedzmy < 10^ bitów) interpreter/kompilator na raz na zawsze ustalonej stałej referencji uniwersalnej maszynie Turinga. Tezy byłyby wtedy takie, że te definicje są rozsądne. Z przyczyn technicznych potrzebujemy następujących wariantów maszyny Turinga.

Definicja 2.6 (Prefiksowa/monotoniczna maszyna Turinga). Prefiksowa/monotoniczna maszyna Turinga jest zdefiniowana jako maszyna Turinga z jedną jednokierunkową taśmą wejściową, jedną jednokierunkową taśmą wyjściową i kilkoma dwukierunkowymi taśmami roboczymi. Taśmy wejściowe są tylko do odczytu, taśmy wyjściowe są tylko do zapisu, taśmy jednokierunkowe to takie, w których głowica może się poruszać tylko z lewej do prawej. Wszystkie taśmy są binarne (bez pustego symbolu), taśmy robocze początkowo wypełnione zerami

Prefiks TM. Mówimy, że T zatrzymuje się na wejściu p bez wyjścia x i piszemy T(p) = x, jeśli p znajduje się po lewej stronie głowicy wejściowej, a x po lewej stronie głowicy wyjściowej po zatrzymaniu T. Zbiór p, na którym zatrzymuje się T, tworzy kod prefiksowy. Taki kod p nazywamy programami samoograniczającymi.

Monotone TM. Mówimy, że T wyprowadza/oblicza ciąg zaczynający się od x (lub sekwencję ω) na wejściu p i piszemy T(p) = x* (lub T(p) = ω, jeśli p znajduje się po lewej stronie głowicy wejściowej, gdy wyprowadzany jest ostatni bit x (T odczytuje całe p, ale nie więcej). T może kontynuować działanie i nie musi się zatrzymywać. Dla danego x zbiór takich p tworzy kod prefiksowy. Takie kody nazywamy p programami minimalnymi.

Tablicę reguł maszyny Turinga T można zakodować w sposób kanoniczny jako ciąg binarny, który oznaczamy przez <T>. Stąd zbiór maszyn Turinga $\{T_1,T_2,...\}$ można skutecznie wyliczyć. Istnieją tak zwane uniwersalne maszyny Turinga, które mogą „symulować" wszystkie inne maszyny Turinga. Poniżej zdefiniujemy jedną szczególną, która pozwala również na informacje poboczne y.

Twierdzenie 2.7 (Uniwersalna maszyna Turinga prefiksowa/monotoniczna)

Istnieje uniwersalna maszyna Turinga prefiksowa/monotoniczna U, która symuluje maszynę Turinga prefiksową/monotoniczną T_i z wejściem y'q, jeśli jest zasilana wejściem y'i'q, tj.

$$U(y'i'q) = T_i(y'q) \forall i, q.$$

Nazywamy to konkretne U uniwersalną maszyną Turinga. Zauważ, że dla p niebędącego w formie y'i'q, U(p) nie zatrzymuje się. W przypadku braku informacji pobocznych y = ε pomijamy w poniższym kodzie początkowe y' = ε' = 0. Pomijamy również dodatek „prefiks/monotoniczny", jeśli wynika to z kontekstu, i identyfikujemy obiekty za pomocą ich kodowania <>, tj. pomijamy <>. Ceną, jaką musimy zapłacić za istnienie uniwersalnej maszyny Turinga, jest nierozstrzygalność problemu zatrzymania : Nie istnieje TM T z ∀i,p[T(i'p) = 1 ⇔ T_i(p) nie zatrzymuje się]. Załóżmy, że taka TM istnieje, wówczas R(i):=T{i'i) jest obliczalne, stąd ∃j:Tj ≡ R, stąd R(j)=T(j'j) = 1 ⇔ T_j{j) = R(j) nie zatrzymuje się, co jest sprzecznością.

Złożoność Kołoogomorowa

Aby wykorzystać brzytwę Ockhama poza intuicją, musimy sformalizować koncepcję prostoty i/lub złożoności. Najpierw omówimy przypadek zerowej wiedzy tła y=ε. Intuicyjnie ciąg jest prosty, jeśli można go opisać kilkoma słowami, takimi jak „ciąg miliona jedynek", i jest złożony, jeśli nie ma takiego krótkiego opisu, takiego jak dla losowego ciągu, którego najkrótszym opisem jest określenie go bit po bicie. Interesują nas tylko opisy lub kody, które są skuteczne, a zatem ograniczają dekodery do maszyn Turinga. Mówimy, że (program) p jest opisem ciągu x względem prefiksu maszyny Turinga T, jeśli T(p) = x. Długość najkrótszego opisu jest oznaczana jako $K_T(x) := \min_p \{\ell(p) : T(p) = x\}$. Ta miara złożoności

zależy od T i można zapytać, czy istnieje maszyna Turinga, która prowadzi do najkrótszych kodów spośród wszystkich maszyn Turinga dla wszystkich x. Co ciekawe, istnieje maszyna Turinga (uniwersalna), która „prawie" ma tę własność. Jeśli p jest najkrótszym opisem x przy $T = T_i$, to i'p jest opisem x przy U, stąd

$$K_U(x) \leq K_T(x) + c_{TU} \qquad (2.8)$$

z c_{TU}=^ℓ(i') i podobnie dla innych wyborów uniwersalnych maszyn Turinga. Długość najkrótszego opisu x dla U jest co najwyżej stałą liczbą bitów dłuższą niż najkrótszy opis dla T. Stwierdzenie i dowód tego twierdzenia o niezmienniczości jest często uważane za narodziny algorytmicznej teorii informacji. Co więcej, dla każdej pary uniwersalnych maszyn Turinga U' i U'' spełniających twierdzenie o niezmienniczości złożoności pokrywają się aż do stałej addytywnej $(|K_{U'}(x) - K_{U''}(x)| \leq c_{U'U''})$. .Ponieważ $c_{U'U''}$ jest zasadniczo stałą kompilatora/interpretera, przypominamy sobie Założenie 2.5 i interpretujemy to założenie jako $c_{U'U''}$ będące małym dla naturalnych uniwersalnych maszyn Turinga U' i U''. Odtąd zapisujemy O(1) dla terminów takich jak $c_{U'U''}$, które zależą tylko od wyboru uniwersalnych maszyn Turinga, ale są niezależne od rozważanych ciągów. Rozszerzamy definicję złożoności, aby uwzględnić informacje poboczne y.

Definicja 2.9 (Złożoność Kołomogorowa). Niech U będzie uniwersalną maszyną Turinga prefiksową odniesienia U Twierdzenia 2.7. (Warunkowa) złożoność prefiksowa Kolmogorowa jest zdefiniowana jako najkrótszy program p, dla którego U daje x (przy danym y).

$$K(x) := \min_p\{\ell(p) : U(p) = x\}, \quad K(x|y) := \min_p\{\ell(p) : U(y'p) = x\}$$

Dla ogólnych (nie-stringowych) obiektów (jak funkcje obliczalne) można określić pewne domyślne kodowanie i zdefiniować K(object) :=K((object)), szczególnie dla liczb i par, np. skracamy K{x,y) := K(<x,y>) = K{x'y). Poniżej wymieniono najważniejsze własności informacyjno-teoretyczne K.

Twierdzenie 2.10 (Właściwości złożoności Kolmogorowa)

$$\begin{aligned}
&(i)\ K(x) \stackrel{+}{\leq} \ell(x)+2\log_2\ell(x), \qquad K(n) \stackrel{+}{\leq} \log_2 n+2\log_2\log n\\
&(ii)\ \textstyle\sum_x 2^{-K(x)} \leq 1, \quad K(x) \geq \ell(x) \text{ for 'most' } x, \quad K(n)\to\infty \text{ for } n\to\infty\\
&(iii)\ K(x|y) \stackrel{+}{\leq} K(x) \stackrel{+}{\leq} K(x,y)\\
&(iv)\ K(xy) \stackrel{+}{\leq} K(x,y) \stackrel{+}{\leq} K(x)+K(y|x) \stackrel{+}{\leq} K(x)+K(y)\\
&(v)\ K(x|y,K(y))+K(y) \stackrel{+}{=} K(x,y) \stackrel{+}{=} K(y,x) \stackrel{+}{=} K(y|x,K(x))+K(x)\\
&(vi)\ K(f(x)) \stackrel{+}{\leq} K(x)+K(f) \text{ for recursive } f:\mathbb{B}^*\to\mathbb{B}^*\\
&(vii)\ K(x) \stackrel{+}{\leq} -\log_2 P(x)+K(P) \text{ if } P:\mathbb{B}^*\to[0,1] \text{ is enum. and } \textstyle\sum_x P(x)\leq 1
\end{aligned}$$

Wszystkie (nie)równości pozostają ważne, jeśli K jest (dalej) warunkowane przy pewnym z, tj. $K(...) \rightsquigarrow K(...|z)$ and $K(...|y) \rightsquigarrow K(...|y,z)$. Wszystkie podane są ważne w ramach stałej addytywnej o rozmiarze O(1), ale są inne, które są ważne tylko do dokładności logarytmicznej. K ma wiele wspólnych właściwości z entropią Shannona, jak być powinno, ponieważ obie mierzą zawartość informacyjną ciągu. Własność (i) podaje górną granicę K, a własność (ii) jest nierównością Krafta, która implikuje dolną granicę K ważną dla „większości" n, gdzie „większość" oznacza, że istnieją tylko o{N) wyjątków dla $n \in \{1,...,N\}$. Podanie informacji pobocznej y nigdy nie może zwiększyć długości kodu, a wymaganie dodatkowych informacji y nigdy nie może zmniejszyć długości kodu (iii). Kodowanie x i y

oddzielnie nigdy nie pomaga (iv), a transformacja x nie zwiększa jego zawartości informacyjnej (vi). Własność (vi) pokazuje również, że jeśli x koduje jakiś obiekt o, przełączając się z jednego schematu kodowania do innego za pomocą rekurencyjnej bijekcji f pozostawia K niezmienione w obrębie addytywnych terminów O(1). Pierwszym nietrywialnym wynikiem jest symetria informacji (v), która jest odpowiednikiem reguły łańcuchowej. Własność (vii) leży u podstaw zasady MDL, która aproksymuje K{x) przez $-\log_2 P(x) + K(P)$. Wszystkie górne ograniczenia na K(z) można łatwo udowodnić, opracowując pewien (efektywny) kod dla z o długości prawej strony nierówności i zauważając, że K{z) jest długością najkrótszego kodu spośród wszystkich możliwych efektywnych kodów. Na przykład, jeśli T_{i0} z i_0 = O(1) jest maszyną Turinga z $T_{i0}(\varepsilon' x') = x$, wtedy $U(\epsilon' i_0' x') = x$; stąd $K(x) \leq \ell(\epsilon' i_0' x') \stackrel{+}{=} \ell(x') \stackrel{+}{\leq} \ell(x) + 2\log_2 \ell(x)$, co dowodzi (i). W {vii) używa się kodu Shannona-Fano opartego na rozkładzie prawdopodobieństwa P. Dolne granice są zwykle dowodzone przez zliczanie argumentów (łatwe dla (ii) przy użyciu (2.1) i trudniejsze dla (v)).

Koncepcje obliczalności

Potrzebujemy kilku koncepcji obliczalności słabszych niż te, które można uchwycić, zatrzymując maszyny Turinga.

Definicja 2.12 (Funkcje obliczalne). Rozważamy funkcje f : N -> R:

f jest skończenie obliczalna lub rekurencyjna wtedy i tylko wtedy, gdy istnieje maszyna Turinga T z T(x') = n'd i n/d = f(x)

f jest aproksymowalna wtedy i tylko wtedy, gdy istnieje maszyna Turinga skończenie obliczająca $\Phi(.,.)$ taka, że $\lim_{t\to\infty} \phi(x,t) = f(x)$

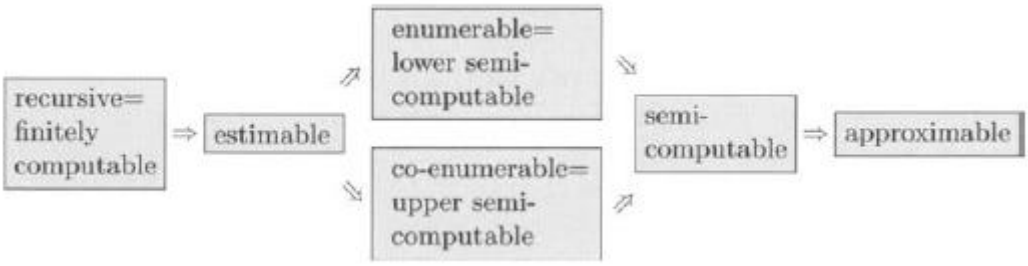
f jest dolną półobliczalną lub wyliczalną liczbą, jeśli dodatkowo $\phi(x,t) \leq \phi(x,t+1)$

f jest półobliczalna górna lub współprzeliczalna, jeśli [-f] jest półobliczalna dolna

f jest półobliczalna, jeśli f jest półobliczalna dolna lub górna

f jest szacowalna, jeśli f jest półobliczalna dolna i górna

Jeśli f jest szacowalne, możemy skończenie obliczyć ε-przybliżenie z f przez górne i dolne półobliczanie f i kończyć, gdy różni się o mniej niż ε. Oznacza to, że istnieje maszyna Turinga, która, biorąc pod uwagę x i ε, skończenie oblicza $\hat{y}$ takie, że $|\hat{y} - f(x)| < \varepsilon$. Co więcej, daje ona oszacowanie przedziałowe $f(x) \in [\hat{y} - \varepsilon, \hat{y} + \varepsilon]$. Oszacowalna funkcja o wartościach całkowitych jest skończenie obliczalna (przyjmij dowolne ε<1). Należy zauważyć, że jeśli f jest tylko przybliżane lub półobliczalne, nadal możemy zbliżyć się dowolnie do f(x), ale nie możemy opracować algorytmu kończącego, który generuje przybliżenie ε. W przypadku dolnej/górnej półobliczalności możemy przynajmniej skończenie obliczyć dolne/górne granice do f(x). W przypadku przybliżalności, najsłabszej formy obliczalności, nawet ta zdolność jest tracona. Analogicznie do niższej/wyższej zdolności półobliczeniowej, można pomyśleć o pojęciach takich jak niższa/wyższa estymowalność, ale łatwo wykazać, że pokrywają się one z estymowalnością. Następujące implikacje są ważne:



Główną właściwością algorytmiczną K jest:

Twierdzenie 2.13 ((Nie)obliczalność złożoności Kolmogorowa). Złożoność Kolmogorowa K:B^*->IN jest współprzeliczalna, ale nie skończenie obliczalna

Współwyliczalność K jest oczywista z definicji K. Nieobliczalność wynika z argumentu diagonalizacji: Załóżmy, że K jest obliczalna. Wówczas f(m) := min{n : K{n) ≥ m} istnieje na mocy twierdzenia 2.10 (ii) i jest obliczalna, K(f{m)) ≥ m na mocy definicji f oraz $K(f(m)) \stackrel{+}{\leq} K(m)+K(f) \stackrel{+}{\leq} 2\log_2 m$ na mocy twierdzenia 2.10(i,vi). Stąd m ≤ $\log_2$m + c dla pewnego c, ale jest to fałszem dla dostatecznie dużego m. W dalszej części używamy terminu obliczalny jako synonimu skończenie obliczalnego, ale czasami także generycznie dla niektórych form obliczalności z definicji 2.12. To, co nazywamy szacowalnym, jest często nazywane po prostu obliczalnym, ale ma sens oddzielenie tutaj pojęć skończonej obliczalności i szacowalności, ponieważ pierwsze jest koncepcyjnie łatwiejsze.

Niepewność i: Prawdopodobieństwa

Celem teorii prawdopodobieństwa jest opisanie niepewności. Istnieją różne źródła niepewności, a zatem różne interpretacje prawdopodobieństw. Istnieją co najmniej trzy „szkoły”:

• prawdopodobieństwa częstotliwościowe to względne częstotliwości. (np. względna częstotliwość rzucania głową)

• obiektywistyczne: prawdopodobieństwa to rzeczywiste aspekty świata. (np. prawdopodobieństwo, że jakiś atom rozpadnie się w ciągu następnej godziny)

• subiektywistyczne: prawdopodobieństwa opisują stopień wiary agenta w coś (np. jest (nie)prawdopodobne, że istnieją istoty pozaziemskie)

Poniższe podsekcje opisują te interpretacje i podejścia iscuss do uzyskiwania prawdopodobieństw a priori.

Uwaga. W niektórych społecznościach dziedzina stosowalności oraz prawidłowa interpretacja i forma teorii prawdopodobieństwa są nadal kontrowersyjne. Dla tych czytelników chcemy podkreślić, że prawdopodobieństwa można całkowicie pominąć bez trywializacji jej celów i wyników. Terminologia prawdopodobieństw subiektywnych jest używana w tej książce wyłącznie w celach motywacyjnych i ilustracyjnych. Nie opieramy się na uzasadnieniu Coxa , ale podajemy uzasadnienia teoretyczno-decyzyjne. Nawet pojęcie prawdopodobieństw obiektywnych może zostać porzucone przez założenie deterministycznych środowisk. Niektóre wyniki w książce upraszczają się w tym przypadku, ale zachowują swoje znaczenie. Czytelnicy niewierzący w prawdopodobieństwa obiektywne i/lub subiektywne nadal mogą uznać tę książkę za interesującą.

Interpretacja częstotliwości: zliczanie

Frekwencjonista interpretuje prawdopodobieństwa jako częstości względne. Jeśli w ciągu n niezależnych, identycznie rozłożonych (i.i.d.) eksperymentów (prób) zdarzenie występuje k(n) razy, częstość względna zdarzenia wynosi k(n)/n. Granica $\lim_{n\to\infty} k(n)/n$ jest zdefiniowana jako prawdopodobieństwo zdarzenia. Była to najwcześniejsza matematyczna definicja prawdopodobieństw autorstwa Bernoulliego, opublikowana w 1713 r. Na przykład prawdopodobieństwo, że zdarzenie wypadnie orłem w ciągu wielokrotnego rzucania uczciwą monetą, wynosi 1/2. Stanowisko frekwencyjne jest najłatwiejsze do zrozumienia, ale ma kilka wad:

• Frekwencjonista uzyskuje prawdopodobieństwa z procesów fizycznych, jak opisano powyżej. Aby naukowo rozumować o prawdopodobieństwach, potrzebna jest teoria matematyczna. Problem polega na tym, jak zdefiniować sekwencje losowe. Jest to o wiele bardziej skomplikowane niż można by sądzić i zostało rozwiązane dopiero w latach 60-tych przez Kolmogorova i Martina-Lofa. Naiwna definicja prawdopodobieństwa jest kołowa: Prawdopodobieństwo zdarzenia E wynosi $p := \lim_{n\to\infty} \frac{k_n(E)}{n}$ gdzie $k_n(E)$ jest liczbą wystąpień zdarzenia E w pierwszych n niezależnych od siebie próbach. Problem polega na tym, że granica może być dowolna lub nawet nie istnieć: np. uczciwa moneta może dać: orzeł, orzeł, orzeł, orzeł, ... tj. p=1. Oczywiście ta sekwencja jest „mało prawdopodobna". W przypadku uczciwej monety p=1/2 z „wysokim prawdopodobieństwem". Ale aby uczynić to stwierdzenie ścisłym, musimy formalnie zdefiniować, co oznacza „wysokie prawdopodobieństwo". Oto kołowość!

• Filozoficznie, a także często w rzeczywistych eksperymentach, trudno jest uzasadnić wybór tak zwanej klasy odniesienia. Na przykład lekarz chce określić prawdopodobieństwo, że pacjent ma określoną chorobę, licząc częstość występowania choroby u „podobnych" pacjentów. Ale gdyby lekarz wziął pod uwagę wszystko, co wie o pacjencie (objawy, wagę, wiek, pochodzenie, ...), nie byłoby już innych porównywalnych pacjentów.

• Podejście częstotliwościowe jest ograniczone do (wystarczająco dużej) próbki danych i.i.d.

Interpretacja obiektywna

Prawdopodobieństwa opisujące niepewne zdarzenia

Dla obiektywisty prawdopodobieństwa są rzeczywistymi aspektami świata. Wynik obserwacji lub eksperymentu nie jest deterministyczny, ale obejmuje fizyczne procesy losowe. Zbiór Ω wszystkich możliwych wyników nazywa się przestrzenią próby. Mówi się, że zdarzenie $E \subset \Omega$ wystąpiło, jeśli wynik jest w E. W przypadku eksperymentów i.i.d. prawdopodobieństwa przypisane zdarzeniom powinny być interpretowane jako częstości graniczne, ale zastosowanie nie ogranicza się do tego przypadku. Aksjomaty Kołmogorowa formalizują właściwości, jakie powinny mieć prawdopodobieństwa.

Aksjomaty 2.14 (aksjomaty teorii prawdopodobieństwa Kołmogorowa) Niech Ω będzie przestrzenią próby. Zdarzenia są podzbiorami Ω.

* Jeśli A i B są zdarzeniami , to przecięcie $A \cap B$, suma $A \cup B$ i różnica $A \backslash B$ są również zdarzeniami.

* Przestrzeń próby Ω i zbiór pusty {} są zdarzeniami .

* Istnieje funkcja p, która przypisuje nieujemne liczby rzeczywiste , zwane prawdopodobieństwami , do każdego zdarzenia.

* $p(\Omega) = 1$, $p\{\} = 0$

* $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

* Dla malejącego ciągu $A_1 \supset A_2 \supset A_3 \ldots$ ... zdarzeń z $\bigcap_n A_n = \{\},$ mamy $\lim_{n\to\infty} p(A_n) = 0.$

Funkcja p jest nazywana funkcją masy prawdopodobieństwa lub miarą prawdopodobieństwa, lub, luźniej, rozkładem prawdopodobieństwa. Prawdopodobieństwa warunkowe są definiowane w następujący sposób:

Definicja 2.15 (Prawdopodobieństwo warunkowe). Jeżeli A i B to zdarzenia z p(A) > 0, to prawdopodobieństwo, że wystąpi także B pod warunkiem, że wystąpiło także A, jest zdefiniowane jako

$$p(B|A) := \frac{p(A \cap B)}{p(A)}$$

Łatwo zauważyć, że p(·|A) (jako funkcja pierwszego argumentu) jest również miarą prawdopodobieństwa, jeśli p(·) spełnia aksjomaty Kołmogorowa. Można „zweryfikować poprawność" aksjomatów Kołmogorowa i definicję prawdopodobieństw warunkowych w przypadku, gdy prawdopodobieństwa są identyfikowane z częstościami granicznymi. Ale chodzi o to, aby przyjąć aksjomaty jako punkt wyjścia, aby uniknąć problemów frekwencyjnych. Relacja $p(A \cap B) = p(B|A) \cdot p(A)$ nazywana jest regułą mnożenia (prawdopodobieństw warunkowych), która jest szczególnym przypadkiem reguły łańcuchowej.

Twierdzenie 2.16 (reguła Bayesa 1). Jeżeli A i B to zdarzenia z p(A) > 0 i p(B) > 0, to

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Twierdzenie Bayesa można łatwo udowodnić, stosując definicję 2.15 dwukrotnie.

Subiektywna interpretacja : Prawdopodobieństwa opisujące stopnie wiary

Subiektywista używa prawdopodobieństw do scharakteryzowania stopnia wiary agenta w coś, a nie do scharakteryzowania fizycznych procesów losowych. Jest to najbardziej odpowiednia interpretacja prawdopodobieństw w AI. Definiujemy prawdopodobieństwo zdarzenia jako stopień wiary w zdarzenie lub subiektywne prawdopodobieństwo zdarzenia. Problem z subiektywnym poglądem polega na tym, że o wiele bardziej dyskusyjne jest, jak zdefiniować prawdopodobieństwo, w porównaniu z prawdopodobieństwem obiektywnym. Obiektywista może motywować aksjomaty Kolmogorowa za pomocą analizy częstotliwości, ale nie ma interpretacji częstotliwości dla prawdopodobieństw. Jeśli agent wierzy w kosmitów i przypisuje prawdopodobieństwo 0,9 ich istnieniu, nie ma większego sensu interpretować tego jako „w 90 na 100 równoległych wszechświatów istnieją kosmici" lub „90 na 100 podobnych agentów wierzy w kosmitów". Ten problem doprowadził do powstania wielu różnych systemów zajmujących się niepewnym rozumowaniem . Wszystkie mają swoje własne problemy. Najbardziej spójny i skuteczny system opiera się, ponownie, na aksjomatach Kołmogorowa, chociaż nie wszyscy zgodziliby się z tym stwierdzeniem. Zaskakujące jest, że prawdopodobieństwa podlegają tym samym regułom, co częstości graniczne. Aksjomaty Kołmogorowa można wyprowadzić z kilku prawdopodobnych jakościowych reguł, których powinny przestrzegać. Naturalne jest założenie, że prawdopodobieństwa można przedstawić za pomocą liczb rzeczywistych, że reguły jakościowo odpowiadają zdrowemu rozsądkowi i że reguły są matematycznie spójne. Cox] zaczyna od następujących (naturalnych) założeń dotyczących przekonań:

Aksjomaty 2.17 (aksjomaty Coxa dla przekonań)

* Stopień przekonania o zdarzeniu B (prawdopodobieństwie zdarzenia B), zakładając, że zdarzenie A miało miejsce, można scharakteryzować za pomocą funkcji o wartościach rzeczywistych Bel(B|A)

* Bel(Ω\B|A) jest dwukrotnie różniczkowalną funkcją Bel(B|A) dla A ≠ {}

* Bel(B∩C|A) jest dwukrotnie ciągle różniczkowalną funkcją Bel(C|B∩A) i Bel (B|A) dla B∩A ≠ {}

Cox pokazuje, że każda funkcja Bel(·|·) spełniająca te aksjomaty jest izomorficzna z (warunkową) funkcją prawdopodobieństwa. Można uzasadnić związek funkcyjny w aksjomatach Coxa, analizując wszystkie inne możliwości i pokazując, że naruszają one zdrowy rozsądek. Nieco silne założenia różniczkowalności można osłabić do bardziej naturalnych założeń ciągłości i monotoniczności. Dopiero niedawno wykazano lukę w wyprowadzeniach Coxa. Zasugerowano kilka poprawek poprzez wprowadzenie dodatkowych założeń. Większość z nich wymaga, aby zakres Bela, a zatem zbiór zdarzeń, był wystarczająco bogaty. Parafrazujemy je jako „dodatkowe warunki gęstości".

Twierdzenie 2.18 (twierdzenie Coxa). Zgodnie z aksjomatami 2.17 i pewnymi dodatkowymi warunkami gęstości. Bel (·|A) jest izomorficzny z funkcją prawdopodobieństwa w tym sensie, że istnieje ciągła funkcja różnowartościowa g:R->[0,1] taka, że p:= g ∘ Bel spełnia aksjomaty Kołmogorowa 2.14 i jest zgodny z definicją 2.15.

Wynik Coxa wzbudził duże zainteresowanie, szczególnie w społeczności zajmującej się maksymalną entropią i sztuczną inteligencją. Jakościowa motywacja aksjomatów Coxa i wyprowadzenie z nich twierdzenia Coxa jest głównym teoretycznym uzasadnieniem, że subiektywne „stopnie przekonań" powinny spełniać te same aksjomaty Kołmogorowa, co częstotliwości graniczne. Inne podejścia do przekonań nie mają tej silnej podstawy teoretycznej i spójności. Wykorzystując fakt, że subiektywne prawdopodobieństwa podlegają tym samym regułom, co obiektywne prawdopodobieństwa, możemy przedstawić regułę Bayesa w szczególnie przydatnej formie: jak aktualizować przekonania w obliczu nowych obserwacji.

Twierdzenie 2.19 (reguła Bayesa 2). Niech D będzie pewnymi możliwymi danymi (tj. D jest zdarzeniem z p(D) > 0), a $\{H_i\}_{i\in I}$ będzie przeliczalną zupełną klasą wzajemnie wykluczających się hipotez (tj. H_i są zdarzeniami z $H_i \cap H_j = \{\}\ \forall i \neq j$ a $\bigcup_{i\in I} H_i = \Omega$).

Założono: $p(H_i)$ = a priori prawdopodobieństwo hipotez H_i (subj. prob)

Założono: $p(D|H_i)$ = prawdopodobieństwo danych D przy hipotezie H_i (subj. prob)

Cel: p(Hi|D) = a posteriori prawdopodobieństwo hipotezy H_i (subj. prob)

Rozwiązanie:

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{\sum_{i\in I} p(D|H_i)p(H_i)}$$

Dowód. Dowód opiera się na wszystkich i tylko na Aksjomatach 2.14. p{AuB}=p{A}+p{B} jeśli A n 5 = {}, ponieważ p{{}) — 0. Dla skończonego I, przez indukcję, implikuje to $\sum_{i\in I} p(H_i) = p(\bigcup_i H_i) = p(\Omega) = 1$. - Dla przeliczalnie nieskończonego I = {1,2,3,...} z $S_n := \bigcup_{i=n}^{\infty} H_i$ mamy:

$$\sum_{i=1}^{n-1} p(H_i) + p(S_n) \;=\; p\Big(\bigcup_{i=1}^{n-1} H_i \cup S_n\Big) = p(\Omega) = 1. \qquad (2.20)$$

Wykorzystując $S_1 \supset S_2 \supset S_3 \cdots,$ dla dowolnego $\omega \in \Omega$ mamy: $\exists n : \omega \in H_n \Rightarrow \omega \notin H_i \forall i > n$ u) $\Rightarrow \omega \notin S_i \forall i > n \Rightarrow \omega \notin \bigcap_n S_n \Rightarrow \bigcap_n S_n = \{\}$ ponieważ ω było dowolne => biorąc n -> ∞ w (2.20) pokazuje $\sum_{i=1}^{\infty} p(H_i) = 1$. Ponieważ prawdopodobieństwa warunkowe spełniają również Aksjomaty 2.14, mamy również $\sum_{i \in I} p(H_i|D) = 1$ (zarówno dla skończonych, jak i nieskończonych I). Z Definicji 2.15 prawdopodobieństwa warunkowego mamy

$$p(H_i|D)p(D) = p(H_i \cap D) = p(D|H_i)p(H_i).$$

Podsumowując wszystkie hipotezy, H_i daje

$$\sum_{i \in I} p(D|H_i)p(H_i) = \sum_{i \in I} p(H_i|D) \cdot p(D) = 1 \cdot p(D)$$

$$\Rightarrow \quad p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D)} = \frac{p(D|H_i)p(H_i)}{\sum_{i \in I} p(D|H_i)p(H_i)}$$

Określanie prawdopodobieństw a priori

Aksjomaty Kolmogorova 2.14 prawdopodobieństwa pozwalają nam powiązać prawdopodobieństwa i prawdopodobieństwa różnych zdarzeń, ale nie ustalają jednoznacznie wartości liczbowej dla każdego zdarzenia, z wyjątkiem pewnego zdarzenia Ω i pustego zdarzenia {}. Potrzebujemy nowych zasad określania wartości dla co najmniej niektórych zdarzeń bazowych, z których inne można obliczyć za pomocą tych aksjomatów. Wydaje się, że istnieją tylko trzy ogólne zasady:

• zasada obojętności — zasada symetrii,

• zasada maksymalnej entropii,

• brzytwa Ockhama — zasada prostoty.

Podczas gdy pierwsze dwie zasady opierają się na podstawach fizyki statystycznej, zobaczymy, że tylko brzytwa Ockhama (zachowaj tylko najprostszą spójną hipotezę) w połączeniu z zasadą Epikura dotyczącą wielokrotnych wyjaśnień (zachowaj wszystkie spójne hipotezy) jest wystarczająco ogólna, aby przypisać prawdopodobieństwa a priori w każdej sytuacji, szczególnie w przypadku indukcji i innych domen typowych dla AL. Pomysł polega na przypisaniu wysokiego (subiektywnego) prawdopodobieństwa zdarzeniom prostym, a niskiego prawdopodobieństwa zdarzeniom złożonym: Zdarzenia proste (ciągi) są bardziej prawdopodobne a priori niż zdarzenia złożone. Daje to (przybliżoną) sprawiedliwość zarówno brzytwie Ockhama, jak i zasadzie Epikura. W dalszej części odnosimy się również do tej ogólnej idei jako do brzytwy Ockhama. Użycie K do pomiaru prostoty/złożoności prowadzi do uniwersalnego a priori M Solomonoffa. W następnej sekcji kontynuujemy to podejście.

Algorytmiczne prawdopodobieństwo i indukcja uniwersalna

Oprócz notacji wprowadzonej w rozdziale 2.2.1, ciągi binarne o długości n oznaczamy jako $x=x_1x_2...x_n$ z $x_t \in IB$, a następnie skracamy $x_{1:n} := x_1x_2...x_{n-1}x_n$ i $x_{<n} : x_1 \ldots x_{n-1}$.

Uniwersalny Prior M

Złożoność prefiksowa Kolmogorowa K(x) została zdefiniowana jako najkrótszy program p, dla którego uniwersalna maszyna prefiksowa Turinga U wyprowadza ciąg x, i podobnie K(x|y) w przypadku informacji pobocznych y (Definicja 2.9). Solomonoff zdefiniował ściśle powiązaną wielkość, uniwersalny prior M{x). Uniwersalny prior jest zdefiniowany jako prawdopodobieństwo, że wyjście uniwersalnej monotonicznej maszyny Turinga zaczyna się od x, gdy na taśmie wejściowej znajdują się uczciwe krawędzie monety. Formalnie M można zdefiniować jako

$$M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}, \qquad (2.21)$$

gdzie suma przekracza minimalne programy p, dla których U generuje ciąg zaczynający się od x (patrz Definicja 2.6). Ponieważ najkrótsze programy p dominują nad sumą, M(x) wynosi w przybliżeniu $2^{-K(x)}$ $(M(x)=2^{-K(x)+O(K(\ell(x))})$. Zanim będziemy mogli omówić stochastyczne własności M, potrzebujemy pojęcia (pół)miar dla ciągów.

Definicja 2.22 ((Pół)miary). μ(x) oznacza prawdopodobieństwo, że ciąg binarny zaczyna się od ciągu x. Nazywamy μ ≥ 0 półmiarą, jeśli μ(ε) ≤1 i μ(x) ≥ μ(x0) + μ(x1), a miarą prawdopodobieństwa, jeśli zachodzą równości

Powodem, dla którego μ z powyższą własnością nazywamy miarą prawdopodobieństwa, jest to, że spełnia ona Aksjomaty 2.14 prawdopodobieństwa Kolmogorowa w następującym sensie: Przestrzeń próby to IB^{∞} z elementami ω = ω1ω2ω3… ∈ IB^{∞} będącymi nieskończonymi ciągami binarnymi. Zbiór zdarzeń (σ-algebra) jest zdefiniowany jako zbiór wygenerowany ze zbiorów cylindrycznych $\Gamma_{x1:n}$:={ω:$\omega_{1:n}$ = $x_{1:n}$} przez przeliczalną unię i dopełnienie. Miarę prawdopodobieństwa μ definiuje się jednoznacznie, nadając jej wartości $\mu(\Gamma_{x_{1:n}})$ zbiorom cylindrycznym, które w skrócie nazywamy μ{$x_{1:n}$). Będziemy również nazywać μ miarą lub jeszcze luźniej rozkładem prawdopodobieństwa. Powodem rozszerzenia definicji na półmiary jest to, że samo M niestety nie jest miarą prawdopodobieństwa. Mamy M(x0)+-M(x1) < M(x), ponieważ istnieją programy p, które wyprowadzają x, po którym nie następuje ani 0, ani 1. Po prostu zatrzymują się po wyprowadzeniu x lub kontynuują w nieskończoność bez dalszego wyprowadzania. W Problemie 2.7 defekt jest skwantyfikowany. Ponieważ M(ε) = 1, M jest co najmniej półmiarą. Teraz możemy podać podstawową własność M.

Twierdzenie 2.23. (Uniwersalność M). Uniwersalne wcześniejsze $M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}$ jest przeliczalną półmiarą, która mnoży wszystkie przeliczalne półmiary w tym sensie, że $M(x) \overset{\times}{\geq} 2^{-K(\rho)} \cdot \rho(x)$, jeśli ρ jest przeliczalną półmiarą. M jest przeliczalna, ale nie jest szacowana ani skończenie obliczalna.

Złożoność Kołmogorowa funkcji takiej jak p jest zdefiniowana jako długość najkrótszego samoograniczającego się kodu maszyny Turinga obliczającej tę funkcję w sensie definicji 2.12. Do stałej mnożnikowej M przypisuje wyższe prawdopodobieństwo wszystkim x niż jakikolwiek inny obliczalny rozkład prawdopodobieństwa. Możliwe jest znormalizowanie M do prawdziwej miary prawdopodobieństwa M_{norm} (2.30) z nadal prawdziwą dominacją, ale kosztem rezygnacji z wyliczalności (M_{norm} jest nadal aproksymowalny). Zobaczymy, że M jest wygodniejsza przy badaniu zagadnień algorytmicznych, ale prawdziwa miara prawdopodobieństwa taka jak M_{norm} JEST bardziej wygodna przy badaniu zagadnień stochastycznych.

Przewidywanie sekwencji uniwersalnej

W jakim sensie M uwzględnia brzytwę Ockhama i zasadę Epikura dotyczącą wielu wyjaśnień? Z $M(x) \approx 2^{-K(x)}$ widzimy, że M przypisuje wysokie prawdopodobieństwo prostym ciągom znaków (Ockham). Bardziej użyteczne jest myślenie o x jako o obserwowanej historii. Z Definicji (2.21) widzimy, że każdy program p zgodny z historią x może przyczyniać się do M (Epikur). Z drugiej strony, krótsze programy dają znacznie większy wkład (Ockham). Jak to wszystko wpływa na przewidywanie? Jeśli M{x) poprawnie opisuje nasze (subiektywne) wcześniejsze przekonanie o x, to M(y|x) :=M{xy)/M(x) musi być naszym późniejszym przekonaniem o y. Z symetrii informacji algorytmicznej $K(x,y) \stackrel{+}{=} K(y|x,K\{x\})) + K(x)$ (Twierdzenie 2.10(v)), i zakładając K(x,y) ≈ K{xy), i przybliżając K{y|x,K{x)) ≈ K(y|x), $M(x) \approx 2^{-K(x)}$ i $M(xy) \approx 2^{-K(xy)}$ otrzymujemy $M(y|x) \approx 2^{-K(y|x)}$. To mówi nam, że M przewiduje y z dużym prawdopodobieństwem, jeśli y ma łatwe wyjaśnienie, biorąc pod uwagę x (Ockham i Epikur). Powyższa dyskusja jakościowa nie powinna stwarzać wrażenia, że M{x) i $2^{-K(x)}$ zawsze prowadzą do predyktorów o porównywalnej jakości. Rzeczywiście, w środowisku online/przyrostowym badanym tu K(y) = O{1) unieważnia powyższe rozważania. Na przykład ważność poniższego twierdzenia 2.25 zależy od tego, czy M jest półmiarą, a reguła łańcuchowa jest dokładnie prawdziwa, żadna z nich nie jest spełniona przez $2^{-K(x)}$ (patrz Problem 2.8). Algorytmy przewidywania sekwencji (binarnych) próbują przewidzieć kontynuację $x_n \in \mathbb{B}$ danej sekwencji $x_1 \ldots x_{n-1}$. Wyprowadzamy następującą granicę:

$$\sum_{t=1}^{\infty}(1-M(x_t|x_{<t}))^2 \leq -\tfrac{1}{2}\sum_{t=1}^{\infty}\ln M(x_t|x_{<t}) = -\tfrac{1}{2}\ln M(x_{1:\infty}) \leq \tfrac{1}{2}\ln 2 \cdot Km(x_{1:\infty}) \qquad (2.24)$$

gdzie złożoność monotoniczna $Km(x_{1:\infty})$ jest zdefiniowana jako długość najkrótszego (nieprzerwanego) programu obliczającego $x_{1:\infty}$. W pierwszej nierówności użyliśmy $(1-a)^2 \leq -1/2\ln a$ dla $0 \leq a \leq 1$. W równości zamieniliśmy sumę na logarytm i wyeliminowaliśmy wynikowy iloczyn za pomocą reguły łańcuchowej. W ostatniej nierówności użyliśmy $M(x) \geq 2^{-Km(x)}$, co wynika z definicji (2.21) przez pominięcie wszystkich wyrazów w Σ_p z wyjątkiem najkrótszego p obliczającego X. Jeśli $x_{1:\infty}$ jest ciągiem obliczalnym, to $Km\{x_{1:\infty})$ jest skończony, co implikuje $M(x_t|x_{<t}) \to 1$ $(\sum_{t=1}^{\infty}(1-a_t)^2 < \infty \Rightarrow a_t \to 1)$. Oznacza to, że jeśli środowisko jest obliczalną sekwencją (którakolwiek, np. cyfry π lub e w reprezentacji binarnej), po zobaczeniu pierwszych kilku cyfr, M poprawnie przewiduje następną cyfrę z dużym prawdopodobieństwem, tj. rozpoznaje strukturę sekwencji. Załóżmy teraz, że prawdziwa sekwencja jest losowana z obliczalnego rozkładu prawdopodobieństwa μ, tj. prawdziwe (obiektywne) prawdopodobieństwo $x_{1:n}$ wynosi $\mu(x_{1:n})$. Prawdopodobieństwo x_n przy założeniu $x_{<n}$ wynosi zatem $\mu(x_n|x_{<n}) = \mu(x_{1:n})/\mu(x_{<n})$). Centralnym wynikiem Solomonoffa jest to, że M zbiega się do μ:

Twierdzenie 2.25. (Zbieżność a posteriori M do μ)

$$\sum_{t=1}^{\infty}\sum_{x_{<t}\in\mathbb{B}^{t-1}}\mu(x_{<t})\Big(M(0|x_{<t})-\mu(0|x_{<t})\Big)^2 \stackrel{+}{\leq} \tfrac{1}{2}\ln 2 \cdot K(\mu) < \infty$$

Suma nieskończona może być skończona tylko wtedy, gdy różnica $M(0|x_{<t})-\mu(0|x_{:<t})$ dąży do zera dla $t \to \infty$ z μ-prawdopodobieństwem 1 (patrz Definicja 3.8(z) i Problem 2.7). Dotyczy to dowolnego obliczalnego rozkładu prawdopodobieństwa μ. Powodem zadziwiającej właściwości pojedynczej (uniwersalnej) funkcji zbieżnej do dowolnego obliczalnego rozkładu prawdopodobieństwa jest fakt, że zbiór μ-losowych sekwencji różni się dla różnych μ. Dane z przeszłości $x_{<t}$ są wykorzystywane do

uzyskania (przy t -> ∞) poprawiającej się oceny $M(x_t|x_{<t})$ z $\mu(x_t|x_{<t})$. Własność uniwersalności (Twierdzenie 2.23) jest centralnym składnikiem dowodu Twierdzenia 2.25. Dowód Twierdzenia 2.23 obejmuje konstrukcję półmiary ξ, której dominacja jest oczywista. Trudność polega na wykazaniu jego wyliczalności i równoważności z M. Niech M będzie (przeliczalnym) zbiorem wszystkich przeliczalnych półmiar i zdefiniuj

$$\xi(x) := \sum_{\nu\in\mathcal{M}} 2^{-K(\nu)}\nu(x). \qquad (2.26)$$

Następnie dominacja

$$\xi(x) \geq 2^{-K(\nu)}\nu(x) \quad \forall\,\nu\in\mathcal{M} \qquad (2.27)$$

jest oczywiste (bez 0(1) fałszowania). Czy ξ jest dolna półobliczalna? Aby odpowiedzieć na to pytanie, musimy być bardziej precyzyjni. Levin pokazał, że istnieje maszyna Turinga T taka, że dla każdej dolnej półobliczalnej półmiary v istnieje i takie, że T(i'x') dolna półoblicza $v_1 \equiv v$ tj. T wylicza wszystkie dolne półobliczalne półmiary, być może z powtórzeniem. Dla (uporządkowanego zbioru wielowymiarowego) $M = M_U := \{v_1,v_2,v_3...\}$ i $K(v_i):=K\{i)$, można łatwo zobaczyć, że ξ jest dolną półobliczalną. Na koniec, udowodnienie $M(x)\stackrel{\times}{=}\xi(x)$ również ustanawia uniwersalność M. Zaletą ξ nad M jest to, że natychmiast uogólnia się do dowolnych ważonych sum (pół)miar w M dla dowolnego przeliczalnego M. Większość dowodów przechodzi dla generycznych M. i wag. Udowodnimy (to uogólnienie) Twierdzenia 2.25 później

Uniwersalne (pół)miary

Co jest takiego szczególnego w zbiorze wszystkich przeliczalnych półmiar M_U? Im większy wybierzemy M, tym mniej restrykcyjne będzie założenie, że M powinien zawierać prawdziwy rozkład μ, który będzie istotny w całym tekście. Dlaczego nie ograniczyć się do wciąż dość ogólnej klasy szacowanych lub skończenie obliczalnych (pół)miar? Jest oczywiste, że dla każdego przeliczalnego zbioru M, $\xi(x) := \xi_{\mathcal{M}}(x) := \sum_{\nu\in\mathcal{M}} w_\nu\nu(x)$ z $\sum_\nu w_\nu \leq 1$ i $w_\nu > 0$ dominuje nad wszystkimi $\nu \in M$. Ta dominacja jest konieczna dla pożądanej zbieżności ξ -> μ, podobnie jak w twierdzeniu 2.25. Pytanie brzmi, jakie właściwości posiada ξ. Cechą wyróżniającą M_U jest to, że $\xi=\xi_U\equiv\xi_{\mathcal{M}_U}\stackrel{\wedge}{=}M$ samo w sobie jest elementem M_U. Tu $\xi_M \in M$ samo w sobie nie jest ważną własnością, ale to, czy ξ jest obliczalne w jednym ze znaczeń Definicji 2.12.

Definiujemy

$\mathcal{M}_1 \stackrel{\times}{\geq} \mathcal{M}_2 \;:\Leftrightarrow$ istnieje element M_1, który dominuje nad wszystkimi elementami M_2

$:\Leftrightarrow \;\exists\rho\in\mathcal{M}_1\;\forall\nu\in\mathcal{M}_2\;\exists w_\nu>0\;\forall x:\rho(x)\geq w_\nu\nu(x)$

Relacja $\stackrel{\times}{\geq}$ jest przechodnia (ale niekoniecznie zwrotna) w tym sensie, że $\mathcal{M}_1\stackrel{\times}{\geq}\mathcal{M}_2\stackrel{\times}{\geq}\mathcal{M}_3$ implikuje $\mathcal{M}_1\stackrel{\times}{\geq}\mathcal{M}_3$, a $\mathcal{M}_0\supseteq\mathcal{M}_1\stackrel{\times}{\geq}\mathcal{M}_2\supseteq\mathcal{M}_3$ implikuje $\mathcal{M}_0\stackrel{\times}{\geq}\mathcal{M}_3$. Dla koncepcji obliczalności wprowadzonych w rozdziale 2.2.4 mamy następujące właściwe inkluzje zbiorów

$$\begin{array}{ccccccc} \mathcal{M}^{msr}_{comp} & \subset & \mathcal{M}^{msr}_{est} & \equiv & \mathcal{M}^{msr}_{enum} & \subset & \mathcal{M}^{msr}_{appr} \\ \cap & & \cap & & \cap & & \cap \\ \mathcal{M}^{semi}_{comp} & \subset & \mathcal{M}^{semi}_{est} & \subset & \mathcal{M}^{semi}_{enum} & \subset & \mathcal{M}^{semi}_{appr} \end{array}$$

gdzie M^{msr}_c oznacza zbiór wszystkich miar prawdopodobieństwa odpowiedniego typu obliczalności c ∈ {comp=skończenie obliczalny, est—szacowalny, enum=przeliczalny, appr=przybliżony}, i podobnie dla półmiar M^{semi}_c. W przypadku wyliczenia miary ρ można skonstruować współwyliczenie wykorzystując $\rho(x_{1:n}) = 1 - \sum_{y_{1:n} \neq x_{1:n}} \rho(y_{1:n})$ pokazuje, że każda przeliczalna miara jest również współprzeliczalna, a zatem szacowalna, co dowodzi tożsamości = powyżej. Przy takim oznaczeniu można odczytać $\mathcal{M}^{semi}_{enum} \overset{\times}{\geq} \mathcal{M}^{semi}_{enum}$. Przechodniość pozwala na przykład wnioskować, że $\mathcal{M}^{semi}_{appr} \overset{\times}{\geq} \mathcal{M}^{msr}_{comp}$, tj. istnieje przybliżona półmiara, która dominuje nad wszystkimi obliczalnymi miarami. Standardowym sposobem „diagonalizacji" dowodzenia $\mathcal{M}_1 \overset{\times}{\not\geq} \mathcal{M}_2$ jest wzięcie dowolnego μ ∈ M_1 , zwiększenie" go do ρ tak, aby $\mu \overset{\times}{\not\geq} \rho$ i pokazanie, że ρ ∈ M_2 Istnieje 7 x7 kombinacji (pół)miar M_1 z M_2 dla którego $\mathcal{M}_1 \overset{\times}{\geq} \mathcal{M}_2$ może być prawdą lub fałszem. Istnieją cztery podstawowe przypadki, wyjaśnione w następującym twierdzeniu, z których pozostałe 49 kombinacji przedstawionych w Tabeli wynika przez przechodniość.

↖	$\mathcal{M}$	semimeasure				measure		
ρ	↘	comp.	est.	enum.	appr.	comp.	est.	appr.
s	comp.	no^{iii}	no^{iii}	no^{iii}	no^{iv}	no^{iii}	no^{iii}	no^{iv}
e	est.	no^{iii}	no^{iii}	no^{iii}	no^{iv}	**no^{iii}**	no^{iii}	no^{iv}
m	enum.	yes^{i}	yes^{i}	**yes^{i}**	no^{iv}	yes^{i}	yes^{i}	no^{iv}
i	appr.	yes^{i}	yes^{i}	yes^{i}	no^{iv}	yes^{i}	yes^{i}	**no^{iv}**
m	comp.	no^{iii}	no^{iii}	no^{iii}	no^{iv}	no^{iii}	no^{iii}	no^{iv}
s	est.	no^{iii}	no^{iii}	no^{iii}	no^{iv}	no^{iii}	no^{iii}	no^{iv}
r	appr.	yes^{ii}	yes^{ii}	**yes^{ii}**	no^{iv}	yes^{ii}	yes^{ii}	no^{iv}

Twierdzenie 2.28. (Uniwersalne) półmiary. Półmiary RHO nazywamy uniwersalnymi dla M, jeżeli dominują multiplikatywnie nad wszystkimi elementami M w tym sensie, $\forall \nu \exists w_\nu > 0 : \rho(x) \geq w_\nu \nu(x) \forall x$. Prawda jest taka, że:

(o) $\exists \rho : \{\rho\} \overset{\times}{\geq} \mathcal{M}$: Dla każdego policzalnego zbioru (pół)miar M istnieje (pół)miara, która dominuje nad wszystkimi elementami M

(i) $\mathcal{M}^{semi}_{enum} \overset{\times}{\geq} \mathcal{M}^{semi}_{enum}$: Klasa przeliczalnych półmiar zawiera uniwersalne elementy

(ii) $\mathcal{M}^{msr}_{appr} \overset{\times}{\geq} \mathcal{M}^{semi}_{enum}$: Istnieje przybliżona miara, która dominuje nad wszystkimi przeliczalnymi półmiarami

(iii) $\mathcal{M}^{semi}_{est} \overset{\times}{\not\geq} \mathcal{M}^{msr}_{comp}$: Nie istnieje szacowana półmiara, która dominuje nad wszystkimi obliczalnymi miarami

(iv) $\mathcal{M}_{appr}^{semi} \not\gtrsim \mathcal{M}_{appr}^{msr}$: Nie istnieje przybliżona półmiara, która dominuje nad wszystkimi przybliżalnymi miarami

Jeżeli poprosimy o uniwersalną (pół)miarę, która przynajmniej spełnia najsłabszą formę obliczalności, mianowicie jest aproksymowalna, zobaczymy, że największym zdominowanym zbiorem spośród siedmiu zbiorów zdefiniowanych powyżej jest zbiór przeliczalnych półmiar. To jest powód, dla którego M^{semi}_{enum} odgrywa szczególną rolę w tej (i innych) pracy. Z drugiej strony, M^{semi}_{enum} to nie największy zbiór zdominowany przez aproksymowalną półmiarę, i rzeczywiście żaden taki największy zbiór nie istnieje. Można zatem poprosić o „naturalne" większe zbiory M.. Jeden taki zbiór, mianowicie zbiór kumulatywnie przeliczalnych półmiar AICEM, został niedawno odkryty przez Schmidhubera, dla którego zachodzi nawet $\xi_{CEM} \in M_{CEM}$. Twierdzenie 2.28 zachodzi również dla iscrete (semi)miar P zdefiniowanych jako

$$P : \mathbb{N} \to [0,1] \quad \text{with} \quad \sum_{x \in \mathbb{N}} P(x) \stackrel{(<)}{=} 1.$$

Najpierw udowadniamy twierdzenie dla tego dyskretnego przypadku, ponieważ zawiera ono istotne idee w czystszej formie. Następnie przedstawiamy dowód dla „ciągłych" (pół)miar μ (Definicja 2.22). Dowody naturalnie uogólniają się od binarnego do dowolnego skończonego alfabetu. Wartość a: minimalizująca f{x) jest oznaczana przez $\text{argmin}_x f(x)$. Remisy są rozwiązywane w dowolny, ale obliczalny sposób (np. przez wzięcie najmniejszego x).

Dowód (przypadek dyskretny), $(o)\ Q(x) := \sum_{P \in \mathcal{M}} w_P P(x)$ z $w_p>0$ wyraźnie dominuje nad wszystkimi P ∈ M. (ze stałym w_p). Ponieważ $\sum_P w_P = 1$ i wszystkie P są dyskretnymi (pół)miarami, Q jest również dyskretną (pół)miarą.

(i) Niech P będzie elementem uniwersalnym w M^{semi}_{enum} i $\alpha := \sum_x P(x)$. Normalizujemy P przez Q(x):=1/αP(x). Ponieważ α ≤ l mamy Q(x) ≥ P{x), stąd $Q \geq P \stackrel{\times}{\geq} \mathcal{M}_{enum}^{semi}$. Jako stosunek między dwiema funkcjami wyliczalnymi, Q jest nadal aproksymowalne, stąd $\mathcal{M}_{appr}^{msr} \stackrel{\times}{\geq} \mathcal{M}_{enum}^{semi}$.

(ii) Niech $P \in \mathcal{M}_{comp}^{semi}$. Dzielimy IN na fragmenty I_n:={2^{n-1}, …, 2^n-1} (n ≥ 1) o rosnącym rozmiarze. Przy x_n :=$\text{argmin}_{x \in I_n} P(x)$ definiujemy $Q(x_n) := \frac{1}{n(n+1)} \forall n$ i Q{x) :=0 dla wszystkich pozostałych x. Wykorzystując fakt, że minimum jest mniejsze od średniej, otrzymujemy

$$P(x_n) = \min_{x \in I_n} P(x) \leq \frac{1}{|I_n|} \sum_{x \in I_n} P(x) \leq \frac{1}{|I_n|} = \frac{1}{2^{n-1}} = \frac{n(n+1)}{2^{n-1}} Q(x_n).$$

Ponieważ $n(n+1)/2^{n-1}$ -> 0 dla n->∞, P nie może dominować nad $Q\ (P \not\gtrsim Q)$. Z P również Q jest obliczalne. Ponieważ P było dowolną obliczalną półmiarą, a Q jest obliczalną miarą $(\sum Q(x) = \sum[\frac{1}{n(n+1)}] = \sum[\frac{1}{n} - \frac{1}{n+1}] = 1)$, to implikuje $\mathcal{M}_{comp}^{semi} \not\gtrsim \mathcal{M}_{comp}^{msr}$. Załóżmy teraz, że istnieje szacowana półmiara $S \stackrel{\times}{\geq} \mathcal{M}_{comp}^{msr}$. Konstruujemy skończenie obliczalną półmiarę $P \stackrel{\times}{\geq} S$ w następujący sposób. Wybieramy początkowe ε > 0 i skończenie obliczamy ε-przybliżenie $\hat{S}$ z S(x). Jeśli $\hat{S} > 2\varepsilon$ definiujemy P(x) = 1/2$\hat{S}$ w przeciwnym razie dzielimy ε na pół i powtarzamy proces. Ponieważ

S{x) > 0 (w przeciwnym razie nie mogłoby dominować, np. $T(x) := \frac{1}{x(x+1)} \in \mathcal{M}^{msr}_{comp}$) kończy się po skończonym czasie. Tak więc P jest skończenie obliczalne. Wstawiając $\hat{S}$ = 2P(x) i ε < 1/2$\hat{S}$ = P(x) do |S{x)-$\hat{S}$| 0 < ε, otrzymujemy |S{x)-2P{x)| < P{x), co implikuje S{x) ≥ P{x) i S{x) ≤ 3P(x). Pierwsze implikuje $\sum_x P(x) \leq \sum_x S(x) \leq 1$, ,tj P jest półmiarą. Drugie implikuje $P \geq \frac{1}{3} S \overset{\times}{\geq} \mathcal{M}^{msr}_{comp}$. Stąd P jest obliczalną półmiarą dominującą nad wszystkimi obliczalnymi miarami, co przeczy temu, co udowodniliśmy w pierwszej połowie (iii). Stąd założenie dotyczące S było błędne, co ustanawia $\mathcal{M}^{semi}_{est} \overset{\times}{\not\geq} \mathcal{M}^{msr}_{comp}$.

(iv) Załóżmy, że $P \in \mathcal{M}^{semi}_{appr} \overset{\times}{\geq} \mathcal{M}^{msr}_{appr}$. Konstruujemy przybliżoną miarę Q, która nie jest zdominowana przez P, co zaprzecza założeniu. Niech $P_1, P_2 \ldots$ będzie ciągiem funkcji rekurencyjnych zbieżnych do P. Konstruujemy $x_1, x_2, \ldots$ takie, że $\forall c > 0 \ \exists n \in \mathbb{N} : P(x_n) \not\geq c \cdot Q(x_n)$. W tym celu rekurencyjnie definiujemy ciągi $x^1_n, x^2_n, \ldots$ zbieżne do x_n, a z nich Q_1, Q_2 zbieżne do Q. Niech In := $\{2^{n-1}, \ldots, 2^n-1\}$ i $x^1_n = 2^{n-1} \forall n$. Jeśli $P_t(x^{t-1}_n) > n^{-3}$ wtedy $x^t_n := \operatorname{argmin}_{x \in I_n} P_t(x)$ w przeciwnym razie $x^t_n := x^{t-1}_n$. Pokażemy, że x^t_n jest zbieżne dla t -> ∞, zakładając coś przeciwnego i wykazując sprzeczność. Ponieważ $x^t_n \in$ In pewnej wartości, powiedzmy x^*_n, jest zakładane nieskończenie często. Niezbieżność oznacza, że ciąg opuszcza i powraca do x^*_n nieskończenie często, x^*_n pozostaje tylko $(x^{t-1}_n = x^*_n \neq x^t_n)$ jeśli $P_t(x^*_n) > n^{-3}$. Z drugiej strony, w chwili, gdy x^t_n powraca do x^*_n $(x^{t-1}_n \neq x^*_n = x^t_n)$ mamy $P_t(x^*_n) = P_t(x^t_n) = \min_{x \in I_n} P_t(x) \leq |I_n|^{-1} = 2^{-n+1}$. Stąd $P_t(x^*_n)$ oscyluje (dla n ≥ 12) nieskończenie często pomiędzy ≤ 2^{-n+1} i ≥ n^{-3}, co przeczy założeniu, że P_t jest zbieżny. Stąd założenie o niekonwergentnym x^t_n było błędne, x^t_n jest zbieżne do x^*_n, a $P_t(x^*_n)$ do wartości ≤n^{-3}. Przy x^t_n również miara $Q_t(x^t_n) := \frac{1}{n(n+1)}$ (i $Q_t(x)$ = 0 dla wszystkich innych x) jest zbieżna. Ponieważ $P(x^*_n) \leq n^{-3}$ nie dominuje nad $Q(x^*_n)$, mamy $P \overset{\times}{\not\geq} Q$. Ponieważ P ∈ M^{semi}_{appr} było dowolne, a Q jest miarą aproksymacyjną, otrzymujemy $\mathcal{M}^{semi}_{appr} \overset{\times}{\not\geq} \mathcal{M}^{msr}_{appr}$

Dowód (przypadek ciągły). Główną różnicą w stosunku do przypadku dyskretnego jest to, że trzeba również zadbać o to, aby $\rho(x) \geq \rho(x0) + \rho(x1), \ x \in \mathbb{B}^*$ było respektowane. Z drugiej strony, dzielenie na fragmenty In := B^n jest tutaj bardziej naturalne.

$(o) \ \rho(x) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(x)$ z w_ν > 0 Wjy>0 wyraźnie dominuje nad wszystkimi v ∈ M (ze stałą dominacji w_ν). Ponieważ $\sum_\nu w_\nu = 1$ i wszystkie v są (pół)miarami, to ρ jest również (pół)miarą.

(ii) Niech ξ będzie elementem uniwersalnym w M^{semi}_{enum}. Definiujemy

$$\xi_{norm}(x_{1:n}) \ := \ \prod_{t=1}^{n} \frac{\xi(x_{1:t})}{\xi(x_{<t}0) + \xi(x_{<t}1)}. \qquad (2.30)$$

Przez indukcję można wykazać, że ξ_{norm} jest miarą i że ξnorm(x) ≥ ξ(x)∀x, stąd $\xi_{norm} \geq \xi \overset{\times}{\geq} \mathcal{M}^{semi}_{enum}$. Jako stosunek funkcji wyliczalnych, ξ_{norm} jest nadal aproksymowalny, stąd $\mathcal{M}^{msr}_{appr} \overset{\times}{\geq} \mathcal{M}^{semi}_{enum}$.

(iii) Niech $\mu \in \mathcal{M}^{semi}_{comp}$. Rekurencyjnie definiujemy ciąg $x^*_{1:\infty}$ przez $x^*_k := \text{argmin}_{x_k}\mu(x^*_{<k}x_k)$ i miarę ρ przez $\rho(x^*_{1:k}) = 1 \forall k$ i ρ(x)=0 dla wszystkich x które nie są prefiksami $x^*_{1:k}$. Wykorzystując fakt, że minimum jest mniejsze od średniej i że μ jest półmiarą, otrzymujemy

$$\mu(x^*_{1:k}) = \min_{x_k} \mu(x^*_{<k}x_k) \;\leq\; \tfrac{1}{2}[\mu(x^*_{<k}0)+\mu(x^*_{<k}1)] \;\leq\; \tfrac{1}{2}\mu(x^*_{<k})$$

Stąd $\mu(x^*_{1:n}) \leq (\tfrac{1}{2})^n = (\tfrac{1}{2})^n \rho(x^*_{1:n})$ co pokazuje, że μ nie dominuje nad p. Ponieważ $\mu \in \mathcal{M}^{semi}_{comp}$ i ρ dowolne, a p jest miarą obliczalną, implikuje to $\mathcal{M}^{semi}_{comp} \not\gtrsim \mathcal{M}^{msr}_{comp}$. Załóżmy teraz, że istnieje szacowana półmiara $\sigma \in \mathcal{M}^{msr}_{comp}$. Konstruujemy skończenie obliczalną funkcję $\mu \gtrsim \sigma$ w następujący sposób. Wybieramy początkowe ε > 0 i skończenie obliczamy ε-przybliżenie $\hat{\sigma}$ z σ(x). Jeśli $\hat{\sigma} > 4\varepsilon$ definiujemy $\mu(x) := \hat{\sigma}$, w przeciwnym razie dzielimy ε na pół i powtarzamy proces. Ponieważ σ(x) > 0 (w przeciwnym razie nie mogłoby dominować, np. $2^{-\ell(x)}$), pętla kończy się po skończonym czasie. Więc μ jest skończenie obliczalne. Wstawiając $\hat{\sigma} = \mu(x)$ oraz $\varepsilon < \tfrac{1}{4}\hat{\sigma} = \tfrac{1}{4}\mu(x)$ do $|\sigma(x) - \hat{\sigma}| < \varepsilon$ otrzymujemy $|\sigma(x) - \mu(x)| < \tfrac{1}{4}\mu(x)$), co implikuje $\tfrac{3}{4}\mu(x) \leq \sigma(x) \leq \tfrac{5}{4}\mu(x)$. Niestety μ nie jest półmiarą, ale nadal spełnia słabszą nierówność $\mu(x0)+\mu(x1) \leq \tfrac{4}{3}[\sigma(x0)+\sigma(x1)] \leq \tfrac{4}{3}\sigma(x) \leq \tfrac{4}{3}\cdot\tfrac{5}{4}\mu(x) = \tfrac{5}{3}\mu(x)$. Czy to wystarczy, aby pierwsza połowa dowodu (iii) przebiegła z 1/2 zastąpionym przez $\tfrac{1}{2}\cdot\tfrac{5}{3} = \tfrac{5}{6} < 1$, co pokazuje, że $\mu \not\gtrsim \mathcal{M}^{msr}_{comp}$. Jednakże, to przeczy $\mu \geq \tfrac{4}{5}\sigma \in \mathcal{M}^{msr}_{comp}$ pokazując, że nasza założona szacowalna półmiara σ nie istnieje, tj. $\mathcal{M}^{semi}_{est} \not\gtrsim \mathcal{M}^{msr}_{comp}$

(iv) Załóżmy $\mu \in \mathcal{M}^{semi}_{appr} \not\gtrsim \mathcal{M}^{msr}_{appr}$. Konstruujemy przybliżoną miarę ρ, która nie jest zdominowana przez μ, co przeczy założeniu. Niech $\mu_1\mu_2$,… będzie ciągiem funkcji rekurencyjnych zbieżnych do μ. Rekurencyjnie (w t i n) definiujemy ciągi $y^1_n, y^2_n, \ldots, y^t_n$ zbieżne do y_n i z nich $\rho_1\rho_2$ … zbieżne do ρ. Niech $y^1_n = 0 \forall n$. Jeśli $\mu_t(y^t_{<n}y^{t-1}_n) > \tfrac{2}{3}\mu_t(y^t_{<n})$, wtedy $y^t_n := \text{argmin}_{x_n}\mu_t(y^t_{<n}x_n)$, w przeciwnym razie $y^t_n := y^{t-1}_n$.'Wykazujemy, że y^t_n zbiega się dla t -> ∞, zakładając odwrotność i wykazując sprzeczność. Załóżmy, że k jest najmniejszym n, dla którego $y^t_n \not\to y_n$. Ponieważ y^t_n -> y_n dla wszystkich n < k i y^t_n ∈ B jest dyskretne, istnieje t_0 takie, że $y^t_{<k} = y_{<k}\, \forall t > t_0$. Załóżmy, że t > t_0 w poniższym przykładzie. Ponieważ y^t_k ∈ B pewna wartość, powiedzmy $\tilde{y}_k$, jest zakładana nieskończenie często. Niezbieżność oznacza, że ciąg opuszcza i wchodzi do $\tilde{y}_k$ nieskończenie często. Jeśli $\tilde{y}_k$ jest opuszczony $(y^{t-1}_k = \tilde{y}_k \neq y^t_k)$ mamy

$$\mu_t(y_{<k}\tilde{y}_k) \;=\; \mu_t(y^t_{<k}y^{t-1}_k) \;>\; \tfrac{2}{3}\mu_t(y^t_{<k}) \;=\; \tfrac{2}{3}\mu_t(y_{<k}) \;\stackrel{t\to\infty}{\longrightarrow}\; \tfrac{2}{3}\mu(y_{<k}).$$

Jeśli $\tilde{y}_k$ jest wpisany $(y^{t-1}_k \neq \tilde{y}_k = y^t_k)$ mamy,

$$\mu_t(y_{<k}\tilde{y}_k) \;=\; \mu_t(y^t_{<k}y^t_k) \;=\; \min_{x_k}\mu_t(y^t_{<k}x_k) \;\leq\; \tfrac{1}{2}[\mu_t(y^t_{<k}0)+\mu_t(y^t_{<k}1)] \;\leq$$
$$\leq\; \tfrac{1}{2}\mu_t(y^t_{<k}) \;=\; \tfrac{1}{2}\mu_t(y_{<k}) \;\stackrel{t\to\infty}{\longrightarrow}\; \tfrac{1}{2}\mu(y_{<k}).$$

Stąd $\mu_t(y_{<k}\tilde{y}_k)$ oscyluje nieskończenie często pomiędzy $>\frac{2}{3}\mu(y_{<k})$ i $\leq\frac{1}{2}\mu(y_{<k})$, co przeczy założeniu, że μ_t jest zbieżne. Stąd założenie o niekonwergentnym y^t_k było błędne. Z y^t_k również miara $\rho_t(y^t_{1:n}) := 1$ (i $\rho_t\{x) = 0$ dla wszystkich innych x które nie są prefiksami $y^t_{1:\infty}$ jest zbieżna. Dla każdego dostatecznie dużego t mamy $y_{1:n} = y^t_{1:n}$ $\mu_t(y_{1:n}) = \mu_t(y^t_{1:n}) \leq \frac{2}{3}\mu_t(y^t_{<n}) \leq \ldots \leq (\frac{2}{3})^n$. Ponieważ $\mu(y_{1:n}) \leq (\frac{2}{3})^n$ nie dominują nad $\rho(y_{1:n}) = 1 \;(\forall t > t_0)$ mamy $\mu \not\gtrsim \rho$. Ponieważ $\mu \in M^{semi}_{appr}$ było dowolne, a ρ jest przybliżoną miarą, otrzymujemy $\mathcal{M}^{semi}_{appr} \not\gtrsim \mathcal{M}^{msr}_{appr}$.

Losowość Martina-Löfa

Losowość Martina-Lofa jest bardzo ważną koncepcją losowości pojedynczych sekwencji, która jest ściśle związana ze złożonością Kolmogorowa i uniwersalnym priorem Solomonoffa. Podajemy charakterystykę równoważną oryginalnej definicji Martina-Löfa, aby ominąć konieczność podawania formalnej definicji „efektywnych testów losowości" :

Twierdzenie 2.31 (Losowy ciąg Martina-Löfa). Ciąg $x_{1:\infty}$ nazywamy losowym ciągiem μ-Martina-Löfa, jeśli istnieje stała c taka, że $M(x_{1:n}) \leq c\cdot\mu(x_{1:n})$ dla wszystkich n .

Równoważna formuła dla obliczalnego μ jest następująca:

$$x_{1:\infty} \text{ is } \mu\text{.M.L.-random} \quad\Leftrightarrow\quad Km(x_{1:n}) \stackrel{+}{=} -\log_2\mu(x_{1:n})\;\forall n, \qquad (2.32)$$

gdzie $Km\{x_{1:n}$) jest długością najkrótszego (możliwie nieprzerwanego) programu obliczającego ciąg znaków zaczynający się od $x_{1:n}$. Twierdzenie 2.31 wynika z (2.32) przez potęgowanie, „używając 2-Km ≈ M" i zauważając, że $M \stackrel{\times}{\geq} \mu$ wynika z uniwersalności M. Rozważmy szczególny przypadek, w którym μ jest uczciwą monetą, tj. $\mu(x_{1:n}) = 2^{-n}$, wówczas $x1{:}\infty$ jest losowe M.L. wtedy i tylko wtedy, gdy $Km(x_{1:n}) \stackrel{+}{=} n$, tj. jeśli $x_{1:n}$ jest nieściśliwe. Dla ogólnego μ $-\log_2\mu(x_{1:n})$ jest długością kodu Shannona-Fano z $x_{1:n}$, stąd $x_1 \to \infty$ jest losowe wtedy i tylko wtedy, gdy kod Shannona-Fano jest optymalny. Można pokazać, że losowa sekwencja μ.M.L. $x_{1:\infty}$ przechodzi wszystkie możliwe testy losowości efektywnej, np. prawo wielkich liczb, prawo iterowanego logarytmu itd. W szczególności zbiór wszystkich μM.L.-losowe sekwencje mają μ-miarę 1. Następujące uogólnienie jest naturalne przy rozważaniu ogólnych mieszanin Bayesa ξ:

Definicja 2.33. (μ/ξ - sekwencja losowa). Sekwencja $x_{1:\infty}$ jest nazywana μ/ξ - losowa, jeśli istnieje stała c taka jak $\xi(x_{1:n}) \leq c\cdot\mu(x_{1:n})$ dla wszystkich n

Zwykle ξ jest mieszaniną pewnego M, jak zdefiniowano w (2.26), w którym to przypadku prawdziwa jest również odwrotna nierówność $\xi(x) \stackrel{\times}{\geq} \mu(x)$ (dla wszystkich x). Dla skończonego M lub jeśli $\xi \in M$, definicja losowości μ/ξ zależy tylko od M, a nie od konkretnych wag użytych w ξ. Dla $M=M_U$ losowość μ/ξ jest po prostu losowością μ.M.L. Im większe M, tym więcej wzorców jest rozpoznawanych jako nielosowe. Mówiąc ogólnie, te regularności charakteryzujące się pewnym $\nu \in M$ są rozpoznawane

przez losowość μ/ξ, tj. dla $M \subset M_U$ niektóre losowe ciągi μ/ξ mogą nie być losowe M.L. Inne koncepcje losowości, np. te autorstwa Schnorra, Ko, van Lambalgena, Lutza, Kurtza, von Misesa, Walda i Churcha można by również scharakteryzować w kategoriach losowości μ/ξ dla konkretnych wyborów M.