SNUM – Lab 1

Drzewa decyzyjne

Iwo Błądek, Konrad Miazga

1 Drzewa decyzyjne

Algorytmy indukcji drzew decyzyjnych to jeden z klasycznych algorytmów uczenia maszynowego służący do rozwiązywania problemu *klasyfikacji*. Drzewa decyzyjne reprezentują wiedzę uzyskaną z danych w sposób symboliczny, który jest łatwy do odczytania i analizy dla człowieka (w przeciwieństwie np. do sztucznych sieci neuronowych, gdzie wiedza jest zawarta w wagach połączeń między neuronami).

1.1 Budowa drzewa

Na Rysunku 1 przedstawiony jest diagram przykładowego drzewa decyzyjnego dla problemu przewidywania, czy klient kupi jakiś towar. Na Rysunku 2 przedstawiona jest graficzna interpretacja koncepcji drzewa decyzyjnego w kontekście botaniki. Drzewa decyzyjne przyjęło się rysować "w dół", czyli schodząc od korzenia do liści. Nie ma to jednak większego znaczenia, czy korzeń w drzewie decyzyjnym będziemy rysować na dole (i wtedy kolejne węzły z warunkami będą piąć się w górę), czy też na górze (tak jak na rysunkach).

W ogólności drzewo decyzyjne można traktować jako *skierowany graf acykliczny* (a konkretniej *drzewo*) i reprezentować je za pomocą węzłów oraz krawędzi je łączących. Ważne pojęcia w takim ujęciu to:

węzeł – na rysunku elipsy lub prostokąty. W węźle znajduje się zawsze albo test na wartość pewnego atrybutu (np. elipsa 'Student', odpowiadająca pytaniu, czy klient był studentem), albo klasa decyzyjna (tutaj prostokąty 'TAK', 'NIE'). Konkretne figury geometryczne zostały wprowadzone tylko by ułatwić rozróżnianie węzłów, tzn. nie są elementem konwencji.

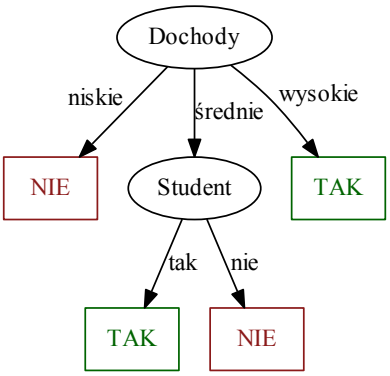
krawędź – łączy dwa węzły, przy czym strzałka określa kierunek połączenia (i zawsze wskazuje w dół, jeżeli korzeń jest na górze). Z każdą krawędzią związana jest pewna **etykieta**, np. *niskie*, *średnie*, *tak*, *nie*. Etykiety to możliwe wartości atrybutu związanego z węzłem, z którego wyszła ta krawędź.

liść – taki węzeł, z którego nie wychodzi żadna krawędź. W liściu *zawsze* znajdować się będzie przypisanie do jakiejś klasy decyzyjnej, w tym wypadku do 'TAK' albo 'NIE'. Na rysunku liście to węzły w kształcie prostokątów.

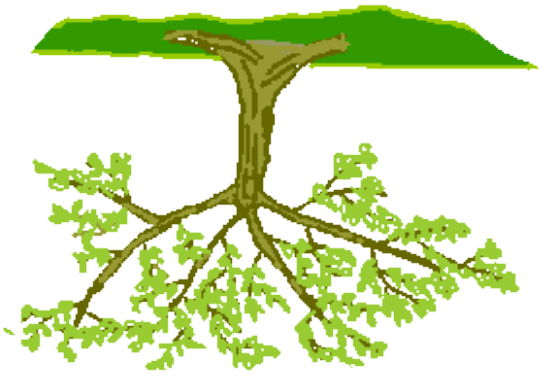
korzeń – od korzenia drzewo zaczyna rosnąć. Jest to węzeł, do którego nie dochodzi żadna krawędź, czyli w tym wypadku *Dochody*.

1.2 Klasyfikacja przykładów

Skupmy się na jakiejś ścieżce w drzewie, np. $Dochody \rightarrow niskie \rightarrow NIE$. Węzeł *Dochody* to test na wartość atrybutu *Dochody*. Nowe przypadki z wartością *niskie* na tym atrybucie przejdą w drzewie po krawędzi *niskie*, *średnie* po krawędzi *średnie* itd. Po przejściu krawędzią przypadek może napotkać na kolejny test, tym razem na innym atrybucie, i procedura się powtarza. Prędzej czy później każdy przypadek skończy w jakimś liściu zawierającym przydział do klasy decyzyjnej



Rysunek 1: Przykładowe drzewo decyzyjne.



Rysunek 2: Drzewa decyzyjne rosną ku dołowi.

(w każdym kroku schodzimy w dół, a drzewo ma skończoną wysokość). Przypadek jest *zawsze* przydzielany do klasy decyzyjnej zawartej w liściu, na którym zakończył „podróż".

Drzewo decyzyjne można alternatywnie zapisać w postaci reguł określających przydział obiektów do klas. Każda ścieżka od korzenia do liścia odpowiada jednej regule.

Przykład 1.1 — Zapisywanie drzewa w postaci reguł. Przedstawimy drzewo z Rysunku 1 w postaci reguł.

jeżeli $Dochody = niskie$ **to** $Decyzja = NIE$
jeżeli $Dochody = wysokie$ **to** $Decyzja = TAK$
jeżeli $Dochody = średnie \wedge Student = tak$ **to** $Decyzja = TAK$
jeżeli $Dochody = średnie \wedge Student = nie$ **to** $Decyzja = NIE$

Równoważnie, reguły można zapisywać w ten sposób:

$$\forall_x \; Dochody(x, niskie) \implies Decyzja(x, NIE),$$

gdzie dziedziną x jest zbiór wszystkich możliwych przypadków. ■

Reguły pozwalają zobaczyć, czym „tak naprawdę" jest drzewo decyzyjne oraz w jaki sposób dokonywana jest klasyfikacja obiektów. W drzewach decyzyjnych wiedza reprezentowana jest zasadniczo w postaci pewnych zdań logicznych. **Tego typu reprezentacja jest łatwa do interpretacji przez człowieka, co jest dużą zaletą tych metod.**

1.3 Entropia

Zanim powiemy, jak tworzone są drzewa decyzyjne, musimy wyjaśnić pojęcie **entropii**. Pochodzi ono z teorii informacji i wykorzystane zostanie jako heurystyka do wyboru najlepszego w danym momencie atrybutu do podziału zbioru

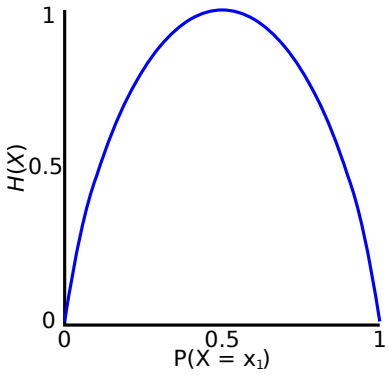
Załóżmy, że mamy zmienną losową X o zbiorze wartości $\{x_1, x_2, ..., x_n\}$. Intuicyjnie, zmienna losowa to funkcja, która zwraca którąś ze swoich dopuszczalnych wartości (zgodnie z pewnym rozkładem prawdopodobieństwa). Entropię tej zmiennej wyliczymy ze wzoru[1]:

$$H(X) = \sum_{i=1}^{n} p(x_i) \cdot \log_r \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \cdot \log_r p(x_i)$$

Podstawa logarytmu r wyraża nam liczbę bazowych stanów, za pomocą których możemy kodować informację. Zazwyczaj przyjmuje się $r = 2$, co odpowiada kodowaniu binarnemu. $p(x_i)$ to z kolei prawdopodobieństwo wylosowania wartości x_i.

[1] Literka H wzięła się z inspiracji *H-theorem* wprowadzoną przez Boltzmanna (dziedzina termodynamiki).

Wykres entropii w przypadku **dwóch** możliwych wartości (x_1, x_2) zmiennej losowej X, dla różnych prawdopodobieństw wylosowania x_1 na osi x wykresu (z założeń wiemy, że $p(x_2) = 1 - p(x_1)$), wygląda następująco:



Jak można zauważyć, entropia osiąga 0 w przypadku, gdy któraś z wartości zmiennej losowej X ma prawdopodobieństwo wylosowania równe 1. Jeżeli wartości są równo prawdopodobne ($p(x_1) = 0.5$, $p(x_2) = 0.5$), to entropia osiąga maksimum, czyli wartość 1.

Przykład 1.2 — Obliczanie entropii. Załóżmy, że zmienna losowa X przyjmuje:

- wartość x_1 z prawdopodobieństwem $\frac{2}{3}$ i x_2 z prawdopodobieństwem $\frac{1}{3}$:

$$H(X) = -\frac{2}{3} \cdot log_2 \frac{2}{3} - \frac{1}{3} \cdot log_2 \frac{1}{3} \approx -\frac{2}{3} \cdot (-0.5849) - \frac{1}{3} \cdot (-1.5849) \approx 0.918$$

- wartość x_1 z prawdop. 1 i x_2 z prawdop. 0:

$$H(X) = -1 \cdot log_2 1 - 0 \cdot log_2 0 = 0 - 0 = 0$$

- wartość x_1 z prawdop. 0.5 i x_2 z prawdop. 0.5:

$$H(X) = -0.5 \cdot log_2 0.5 - 0.5 \cdot log_2 0.5 = -0.5 \cdot (-1) - 0.5 \cdot (-1) = 0.5 + 0.5 = 1$$

■

Warto jeszcze coś wspomnieć o interpretacji entropii. Można ją traktować jako miarę nieprzewidywalności, „nieuporządkowania" zmiennej losowej. W drzewach decyzyjnych chcemy ją minimalizować.

1.4 Uczenie – Algorytm ID3

Informacje wstępne

Celem uczenia drzew decyzyjnych jest utworzenie **jak najmniejszego** drzewa. Łatwo zauważyć, że stworzenie na podstawie przykładów uczących *jakiegoś* drzewa jest bardzo proste – wystarczy utworzyć osobną ścieżkę dla każdego przykładu uczącego. Jednak takie drzewa są obciążające obliczeniowo podczas klasyfikacji i mają słabą zdolność uogólniania (powstały model nie dostarcza nam żadnej interesującej wiedzy). Dlatego jesteśmy zainteresowani tym, by opisać nasze przykłady uczące możliwie niewielkim drzewem (można tu zauważyć pewną analogię do brzytwy Ockhama).

W celu utworzenia minimalnego drzewa można by zastosować algorytm dokładny, jednak złożoność takiego algorytmu byłaby wykładnicza. W praktyce nie musimy posiadać absolutnie najmniejszego drzewa i zadowalamy się „dość małym" drzewem wygenerowanym przez heurystykę. Heurystyka ta oparta jest właśnie na mierze entropii.

Najbardziej podstawowym algorytmem uczenia drzew decyzyjnych jest ID3 (*Iterative Dichotomiser 3*). Dychotomiczny podział zbioru to taki jego podział na 2 podzbiory A i B, że nie mają one części wspólnej i po zsumowaniu dadzą zbiór wyjściowy.

Omówienie na przykładzie

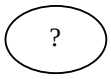
Będziemy rozważać zbiór danych uczących S przedstawiony w poniższej tabelce. Interesujące nas atrybuty *Matematyka*, *Biologia* i *Polski* przyjmują wartości ze zbioru $\{3, 4, 5\}$. *Uczeń* jest identyfikatorem (nie bierze udziału w uczeniu), a *Decyzja* określa prawdziwe klasy decyzyjne przykładów. Problemem jest odpowiedzenie na pytanie, czy uczeń dostanie stypendium biorąc pod uwagę jego oceny.

Uczeń	Matematyka	Biologia	Polski	Decyzja
A	4	4	5	TAK
B	4	5	4	TAK
C	3	4	4	NIE
D	5	3	5	NIE
E	4	4	4	NIE
F	3	5	3	NIE

Czy potrafisz odgadnąć, jaką metodą ktoś podejmował te decyzje? Odpowiedź: stypendia przyznawane były tym studentom, którzy mieli z przynajmniej jednego przedmiotu 5 i nie dostali żadnej oceny niższej niż 4.

Entropia ze względu na decyzje tego całego zbioru ($\{T, T, N, N, N, N\}$) wynosi:

$$H(S) = -\frac{2}{6} \cdot log_2 \frac{2}{6} - \frac{4}{6} \cdot log_2 \frac{4}{6} \approx 0.917$$



Konstrukcję drzewa zawsze zaczynamy od utworzenia korzenia. Sprawdzamy, czy wszystkie przykłady A,B,C,D,E,F są zaklasyfikowane do tej samej klasy decyzyjnej. W tym wypadku nie są, więc szukamy atrybutu najlepiej dzielącego przykłady ze względu na miarę entropii. Obliczenia zaczniemy od *Matematyki*.

Wypisujemy podziały przypadków ze względu na ich wartości na atrybucie *Matematyka*:

wartość atrybutu	częstość	przypadki	entropia
3	$^2/_6$	C(NIE), F(NIE)	0
4	$^3/_6$	A(TAK), B(TAK), E(NIE)	0.917
5	$^1/_6$	D(NIE)	0

Entropie zostały policzone ze względu na rozkład decyzji w przypadkach. Dla wartości 4 entropia przypadkowo wyszła taka sama jak $H(S)$, ale nie ma to szczególnego znaczenia. W ogólności entropia w poszczególnych gałęziach może wyjść dowolna (nawet większa), jednak jak uśrednimy wszystko ważąc po częstościach to i tak $H(S|Matematyka)$ nie przekroczy $H(S)$. Entropia ze względu na podział (który oznaczamy poniżej symbolem '|') na atrybucie *Matematyka* będzie średnią arytmetyczną wszystkich entropii z tabelki ważoną po odpowiadających im częstościach:

$$H(S|Matematyka) = {}^2/_6 \cdot 0 + {}^3/_6 \cdot 0.917 + {}^1/_6 \cdot 0 \approx 0.459$$

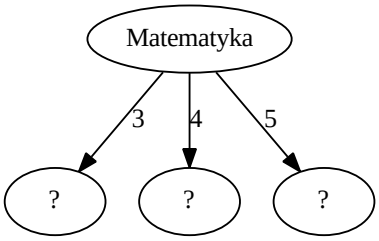
Analogiczne tabelki i wyliczenia możemy wykonać dla atrybutów *Biologia* i *Polski*. Możesz je zrobić jako ćwiczenie, entropie powinny wyjść następujące:

$$H(S|Biologia) = 0.791$$
$$H(S|Polski) = 0.791$$

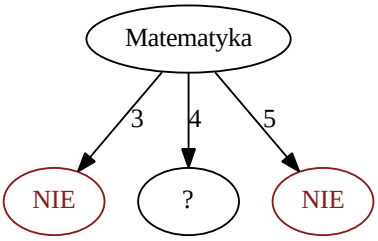
W algorytmie ID3 wykorzystuje się miarę **InformationGain** (zapisywane we wzorach jako *InfoGain*). Oblicza się ją poprzez odjęcie od początkowej entropii ($H(S)$) entropii po podziale na pewnym atrybucie (np. $H(S|Biologia)$). *InformationGain* wyraża, ile informacji „zyskaliśmy" na pojedynczym podziale.

$$InfoGain(S|Matematyka) = H(S) - H(S|Matematyka) = 0.917 - 0.459 = 0.458$$
$$InfoGain(S|Biologia) = H(S) - H(S|Biologia) = 0.917 - 0.791 = 0.126$$
$$InfoGain(S|Polski) = H(S) - H(S|Polski) = 0.917 - 0.791 = 0.126$$

Jak widzimy, zysk największy był w podziale na *Matematykę*, więc ten atrybut umieszczamy w korzeniu i rysujemy odpowiednie krawędzie pamiętając o etykietach.



Umieszczamy w korzeniu atrybut *Matematyka* i tworzymy gałęzie dla wszystkich przyjmowanych przez niego wartości (3, 4, 5).



Sprawdzamy, czy w danym podziale ze względu na ocenę z *Matematyki* elementy należą do tej samej klasy decyzyjnej. Tak jest dla zbiorów zawierających przykłady z ocenami 3 i 5, więc możemy utworzyć liście z decyzjami. W zbiorze przykładów z oceną 4 są dwa przypadki TAK i jeden NIE, tak więc będziemy musieli rozpatrywać podział dla zbioru przypadków $\{A,B,E\}$ ze względu na pozostałe atrybuty: *Biologia*, *Polski*.

Rozpatrujemy więc podziały na zbiorze $S = \{A,B,E\}$ (tylko one miały 4 z *Matematyki*). Entropię ($H(S)$)w tym zbiorze możemy odczytać w pierwszej tabelce i jest to 0.917. Możemy zrobić kolejne tabelki podziałów:

Biologia:

wartość atrybutu	częstość	przypadki	entropia
3	0/3		
4	2/3	A(TAK), E(NIE)	1
5	1/3	B(TAK)	0

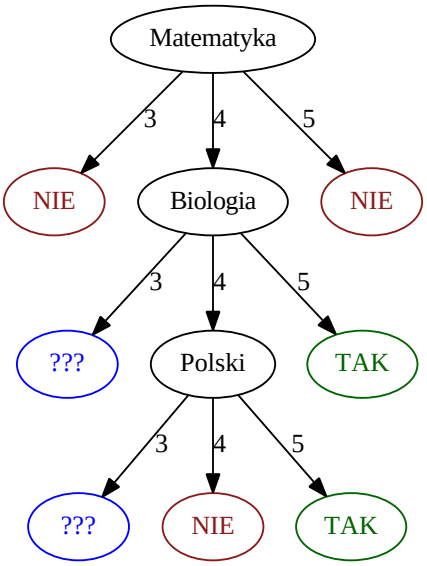
Polski:

wartość atrybutu	częstość	przypadki	entropia
3	0/3		
4	2/3	B(TAK), E(NIE)	1
5	1/3	A(TAK)	0

$$InfoGain(S|Biologia) = H(S) - H(S|Biologia) = 0.917 - 2/3 \cdot 1 \approx 0.25$$
$$InfoGain(S|Polski) = H(S) - H(S|Polski) = 0.917 - 2/3 \cdot 1 \approx 0.25$$

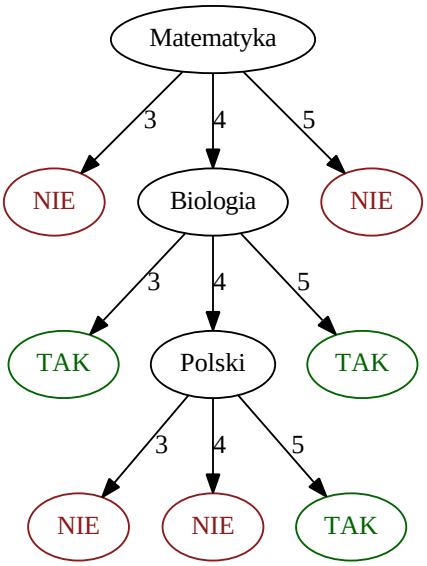
Nie ma znaczenia, który atrybut wybierzemy do podziału (wartości $InfoGain$ są równe) – wybieramy więc, zgodnie z kolejnością, *Biologię*. Zbiór dla oceny 4 z *Biologii* nadal nie zawiera przykładów z tylko jednej klasy decyzyjnej, więc wykonujemy dla niego jeszcze podział dla *Polskiego* (ostatni atrybut, jaki nam został).

Prawie-ostateczne drzewo wygląda tak:



Węzły **???** nie zawierają żadnych przykładów, więc nie możemy jednoznacznie stwierdzić, którą klasę decyzyjną umieścić w takim liściu. Domyślnie bierze się klasę najczęściej występującą na danym „poziomie". Dla 4 z *Matematyki* najczęściej występuje TAK, a dla 4 z *Matematyki* i 4 z *Biologii* NIE występuje tak samo często jak TAK. W tym drugim przypadku arbitralnie weźmiemy NIE.

Ostateczne drzewo (klasyfikator) przedstawione jest poniżej.



Komentarz

- Na żadnym etapie uczenia nie było brane pod uwagę uporządkowanie ocen. Wynika z tego, że **algorytm ID3 traktuje wszystkie dane tak jakby były nominalne**.
- Paradoksalnie, gdyby decyzje o stypendium były podejmowane przez nasz wyprodukowany klasyfikator, to ocena 5 z *Matematyki* automatycznie przekreślałaby na nie szanse! :) Z kolei osoba mająca 4 z *Matematyki* i 3 z *Biologii* dostałaby stypendium niezależnie od oceny z polskiego. **Wynika to przede wszystkim z wybrakowanych danych uczących**. W uczeniu maszynowym algorytm nauczy się tylko tego, co mu pokażemy, tak więc odpowiednio liczny i reprezentatywny zbiór uczący to podstawa. W naszym zbiorze uczącym 5 z *Matematyki* miała tylko jedna osoba, jednak nie dostała ona stypendium przez słabą ocenę z innego przedmiotu. Decyzja algorytmu w kontekście danych, z którymi pracował, była słuszna.

1.5 Uczenie – Algorytm C4.5

Algorytm **C4.5** jest ulepszoną wersją ID3. Ulepszenia obejmują:

- wprowadzenie miary *ilorazu przyrostu informacji* zamiast zysku informacji (*InfoGain*) w celu „karania" atrybutów o wielu różnych wartościach (są one niejawnie preferowane w ID3),
- wbudowane radzenie sobie z brakującymi wartościami (nie są brane pod uwagę podczas liczenia miar entropii),
- obsługa ciągłych wartości (podział na przedziały),
- upraszczanie drzewa po jego utworzeniu (post-pruning). Polega ono na usuwaniu któregoś z węzłów i wstawianiu na jego miejsce liścia zawierającego decyzję. Jeżeli taka zmiana poprawia jakość klasyfikacji na zbiorze testowym, to zmiana jest zachowywana.

Iloraz przyrostu informacji

Zanim zdefiniujemy **iloraz przyrostu informacji**, musimy wprowadzić **podział informacji** (*Split*). *Podział informacji* jest po prostu obliczeniem entropii dla zbioru wartości pewnego atrybutu (poprzednio liczyliśmy entropię ze względu na klasy decyzyjne).

Przykład 1.3 — Obliczanie podziału informacji. Wykorzystamy niedawno rozważaną tabelę opisującą uczniów i decyzje o stypendiach. Dla atrybutu *Matematyka* zbiór (właściwie to multizbiór) wartości wygląda następująco: $\{5, 4, 4, 4, 3, 3\}$. Entropię tego zbioru, czyli podział informacji, możemy policzyć jak każdą inną:

$$Split(S|Matematyka) = H(Matematyka) = -1/6 \cdot log_2 1/6 - 3/6 \cdot log_2 3/6 - 2/6 \cdot log_2 2/6 \approx 1.46$$

Entropia wyszła większa niż 1, ale jest to absolutnie normalne w przypadku kiedy mamy więcej niż dwie różne wartości w zbiorze (tu mieliśmy trzy: 5, 4 i 3). ■

Iloraz przyrostu informacji (*GainRatio*) wyraża się wzorem:

$$GainRatio(S|Atrybut) = \frac{InfoGain(S|Atrybut)}{Split(S|Atrybut)}$$

Podobnie jak przy zysku informacji chcemy go maksymalizować. Dzielenie przez *podział informacji*, czyli *Split*, sprawia, że wyrównane będą szanse atrybutów z dużą i małą liczbą różnych wartości na zbiorze przykładów uczących (ID3 niejawnie premiowało atrybuty o wielu przyjmowanych wartościach).