Drzewo decyzyjne

https://analityk.edu.pl/drzewo-decyzyjne-wstep-oraz-przyklady/

https://www.rafalszrajnert.pl/drzewo-decyzyjne/

Definicja

jest to przedstawienie graficzne procesu decyzyjnego. Najczęściej wykorzystuje się go w szkolnictwie. Charakteryzuje się tym, że przypomina budowę drzewa.

Ma **pień** – problem decyzyjny,
gałęzie – warianty wyjścia z sytuacji,
węzły – skutki (inaczej stany natury) pozytywne i negatywne,
koronę drzewa – cele i wartości danych decyzji.

Wykorzystywana także w naukach humanistycznych, gdzie wybór decyzji jest istotny, bądź istnieje wiele jej wariantów oraz w naukach ścisłych, gdzie jako algorytm wspomaga tzw. „uczenie maszynowe" będące podstawą dla np. sztucznej inteligencji.

Zaletą drzewa jest jego czytelność i symboliczne odzwierciedlenie podejmowania decyzji. Dzięki temu z łatwością można ten proces zmniejszać lub rozbudowywać.

Rodzaj drzew decyzyjnych

Rodzajem drzewa decyzyjnego jest diagram decyzyjny, różniący się od drzewka tym, że do danego węzła (skutku) można dojść wieloma drogami. Diagramy, w przeciwieństwie do dużych, rozbudowanych drzew, pozwalają w szybszy sposób zmniejszyć ilość powtarzających się gałęzi, co sprawia, że są użyteczne np. w analizach poprawności oprogramowania.

Do czego może nam posłużyć drzewo decyzyjne

Naszym celem jest zbudowanie algorytmu, który na podstawie danych pasażera, przedstawi nam prawdopodobieństwo czy dana osoba przeżyje czy nie.

Przykładowo, posiadamy dane: Kobieta, wiek 30, klasa biletu – premium, port docelowy – Michigan. W rezultacie chcemy otrzymać informację: 1 – przeżyje, 0 – nie. W idealnym przypadku, nie tylko samo 0 czy 1, ale prawdopodobieństwo przeżycia, czyli jedynki – np 84%.

zero r algorytm

Większość pasażerów Titanica zginęła. Jeżeli spojrzymy na naszą listę pasażerów to dokładnie, 61 %. W związku, jeżeli dostaniemy dane nowego pasażera, możemy w 'ślepo' powiedzieć, że zginie i w około 61% przypadków, będziemy mieć rację.

Nie jest to najlepszy sposób predykcji, czy nowy pasażer zginie czy nie, ale ta wartość może pomóc nam się odnieść do jakości bardziej wyrafinowanych algorytmów, takich jak chociażby drzewo decyzyjne.

one r algorytm

Podejściem, które da nam lepsze efekty, jest podejście w którym wybieramy jeden parametr i dokonujemy predykcji na jego podstawie. Np płeć. Możemy powiedzieć, że jeżeli pasażer jest kobietą, to przeżyje i z 72% możemy spodziewać się, że będziemy mieli rację.

Znowu. Nie jest to jednak, najlepszy sposób predykcji, ale za to dobry wstęp do drzew decyzyjnych.

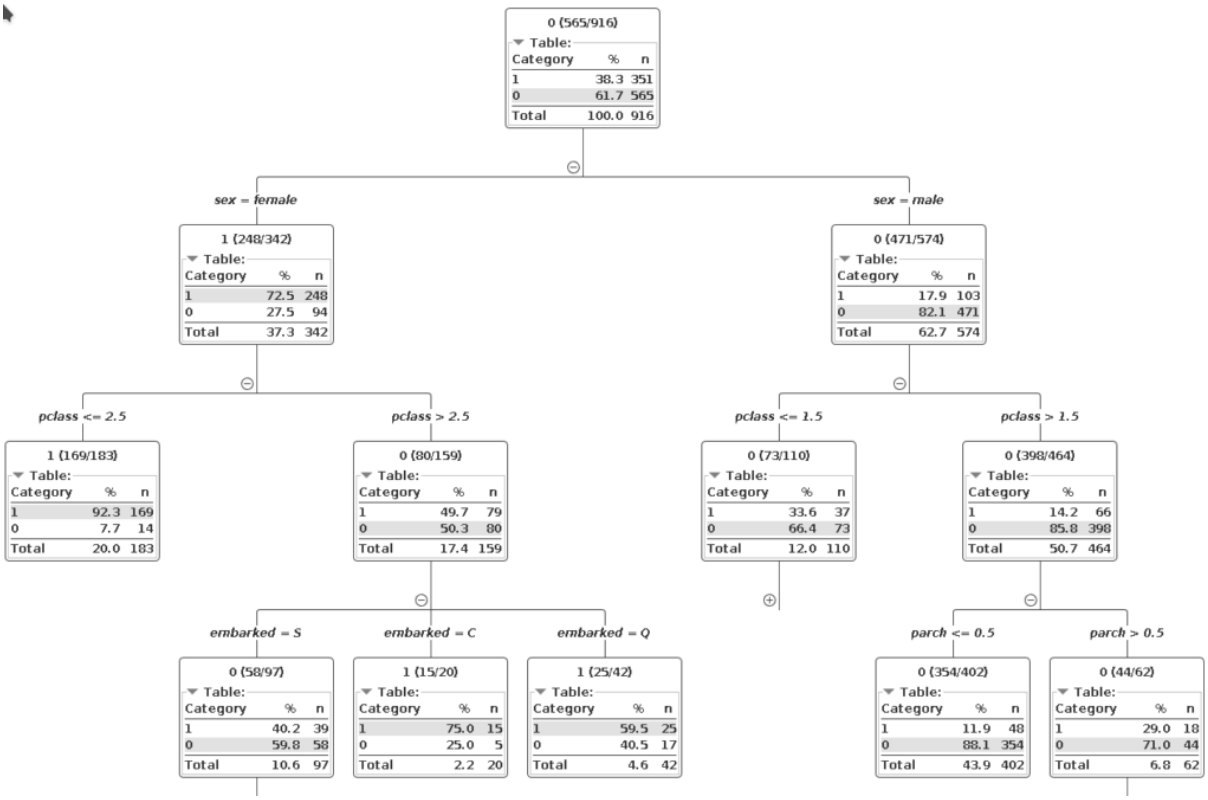
Czym są drzewa decyzyjne

Jest to prosta, lecz użyteczna w uczeniu maszynowym, koncepcja opierająca się o drzewo. Bardzo często stosowana do klasyfikacji, czyli przypisanie obserwacji zbioru danych do jednej z klas.

Na każdym węźle drzewa, dokonujemy podziały zbioru na 2 lub więcej mniejsze zbiory, mając na celu, jak najlepsze odseparowanie obserwacji należących do różnych klas.

Jak wygląda drzewo decyzyjne

Jeżeli dokonamy szybkiej analizy zbioru Titanic, za pomocą drzewa decyzyjnego, możemy otrzymać poniższy rezultat:



Pierwszą zmienną, którą algorytm wybrał do podziału zbioru to płeć. Widzimy że w przypadku kobiet, szanse na przeżycie wynosiły 72%

Następne, zarówno w przypadku kobiet, jak i mężczyzn, następną zmienną po dalszych podziałów zbiorów, mamy klasę biletu

A następnie, w przypadku gałęzi, dla kobiet, oraz klasy ekonomicznej, znaczenie miał port w którym osoba wsiadła na pokład. Możemy się domyślać, że zdecydowało to o odległości od kajut ratunkowych. Natomiast w przypadku gałęzi z mężczyznami, ilość dzieci oraz rodziców. Możemy się domyślać, że jest to skorelowane z wiekiem. Jeżeli mężczyzna, miał rodziców, to znaczy o jego młodym wieku i potencjalnym pierwszeństwie w zajęciu miejsca w kajucie ratunkowej.

Jak buduje się drzewo decyzyjne

Drzewo decyzyjne, dąży do wyboru takiego parametru podziału, oraz takich wartości podziału, aby w konsekwencji otrzymać 'najlepszy' podział.

'Najlepszy', mierzy się często za pomocą Gini Index lub Information Gain, opartym na entropii. Są to miary sprawdzające czystość podziału, czyli czy pozyskane poprzez podział liście drzewa, dają nam

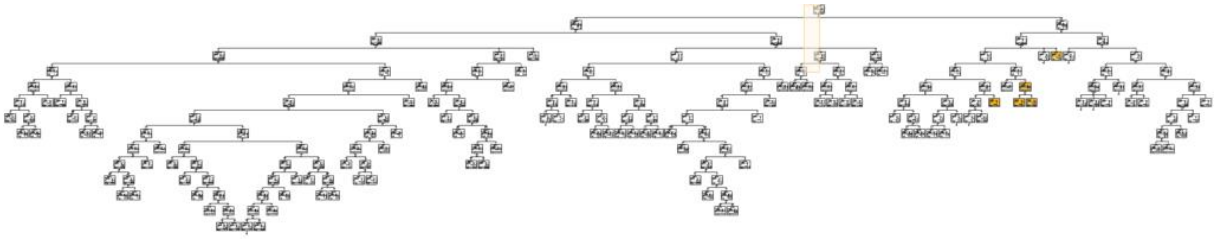
wartość. W idealnym przypadku, w wyniku podziału zbioru, za pomocą wybranego parametru, jeden liść zawierał by tylko i wyłącznie pasażerów którzy przeżyli, a drugi, tych którzy NIE przeżyli.

Następnie ponownie wybieramy parametr, oraz dokonujemy ponownego podziału, aby otrzymać jeszcze lepszy wynik.

Drzewo decyzyjne, możemy budować do momenty, w którym w każdym liść będzie w 100% czysty, czyli będzie zawierać obserwacje tylko i wyłącznie jednak klasy (przeżył lub nie).

W tedy jednak, drzewo będzie przeuczone (overfitted),

Poniżej, znajduje się drzewo decyzyjne, budowane do momentu, aż liście będą zawierać obserwację tylko jednej klasy:



Co oznacza, że osiągnęliśmy 100% skuteczność klasyfikacji czy ktoś zginie czy przeżyje, ale na zbiorze treningowym. Jeżeli użyli byśmy tego drzewa na nowych danych, to wynik był by prawdopodobnie gorszy. Drzewo decyzyjne jest przeuczone, a my dopatrzyliśmy się zależności które występują w zbiorze treningowym, ale prawdopodobnie po za nim, już nie.

Kiedy zatrzymujemy budowę drzewa decyzyjnego

W praktyce, budowę drzewa decyzyjnego, możemy zatrzymać w momencie kiedy:

Zbiór osiągnął minimalną liczbę obserwacji, np w liściu znajduje się mniej niż 5% wszystkich obserwacji

Nie mamy parametru, za pomocą którego, dalszy podział dałby nam lepszy wynik

Zbiór jest 'czysty', czyli zawiera obserwacje tylko jednego typu

Alternatywą jest 'pruning', czyli przycinanie. W tym podejściu, nie zatrzymujemy budowy drzewa, aby móc przyjrzeć się wszystkim znalezionym zależnościom, a następne eliminujemy te które uznamy za przeuczenie.

Reprezentacja drzewa decyzyjnego

Niewątpliwą zaletą drzew decyzyjnych, jest ich prosta reprezentacja w formie reguł if. Każdy liść, możemy zapisać, w przykładowy sposób:

if param1 > X and param2 < Y then probability = Z

W naszym przypadku, możemy powyższe drzewo, zapisać w postaci kilku reguł, wyglądających następująco:

Condition	Outco...	Probability 1
{$parch$ <= 0.5 AND $pclass$ <= 1.5 AND $embarked$ = "S" AND $pclass$ <= 2.5 AND sex = "female"}	1	0.967
{$parch$ > 0.5 AND $pclass$ <= 1.5 AND $embarked$ = "S" AND $pclass$ <= 2.5 AND sex = "female"}	1	0.889
{age <= 30.5 AND $parch$ <= 0.5 AND $pclass$ > 1.5 AND $embarked$ = "S" AND $pclass$ <= 2.5 A...	1	0.811
{age > 30.5 AND $parch$ <= 0.5 AND $pclass$ > 1.5 AND $embarked$ = "S" AND $pclass$ <= 2.5 AN...	1	0.871
{age <= 24.5 AND $parch$ > 0.5 AND $pclass$ > 1.5 AND $embarked$ = "S" AND $pclass$ <= 2.5 AN...	1	0.941
{age > 24.5 AND $parch$ > 0.5 AND $pclass$ > 1.5 AND $embarked$ = "S" AND $pclass$ <= 2.5 AND ...	1	0.933
{age <= 35.5 AND $parch$ <= 0.5 AND $embarked$ = "C" AND $pclass$ <= 2.5 AND sex = "female"}	1	1
{age > 35.5 AND $parch$ <= 0.5 AND $embarked$ = "C" AND $pclass$ <= 2.5 AND sex = "female"}	1	0.887
{$parch$ > 0.5 AND $embarked$ = "C" AND $pclass$ <= 2.5 AND sex = "female"}	1	1
{$embarked$ = "Q" AND $pclass$ <= 2.5 AND sex = "female"}	1	1
{age <= 25.5 AND $parch$ <= 0.5 AND $parch$ <= 1.5 AND $embarked$ = "S" AND $pclass$ > 2.5 AN...	0	0.451
{age > 25.5 AND $parch$ <= 0.5 AND $parch$ <= 1.5 AND $embarked$ = "S" AND $pclass$ > 2.5 AND...	1	0.51
{$parch$ > 0.5 AND $parch$ <= 1.5 AND $embarked$ = "S" AND $pclass$ > 2.5 AND sex = "female"}	1	0.533
{$parch$ > 1.5 AND $embarked$ = "S" AND $pclass$ > 2.5 AND sex = "female"}	0	0.226
{$embarked$ = "C" AND $pclass$ > 2.5 AND sex = "female"}	1	0.75
{$embarked$ = "Q" AND $pclass$ > 2.5 AND sex = "female"}	1	0.595
{age <= 39.0 AND $embarked$ = "S" AND $pclass$ <= 1.5 AND sex = "male"}	0	0.383

Jeżeli pewnej obserwacji, dana reguła jest spełniona, zostaje jej przypisane prawdopodobieństwo, że należy do klasy o numerze 1. W naszym przypadku, oznacza to prawdopodobieństwo że osoba przeżyje.

Podsumowując

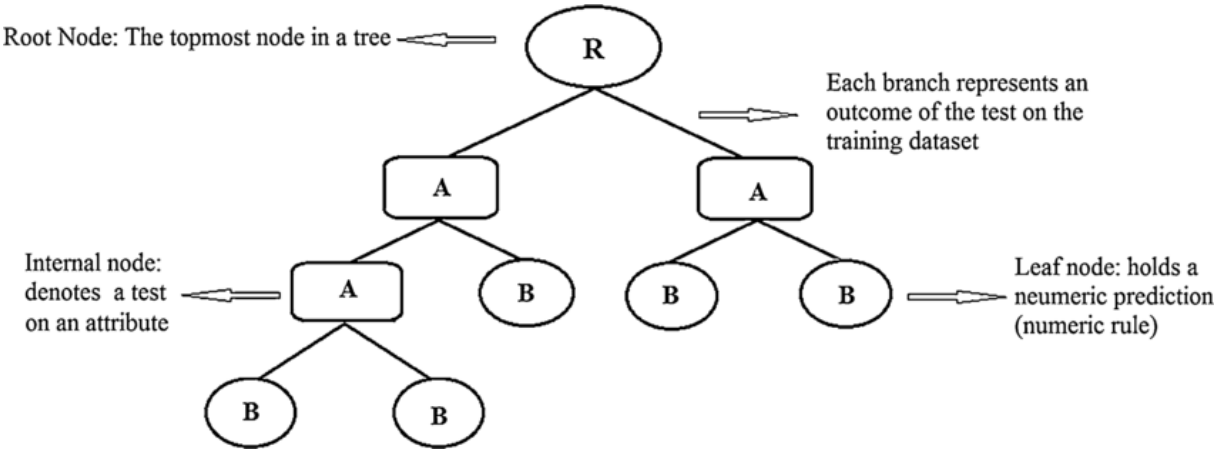
Najważniejszą zaletą drzewa decyzyjnego, jest jego prostota, która powoduje że jest to algorytm, którego działanie łatwo jest wytłumaczyć, a tym samym w łatwy sposób uzyskuje aprobatę i zgodę do implementacji w firmach.

jakość jego działania, oceniamy tak samo jak w przypadku innych algorytmów klasyfikacyjnych. Możemy użyć confusion matrix, wykresów ROC, Lift czy też profit curve.

Jest to ciągle bardzo popularna, obok regresji logistycznej, czy też SVM metoda klasyfikacji oraz podstawa budowy bardziej złożonych algorytmów takich jak lasy losowe.

Uczenie maszynowe a drzewa decyzyjne

Drzewo decyzyjne składa się z trzech typów węzłów: węzłów decyzyjnych, węzłów końcowych i węzłów przypadkowych. Węzły losowe reprezentują okrąg – podkreślają prawdopodobieństwo konkretnego wyniku. Kwadratowy kształt reprezentuje węzeł decyzyjny – wskazuje na wybór, którego musisz dokonać. Na koniec, węzeł końcowy reprezentuje wynik decyzji.



Analiza Przykład drzewa decyzyjnego

Możesz zmniejszyć ryzyko i zmaksymalizować szanse na osiągnięcie pożądanych rezultatów, obliczając przewidywaną wartość lub użyteczność każdego wyboru na drzewie. Jeśli chcesz obliczyć oczekiwaną użyteczność danego wyboru, odejmij ten koszt decyzji od oczekiwanych korzyści. Oczekiwane korzyści są proporcjonalne do ogólnej wartości każdego wyniku, który może wystąpić z danej opcji.

Kiedy próbujesz znaleźć pożądany rezultat, ważne jest, aby wziąć pod uwagę preferencje osoby podejmującej decyzję dotyczące użyteczności. Na przykład, niektórzy są gotowi podjąć ryzyko, aby uzyskać znaczne korzyści, podczas gdy inni chcą podjąć najmniejsze ryzyko.

Tak więc, gdy używasz drzewa decyzyjnego z jego modelem prawdopodobieństwa, może ono być przydatne do obliczania warunkowego prawdopodobieństwa zdarzenia. Może również by¢ idealny do określania, czy będzie on oparty na innych zdarzeniach. Dlatego musisz zacząć od początkowej parzystości i podążać jego ścieżką do docelowego zdarzenia. Następnie pomnóż razem prawdopodobieństwo każdego zdarzenia, aby otrzymać wyniki.

W takich przypadkach możesz użyć drzewa decyzyjnego w postaci konwencjonalnego schematu drzewa, który mapuje prawdopodobieństwo różnych zdarzeń, takich jak dwukrotne przetoczenie kostki.

Zrozumienie Algorytmu Drzewa Decyzyjnego

Algorytm drzewa decyzyjnego w pythonie należy do grupy nadzorowanych algorytmów. Ponadto, w przeciwieństwie do większości nadzorowanych algorytmów uczenia się, algorytm drzewa decyzyjnego może być używany do rozwiązywania problemów klasyfikacji i regresji.

Po raz kolejny, podstawowym celem drzewa decyzyjnego przy opracowywaniu modelu treningu jest przewidywanie wartości lub klasy celu poprzez zrozumienie podstawowych reguł decyzyjnych zaczerpniętych ze starszych danych, które programiści nazywają również danymi treningowymi.

Zacznij od korzenia drzewa, gdy próbujesz przewidzieć etykietę klasy rekordu i porównaj wartość korzenia atrybutu z charakterystyką rekordu. Jeśli chodzi o porównanie, postępuj zgodnie z gałęzią odpowiadającą jego wartości, po czym możesz przejść do drugiego węzła.

Ile jest rodzajów drzew decyzyjnych?

Typy drzew decyzyjnych zależą od zmiennych docelowych. Istnieją dwa rodzaje drzew decyzyjnych:

- **Drzewo decyzyjne o zmiennej ciągłej**
- **Drzewo decyzyjne o zmiennej kategorii**

Na przykład, musimy przewidzieć, czy ktoś zwróci składkę na odnowienie za pośrednictwem firmy ubezpieczeniowej. W tym scenariuszu wiemy, że dochód klienta jest ogromną zmienną.

Jednak usługa ubezpieczeniowa nie posiada wszystkich szczegółów dotyczących klienta. Większość z Państwa będzie wiedziała, że ta zmienna jest krytyczna. Dlatego też możemy opracować drzewo decyzyjne do przewidywania dochodów klienta poprzez inne zmienne, takie jak produkty i zawód. W większości przypadków będziemy spekulować wartościami dla zmiennych ciągłych.

Jakie są plusy i minusy drzewa decyzyjnego?

Mocne strony

- Drzewa decyzyjne oferują jasne wyobrażenie o krytycznych polach do klasyfikacji lub przewidywania
- Drzewo decyzyjne jest zdolne do obsługiwania zmiennych kategorycznych i ciągłych
- Nie wymagają one nadmiernych obliczeń przy dokonywaniu klasyfikacji
- Drzewa te mogą generować łatwo zrozumiałe zasady

Słabe strony

- Błędy są dość powszechne w drzewach decyzyjnych, szczególnie jeśli chodzi o problemy z klasyfikacją i przykłady szkoleń
- Drzewa decyzyjne nie są idealnym rozwiązaniem w przypadku tworzenia zadań szacunkowych do przewidywania wartości ciągłego atrybutu
  Szkolenie drzewa decyzyjnego może być dość kosztowne obliczeniowo.
- Musisz posortować pole plucia każdego kandydata na węzeł, aby określić najbardziej korzystny podział. Niektóre algorytmy wykorzystują kombinacje, które wymagają kompleksowego wyszukiwania w celu określenia odpowiednich wag łączących.
- Przycinanie algorytmów jest dość kosztowne, głównie dlatego, że trzeba porównywać i formować podrzędy.

Podstawowa terminologia drzew decyzyjnych

Węzły dziecięce i rodzicielskie

Każdy węzeł, który dzieli się na podwęzły jest również znany jako węzeł nadrzędny. Węzły podrzędne są natomiast węzłami dziecięcymi.

Drzewo podrzędne/oddział

Podsekcja drzewa decyzyjnego to jego poddrzewo lub gałąź.

Przycinanie

Przycinanie jest procesem, w którym redukujesz rozmiar drzewa decyzyjnego poprzez oskubywanie jego węzłów.

Węzeł terminala / liść

Węzły Leaf lub Terminal nie mają dzieci i nie przechodzą przez dodatkowe podziały.

Węzeł decyzyjny

Gdy jeden podwęzeł dzieli się na wiele węzłów, staje się on węzłem decyzyjnym.

Podział na

Podział jest procesem, który dzieli jeden węzeł na wiele podwęzłów.

Węzeł bazowy

Węzeł korzeniowy reprezentuje całą próbę lub populację każdego węzła. Dzieli się on dalej na wiele jednorodnych zbiorów.

Myśli końcowe

Opracowanie drzewa decyzyjnego w pytonie może rozwiązać wiele kwestii związanych z decyzjami dla dużych i mniejszych organizacji. Może również pomóc jednostkom zdecydować, czy wybór, którego mają zamiar dokonać, będzie opłacalny. Deweloperzy często korzystają z biblioteki sclearnowej pythona, aby stworzyć drzewo decyzyjne sclearn. Jego implementacja i algorytm są bardziej wydajne i dają lepsze wyniki.