SKN DSM Learning Group

Adrianna Wołowiec, 30.10.2018 r.

Zanim przejdziemy do modelowania…

Podział zbioru danych

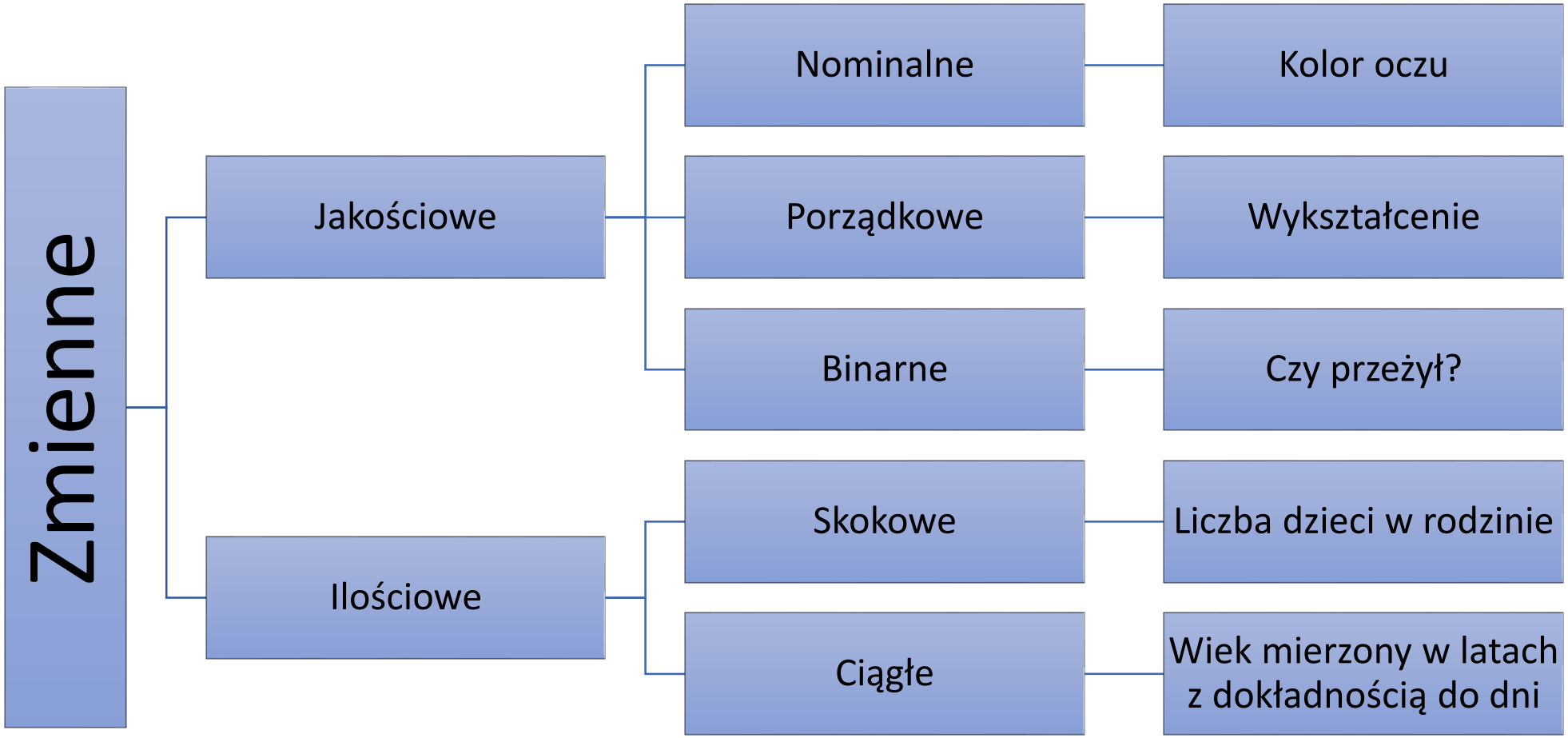
Zbiór uczący (ang. *train set*) - służy do oszacowania konkurujących modeli, ich parametrów

Zbiór walidacyjny (ang. *validation set*) - służy do wyboru jednego z oszacowanych klasyfikatorów - tego, który ma najmniejszy błąd na zbiorze walidacyjnym

Zbiór testowy (ang. *test set*) - służy do nieobciążonej oceny błędu

Przykładowy podział: 70/15/15, 80/20

Typy zmiennych



Wizualizacja danych

https://www.tableau.com/en-nz/learn/whitepapers/which-chart-or-graph-is-right-for-you

Kodowanie zmiennych jakościowych

Kodowanie zmiennych jakościowych sztucznymi zmiennymi zero-jedynkowymi (ang. *dummy coding, dummy variables)*

Zmienna jakościowa o k poziomach wartości jest zastępowana przez k-1 sztucznych zmiennych zero-jedynkowych.

Wartość zmiennej jakościowej, dla której wszystkie sztuczne zmienne przyjmują wartość 0, staje się poziomem referencyjnym.

Np. zmienna płeć o wartościach „K” i „M” będzie reprezentowana przez jedną zmienną sztuczną przyjmującą wartość 1 dla obserwacji, gdzie płeć = „K” i 0, gdy płeć = „M”

Kodowanie zmiennych jakościowych c.d.

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			

Ziarno generatora liczb pseudolosowych

Generowane przez komputer liczby nazywane są pseudolosowymi, ponieważ mają emulować losowość, ale są wyznaczane w deterministyczny, choć często bardzo skomplikowany, sposób.

Generator to funkcja deterministyczna. Do losowania kolejnych liczb wykorzystuje tzw. ziarno (ang. *seed*), całkowicie determinujące wartości kolejnych liczb pseudolosowych. Ziarno to wartość na podstawie której konstruowane będą kolejne liczby losowe.

Dla ustalonego generatora i ziarna generowane będą identyczne liczby losowe bez względu na system operacyjny, nazwę komputera, rasę użytkownika, czy temperaturę w pokoju.

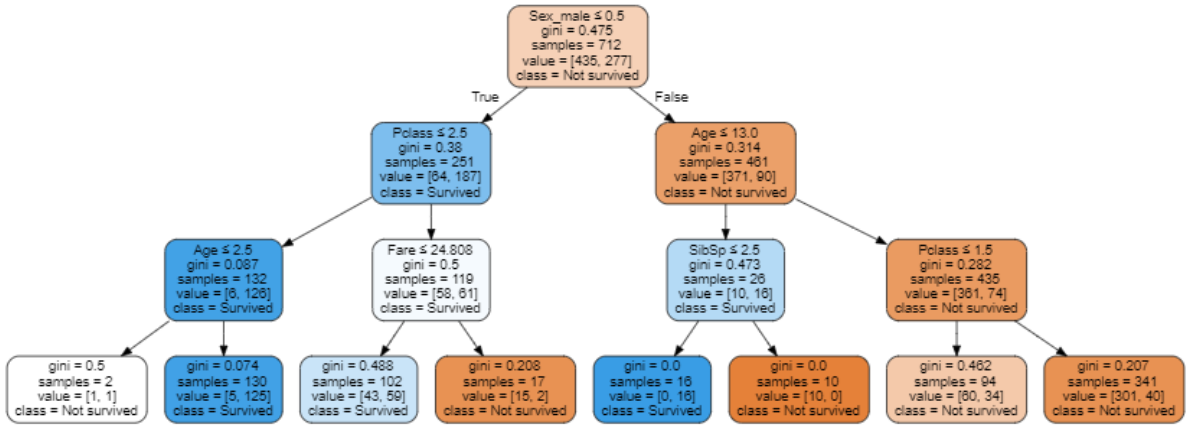
Sterując wyborem ziarna umożliwiamy otrzymywanie identycznych ciągów liczb losowych na różnych komputerach. W ten sposób możemy powtarzać wyniki symulacji, odtwarzać te wyniki na innych komputerach lub kontynuować obliczenia przerwane w wyniku wystąpienia jakiegoś błędu.

Drzewa decyzyjne - podstawy

Model drzewa decyzyjnego

1. Model drzewa decyzyjnego jest jednym z algorytmów uczenia nadzorowanego (czyli ze zdefiniowaną funkcją celu). Można w nim stosować zarówno jakościowe jak i ilościowe zmienne objaśniające. Technika modelu polega na podziale zbioru na dwa lub więcej podzbiory w oparciu o zmienną wejściową.
2. Klasyfikator ten ma strukturę drzewa, które może być przedstawiane graficznie jako zbiór reguł decyzyjnych.
3. Drzewo decyzyjne może być stosowane zarówno w przypadku problemu klasyfikacyjnego (drzewo klasyfikacyjne), jak i problemy regresyjnego (drzewo regresyjne).

Struktura drzewa decyzyjnego



- **Korzeń:** Węzeł będący podstawą drzewa, w którym dokonuje się pierwszy podział zbioru.
- **Podział:** Proces podziału węzła-rodzica na dwa lub więcej węzłów-dzieci
- **Węzeł decyzyjny:** Reprezentuje podzbiór danych. Jego podstawową charakterystyką jest jednorodność.
- **Gałęzie:** Krawędzie łączące węzły. Reprezentują reguły klasyfikacyjne.
- **Liść:** Węzeł końcowy, który nie podlega podziałowi. Liście reprezentują klasy zmiennej celu (w przypadku drzewa klasyfikacyjnego).

Podział węzłów

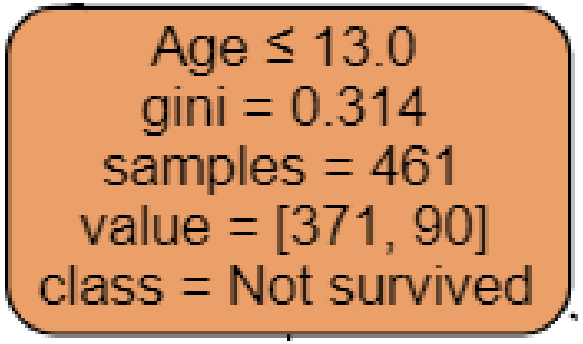
Celem algorytmu uczenia jest znalezienie najlepszego podziału dla każdego węzła w drzewie. Najlepszy to znaczy zapewniający jak największą jednorodność pod względem klas zmiennej celu w podzbiorach utworzonych na podstawie podziału.

Miarami pozwalającymi ocenić podział węzła są przyrost informacji (ang. *information gain)* oraz indeks/wskaźnik Gini'ego (ang. *gini impurity)*.

Przykład podziału

Zmienna A	Zmienna B	Zmienna celu
K	High	1
K	Low	1
M	High	0
K	Low	1
M	Low	0
M	High	0

Schemat drzewa losowego



- Warunek oparty na wartości zmiennej. Odpowiedzią są wartości tak lub nie (prawda lub fałsz), które wskazują czy warunek jest spełniony. W oparciu o spełnienie lub nie warunku obserwacje są przesuwane w głąb drzewa.
- gini - wartość indeksu Giniego w węźle. Wartość ta maleje wraz z kolejnymi poziomami drzewa
- samples – liczba obserwacji w węźle
- value - liczba obserwacji w węźle z podziałem na klasy zmiennej celu. Na podanym przykładzie, w węźle 371 obserwacji należy do klasy 0, a 90 – do klasy 1.
- class - klasa zmiennej celu, która przeważa wśród obserwacji znajdujących się w węźle. W przypadku węzłów, które są liśćmi, oznacza predykcję dla wszystkich obserwacji w liściu.

Dostrajanie hiperparametrów

Ang. *Hyperparameter tuning*

Parametry modelu uzyskuje się podczas procesu uczenia przy założeniu minimalizacji funkcji straty

Przykład: w przypadku regresji liniowej współczynniki *beta* stojące przy zmiennych objaśniających są estymowane przy założeniu minimalizacji sumy kwadratów błędów modelu

Hiperparametry nie są parametrami modelu i nie są ustalane podczas uczenia się modelu na danych.

Dostrajanie hiperparametrów

Jednym z parametrów, które wymagają „tuningu”, jest maksymalna głębokość drzewa (*max_depth*). Wskazuje ona, jak głębokie może być budowane drzewo. Im głębsze drzewo, tym więcej zawiera podziałów i ujmuje więcej informacji ze zbioru. Zbyt duża głębokość prowadzi jednak do przeuczenia modelu.

Inne przykłady: *min_samples_split* (minimalna liczba obserwacji w węźle umożliwiająca dokonanie podziału), *min_samples_leaf* (minimalna liczba obserwacji w liściu)

Przeuczenie modelu

Ang. *overfitting*

Model jest przeuczony, gdy jest zbyt dobrze dopasowany do danych uczących, przez co utracił zdolność do uogólniania. Taki model wyłapuje nie tylko rzeczywiste zależności pomiędzy zmiennymi, ale także szum występujący w zbiorze.

Jak rozpoznać przeuczenie? – Bardzo dobra jakość dopasowania na zbiorze uczącym i słabe wyniki na zbiorze testowym.

Variance-bias tradeoff

Las losowy

Random forest

Model lasu losowego składa się z wielu drzew decyzyjnych (wiele drzew = las). Model wykorzystuje dwie kluczowe koncepcje, które zapewniają element losowości w drzewach.

1. Przy budowie drzew wykorzystywana jest losowa próba obserwacji ze zbioru danych. Losowanie to odbywa się ze zwracaniem (ang. *bootstrapping*).
2. Podczas dzielenia węzłów brany jest pod uwagę losowy podzbiór zmiennych w zbiorze

Las losowy danej obserwacji przypisuje taką wartość zmiennej celu, która najczęściej występowała dla tej obserwacji w wynikach utworzonych drzew.

Gini importance / MeanDecreaseGini

Na podstawie drzew z lasu można wyznaczyć ranking zmiennych tzn. dokonać oceny istotności zmiennych poprzez obliczenie średniej zmiany indeksu Giniego dla każdej zmiennej, a tym samym określić, które zmienne mają lepsze właściwości predykcyjne.

Proces ten polega na obliczeniu różnicy między różnorodnością klas w węźle-rodzicu i węzłach dzieciach dla danej zmiennej, dla każdego drzewa z lasu. Następnie wszystkie te wartości są sumowane. Wyznaczając taką średnią zmianę indeksu dla każdej ze zmiennych otrzymamy ranking zmiennych według ich własności predykcyjnych.