Drzewa klasyfikacyjno-regresyjne (CRT) i lasy losowe

Joanna Karłowska-Pik

11-12 lutego 2020

Klasyfikacja

Zadanie klasyfikacji polega na kojarzeniu obiektów na podstawie ich charakterystycznych cech (zmiennych opisujących, predyktorów) z odpowiednią wartością pewnej jakościowej zmiennej celu, którą nazywamy kategorią lub klasą.

Przykłady

- Bank na podstawie informacji o kliencie decyduje, czy udzielenie kredytu ma duże, średnie czy małe ryzyko.
- Program pocztowy decyduje, czy e-mail to spam.
- Lekarz na podstawie objawów stwierdza, czy pacjent cierpi na pewną chorobę.
- Służby specjalne na podstawie pewnych zachowań decydują, czy dana osoba stanowi zagrożenie dla bezpieczeństwa państwa.

Typ uczenia

Uczenie nadzorowane (ang. *supervised learning*) – dana jest zmienna celu i wiele przykładów ze znaną jej wartością, na podstawie których algorytm uczy się, które wartości zmiennej celu są powiązane z którymi wartościami zmiennych opisujących. Zbiór przykładów powinien być bogaty, różnorodny i powinien zawierać reprezentatywną grupę typów rekordów, których klasyfikacja będzie potrzebna w przyszłości. Zbyt ubogi zbiór może spowodować, że utworzony model klasyfikacji będzie zbyt prosty, tzn. algorytm będzie niedouczony i będzie miał problemy z poprawnym rozpoznawaniem nowych obserwacji.
Przeuczenie (ang. *overfitting*) – algorytm może utworzyć skomplikowany model, który bardzo dokładnie klasyfikuje elementy ze zbioru uczącego, a jego skuteczność w klasyfikacji nowych obserwacji będzie niska.

Podział na zbiory

- Zbiór uczący (ang. *training set*) – zbiór, w oparciu o który konstruowany jest dany model lub modele.
- Zbiór walidacyjny (ang. *validation set*) – zbiór, na którym dobieramy parametry modelu lub wybieramy najlepszy spośród kilku modeli; nie zawsze występuje.
- Zbiór testowy (ang. *testing set*) – zbiór służący do ostatecznej oceny jakości modelu.

Uwaga: Nie należy przywiązywać się do powyższych nazw. Nazwy zbiorów walidacyjnego i testowego bywają stosowane zamiennie. Trzeba raczej rozumieć istotę podziału na trzy zbiory i wiedzieć, do czego służą.

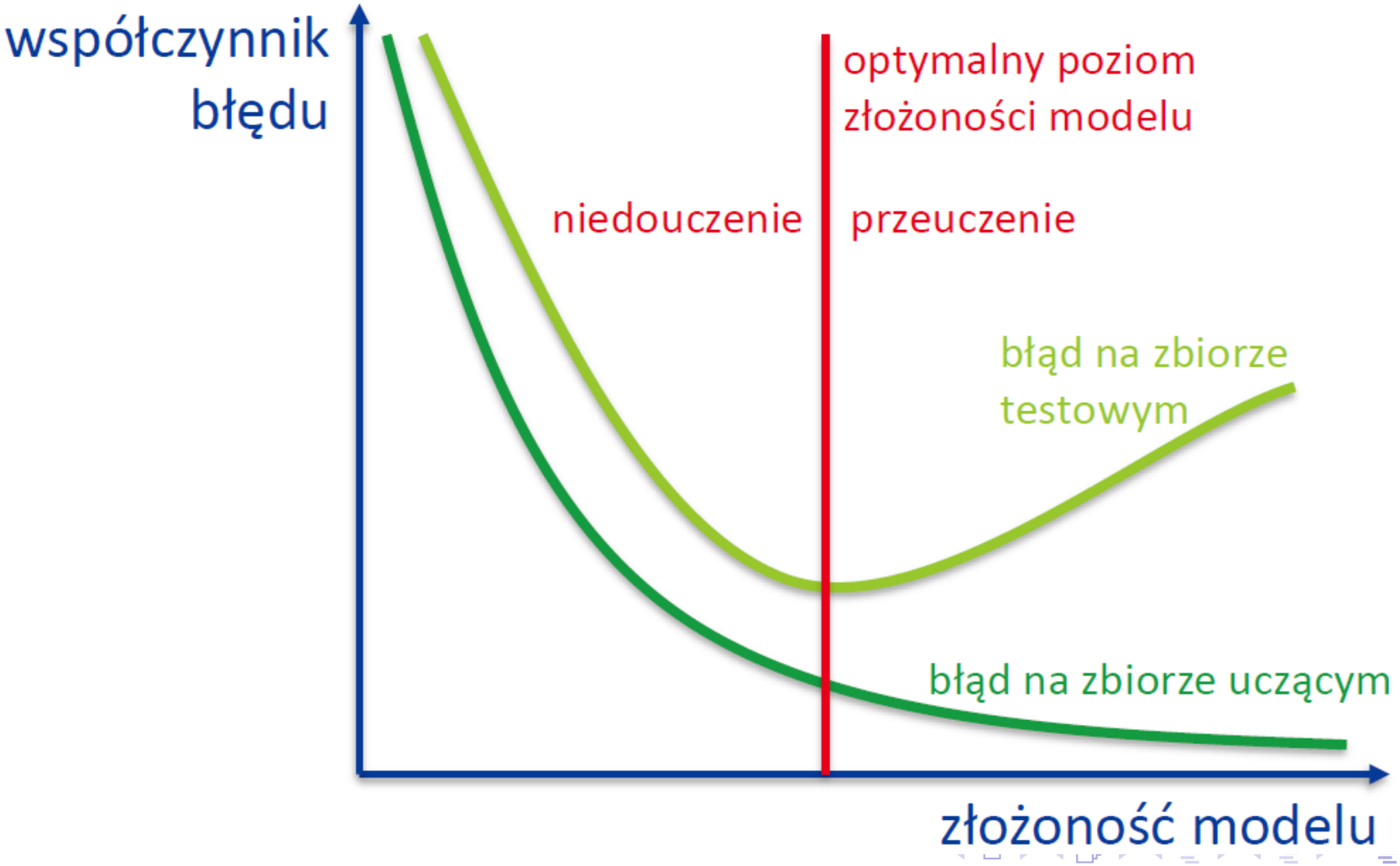
Metoda wydzielania

Metoda wydzielania (ang. *holdout method*): podział losowy na zbiór uczący i testowy w stosunku 2:1 (w praktyce ok. 70% – uczący, a 30% – testowy), a na uczący, walidacyjny i testowy w stosunku 2:1:1 (50%, 25% i 25%).

Skuteczna dla zbiorów z dużą liczbą obserwacji. Nie ma jednego ogólnego zalecenia do wielkości wyodrębnianych zbiorów.

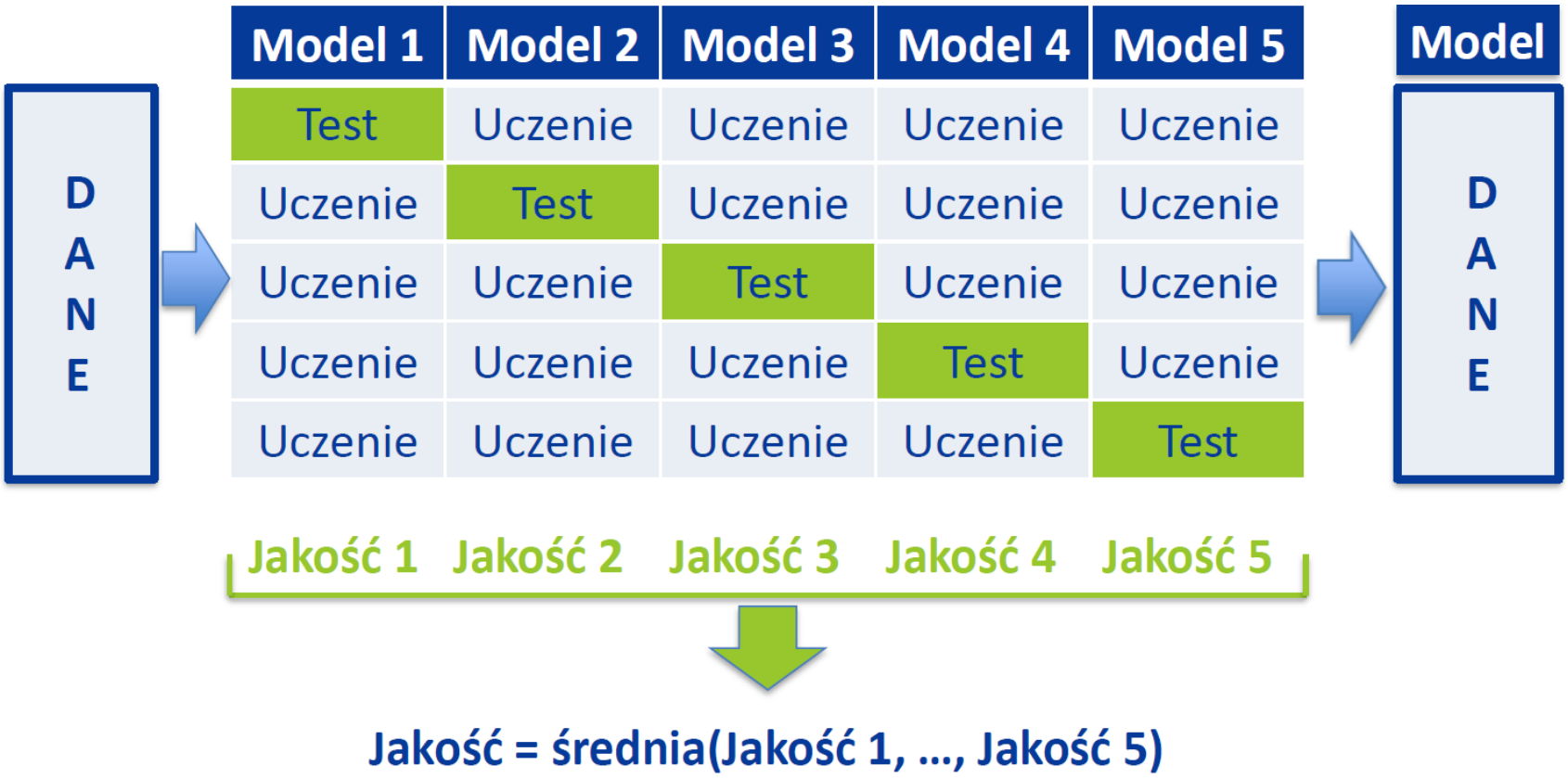
(Patrz: Ćwik, Koronacki (2008), str. 96)

Kompromis obciążeniowo-wariancyjny



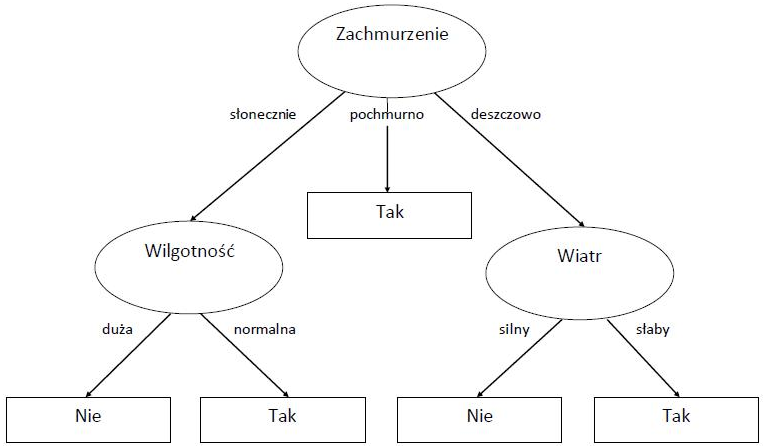
Sprawdzian krzyżowy

v-krotna walidacja krzyżowa (ang. *v-fold cross-validation*)



Drzewa decyzyjne

Drzewem decyzyjnym nazywamy zbiór węzłów decyzyjnych połączonych za pomocą gałęzi (ang. *branch*) rozchodzących się w dół od korzenia (ang. *root node*) aż do kończących je liści (ang. *leaf nodes* lub *terminal nodes*). W węzłach decyzyjnych sprawdzane są wartości zmiennych opisujących. Gałęzie wychodzące z węzłów odpowiadają możliwym wartościom bądź zbiorom wartości tych zmiennych. Każda gałąź prowadzi do kolejnego węzła decyzyjnego lub do liścia. W liściach znajdują się wartości zmiennej celu.



Działanie

- Algorytmy budujące drzewa dążą do sformułowania warunków, które są sprawdzane dla poszczególnych obserwacji.
  W zależności od wyników tego sprawdzania określana jest wartość zmiennej celu.
- Ponieważ w zbiorze uczącym mogą się znaleźć obserwacje, które mają takie same wartości zmiennych opisujących, a różne wartości zmiennej celu, liście mogą nie być „czyste”.
  W takim przypadku drzewo informuje z jakim prawdopodobieństwem są przyjmowane poszczególne wartości zmiennej celu.
- Algorytmy dążą do utworzenia jak najczystszych liści.

Historia

Leo Breiman (1928-2005) -– statystyk amerykański, twórca drzew CART oraz lasów losowych, współautor książki Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen: *Classification and Regression Trees* (1984)

Leo Breiman (1928-2005)

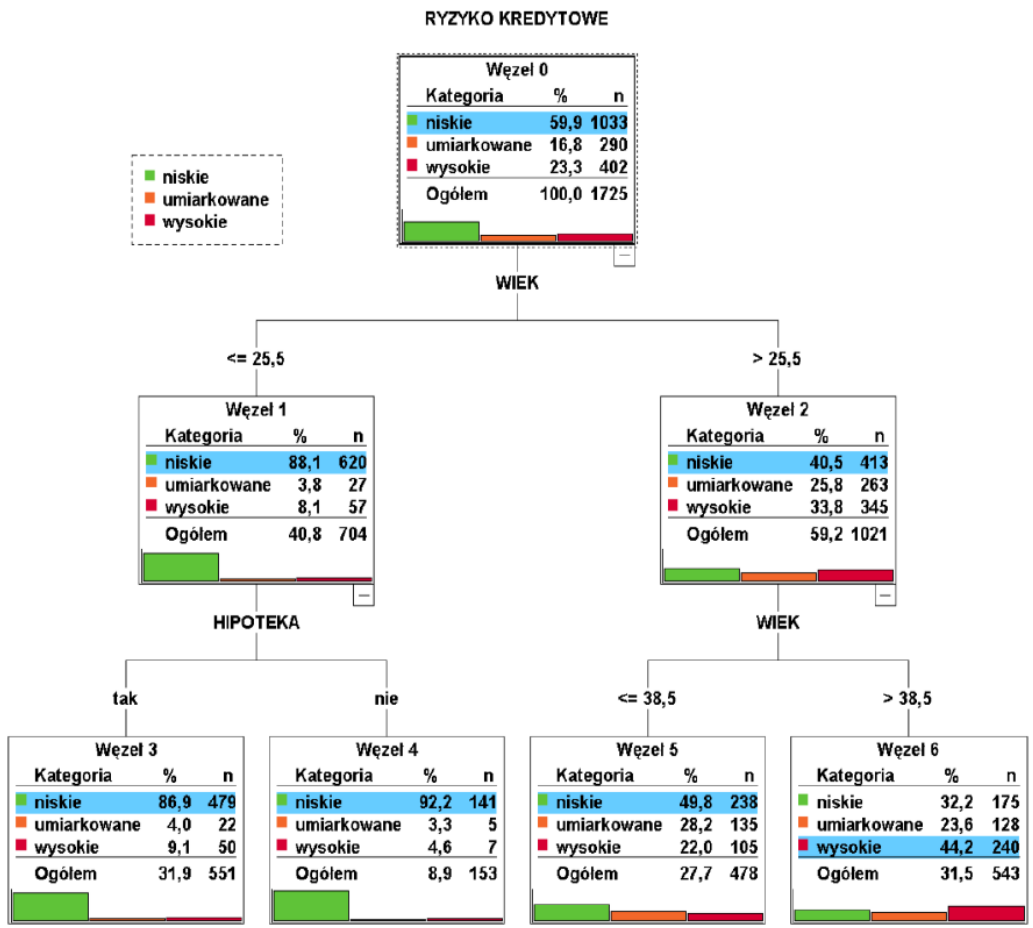
Źródło: http://en.wikipedia.org/wiki/Leo_Breiman

CART

Metoda drzew klasyfikacyjnych i regresyjnych (ang. *classification and regression trees*, ozn. CART - zastrzeżone, CRT lub C&RT):

- Drzewa decyzyjne tworzone przez algorytm CART są ściśle binarne, mają po 2 gałęzie wychodzące z każdego węzła decyzyjnego.
- Z węzła idziemy na lewo, jeśli jest spełniony określony w nim warunek, a na prawo, jeśli nie. Warunki mają postać „$X_i \leqslant C$", gdy zmienna jest ciągła lub „X_i przyjmuje wartości...", gdy zmienna jest dyskretna.
- Dopuszczalne jest kilkukrotne pojawienie się warunków bazujących na tej samej zmiennej.

Przykład drzewa CRT



Miara „nieczystości" węzłów – kryterium Giniego

Miarą nieczystości (ang. *impurity*) węzła jest prawdopodobieństwo, że dwie obserwacje wybrane losowo (ze zwracaniem) z węzła należą do dwóch różnych klas.
Dla zmiennej celu o K klasach od 0 do $K-1$

$$G(t) = \sum_{k=0}^{K-1} p_k(t)(1 - p_k(t)),$$

gdzie $p_k(t)$ oznacza częstość klasy k w węźle t.

Kryterium Giniego przy zastosowaniu kosztów błędnej klasyfikacji

Koszty błędnej klasyfikacji są stosowane w przypadkach, gdy zaklasyfikowanie przypadku z klasy j do klasy i jest gorsze (np. droższe) niż pomyłka w drugą stronę. Np. dla banku droższe może być udzielenie kredytu osobie, która nie ma zdolności kredytowej i kredytu nie spłaci, niż odrzucenie wniosku osoby posiadającej zdolność kredytową i utrata dochodu z tego tytułu.
Wtedy

$$G(t) = \sum_{i \neq j} c_{ij} p_i(t) p_j(t),$$

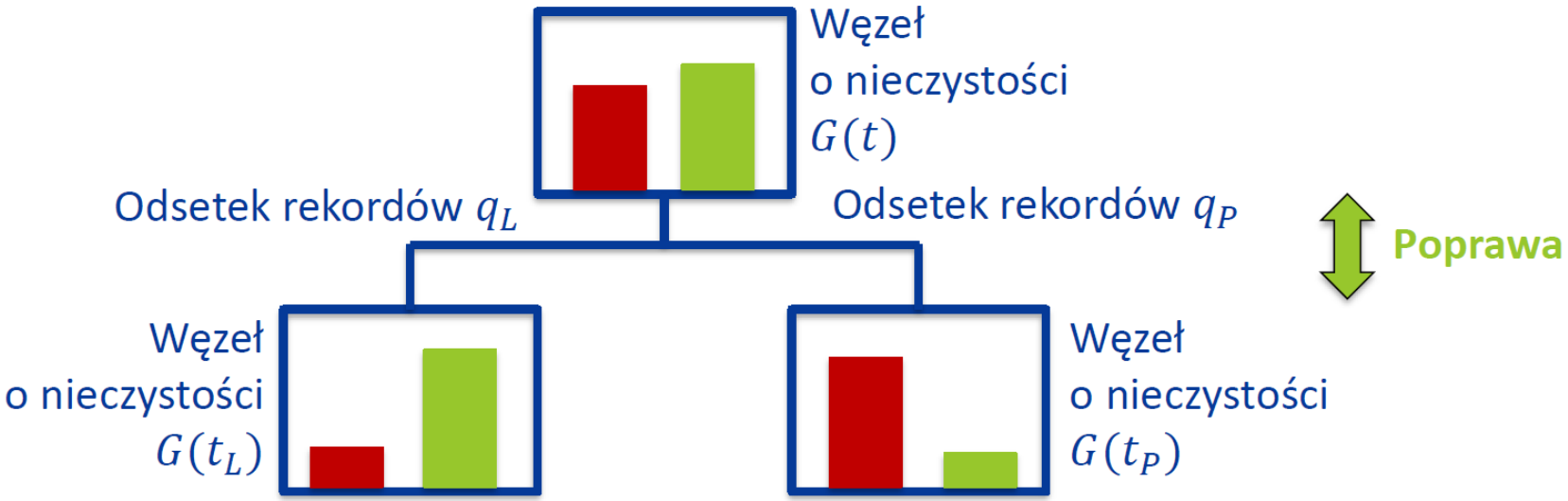
gdzie c_{ij} oznacza koszt błędnej klasyfikacji przypadku z klasy i jako przypadek z klasy j (domyślnie ustawione jako 1).

Miara poprawy

Poprawa (ang. *improvement*) wynikająca z podziału węzła rodzica t na lewy i prawy węzeł dzieci t_L i t_P wynosi

$$I(t) = G(t) - q_L G(t_L) - q_P G(t_P),$$

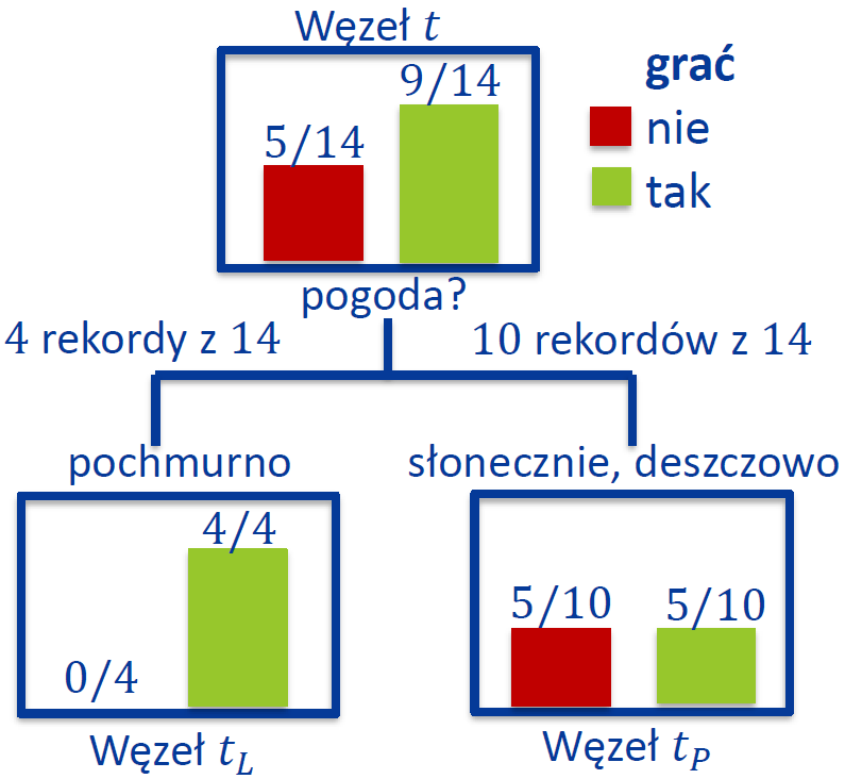
gdzie q_L (q_P) jest frakcją przypadków idących na lewo (prawo). Algorytm w każdym węźle przeszukuje dostępne zmienne i podziały i wybiera podział optymalny, maksymalizując miarę poprawy.



Przykład – gra w golfa

	pogoda	wiatr	temperatura	wilgotność	grać
1	słonecznie	nie	85	85	nie
2	słonecznie	tak	80	90	nie
3	pochmurno	nie	83	78	tak
4	deszczowo	nie	70	96	tak
5	deszczowo	nie	68	80	tak
6	deszczowo	tak	65	70	nie
7	pochmurno	tak	64	65	tak
8	słonecznie	nie	72	95	nie
9	słonecznie	nie	69	70	tak
10	deszczowo	nie	75	80	tak
11	słonecznie	tak	75	70	tak
12	pochmurno	tak	72	90	tak
13	pochmurno	nie	81	75	tak
14	deszczowo	tak	71	80	nie

Przykład – miara poprawy



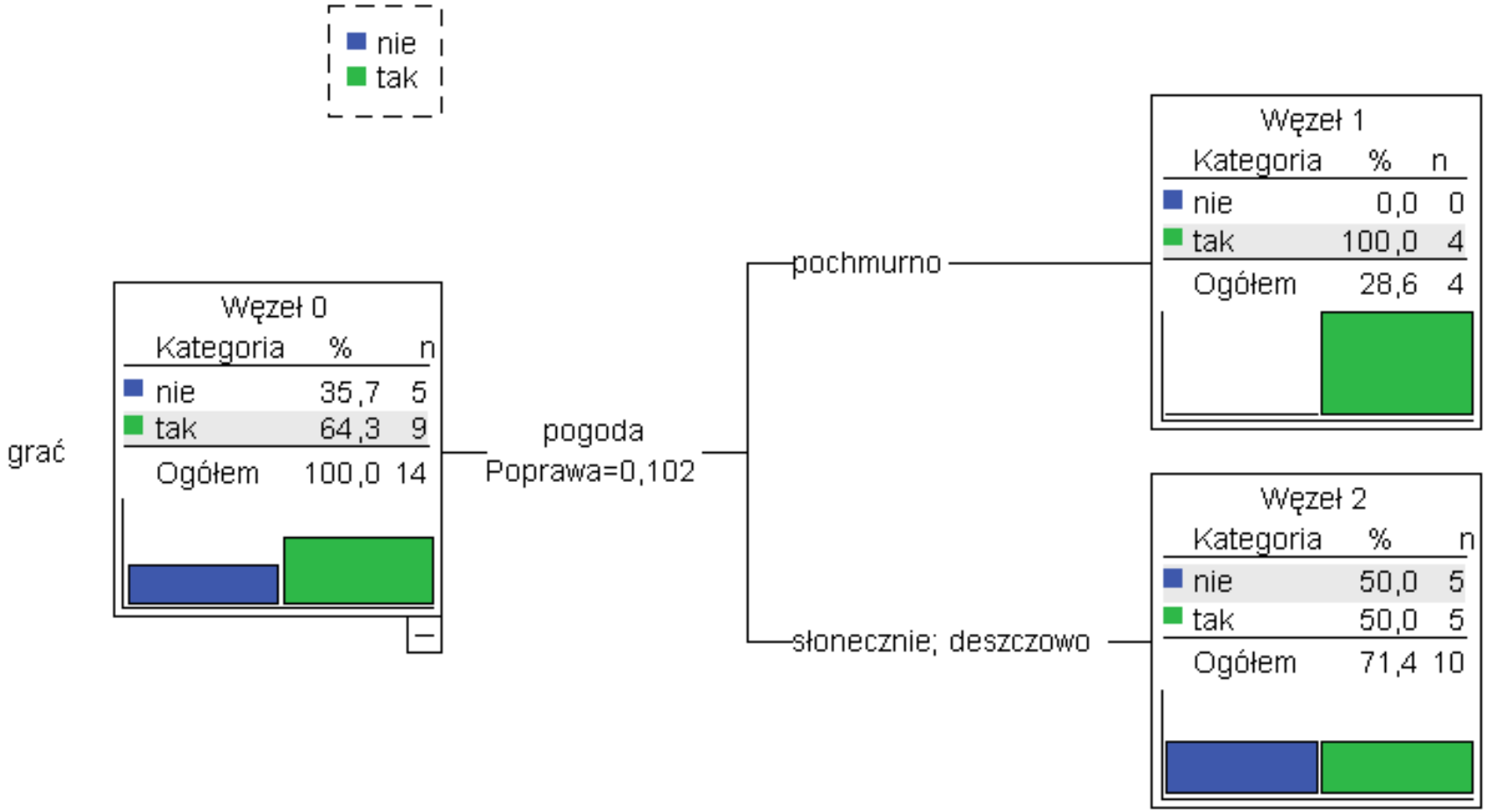
$$G(t) = \frac{9}{14} \cdot \frac{5}{14} + \frac{5}{14} \cdot \frac{9}{14} = \frac{45}{98}$$

$$G(t_L) = 0 \cdot 1 + 1 \cdot 0 = 0$$

$$G(t_P) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

$$I(t) = \frac{45}{98} - \frac{2}{7} \cdot 0 - \frac{5}{7} \cdot \frac{1}{2} = \\ = \frac{10}{98} \approx 0,102$$

Przykład – pierwszy podział



Kryterium *twoing*

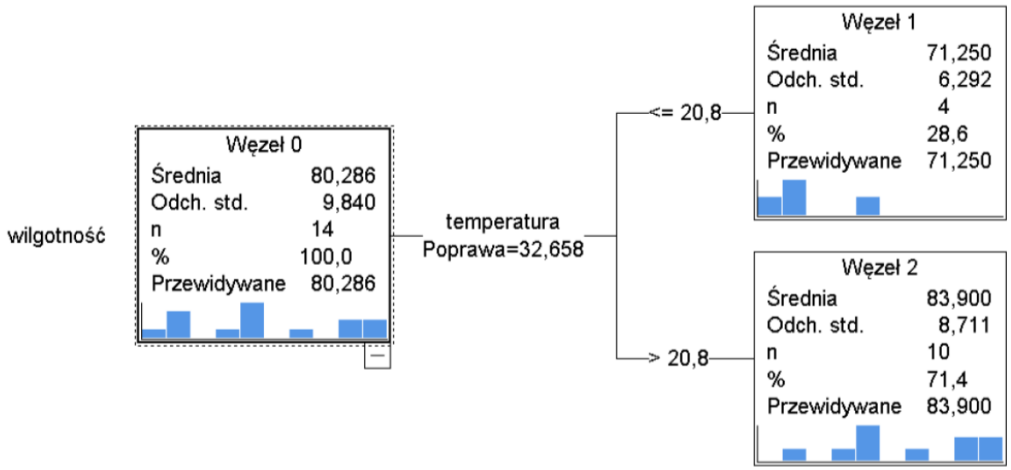
Poprawa jest definiowana jako

$$I(t) = q(1-q)\left(\sum_j |p_j(t_L) - p_j(t_P)|\right)^2$$

(istnieją drobne modyfikacje tego wzoru). W tym przypadku poprawa będzie największa, jeżeli wszystkie przypadki z tej samej klasy trafią do tego samego węzła. Dodatkowo czynnik stojący przed sumą powoduje, że poprawa jest największa dla $q = 0{,}5$. Preferowane są więc podziały, które są jednorodne dla wszystkich klas i mają w przybliżeniu równą liczbę rekordów.
Porządkowe kryterium *twoing* jest stosowane, gdy zmienna celu jest dyskretna i ma poziom porządkowy.

Ciągła zmienna celu

W przypadku **ilościowej zmiennej celu** w każdym węźle podsumowanie wartości zmiennej celu (dla przykładu: szacowanie wilgotności na podstawie temperatury, wiatru i pogody) polega na wyznaczeniu średniej i odchylenia standardowego oraz wykonaniu histogramu. Algorytm stara się odseparować obserwacje, dla których zmienna celu przyjmuje duże wartości, od obserwacji, dla których przyjmuje ona małe wartości.



Odchylenie najmniejszych kwadratów

W procesie budowy drzewa sprawdzana jest każda wartość każdego predyktora, by znaleźć taki predyktor i taką wartość progową, która podzieli rekordy na dwie grupy S_1 i S_2 tak, by zminimalizować sumę kwadratów błędów (odchylenie najmniejszych kwadratów – ang. *least squares deviation* (LSD))

$$SSE = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2,$$

gdzie $\bar{y}_1$ i $\bar{y}_2$ są średnimi wartościami zmiennej celu odpowiednio w grupach S_1 i S_2.

Zapobieganie przeuczeniu

Metody

- zatrzymywanie wzrostu w momencie, gdy wszystkie węzły mają niewielką liczebność ($\leqslant 5$),
- ograniczanie głębokości drzewa,
- ograniczanie liczby węzłów drzewa,
- automatyczne przycinanie (ang. *pruning*).

Przycinanie

Niech T_0 oznacza drzewo maksymalne, $|T|$ – liczbę liści drzewa T, $R(T)$ – odsetek błędnych klasyfikacji w liściach (czasem całkowita ważona nieczystość liści). Dla każdego α istnieje poddrzewo $T \subset T_0$ minimalizujące

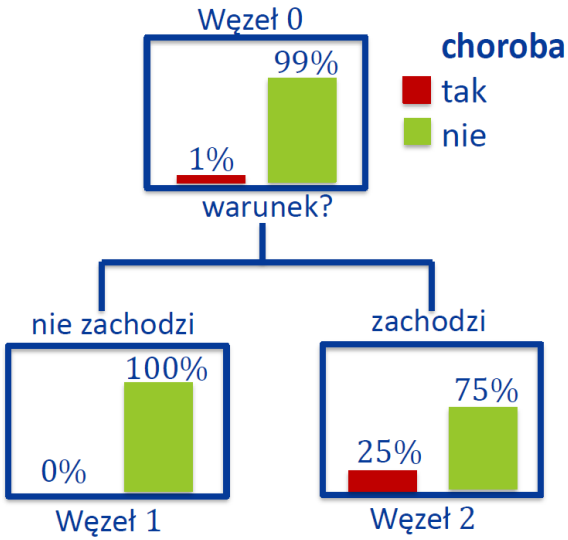
$$R_\alpha(T) = R(T) + \alpha|T|.$$

Do szacowania funkcji straty (wyboru α, dla którego powyższa suma jest najmniejsza) CART używa walidacji krzyżowej.

Reguła 1 SE

Breiman zaproponował regułę jednego błędu standardowego, czyli wybór najmniejszego drzewa z szacowanym błędem w granicach 1 odchylenia standardowego od drzewa minimalnego. Argumentem za regułą 1 SE jest to, że w badaniach symulacyjnych daje ona stabilny rozmiar drzewa w replikacjach, podczas gdy drzewo minimalne może się znacznie różnić w kolejnych replikacjach.

Problem zdarzeń rzadkich



Nowa obserwacja, która trafia do liścia, domyślnie jest klasyfikowana do klasy częstszej. Dlatego drzewo może mieć problem z poprawną klasyfikacją zdarzeń rzadkich (klas niezbalansowanych).

Opcja równe prawdopodobieństwa a priori (ang. *equal priors*) zaklasyfikuje obserwację, która znajdzie się w danym węźle, do klasy rzadszej, jeśli klasa ta będzie w tym węźle reprezentowana z większym prawdopodobieństwem niż w korzeniu.

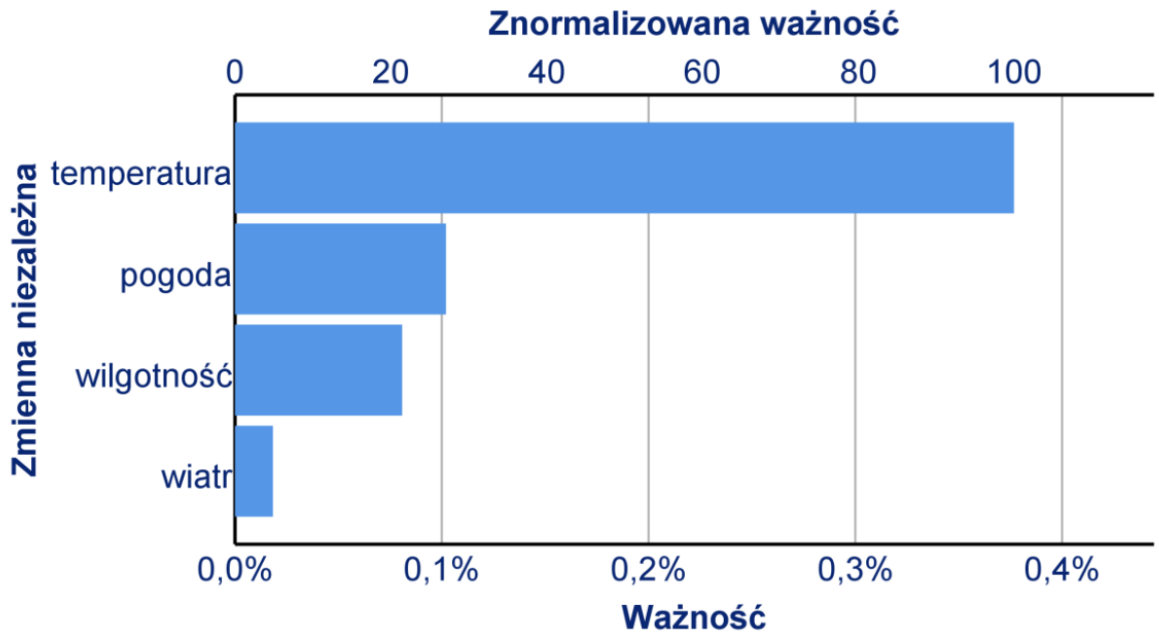
$$\frac{N_1(\text{węzeł})}{N_0(\text{węzeł})} > \frac{N_1(\text{korzeń})}{N_0(\text{korzeń})}.$$

Postępowanie z brakami danych

Nowa obserwacja może „utknąć” w trakcie klasyfikacji z użyciem CART, jeżeli napotka warunek bazujący na zmiennej, w której występuje brak danych. Dlatego CART tworzy surogaty (substytuty, węzły zastępcze, ang. *surrogate splits*), czyli warunki bazujące na innych zmiennych, ale dające podobny podział na węzły.

Ważność zmiennych

Ważność zmiennej (ang. *importance*) to miara bazująca na ważonej (odsetkiem obserwacji w węzłach) sumie popraw we wszystkich (z uwzględnieniem surogatów) węzłach, w których zmienna ta jest wykorzystywana do utworzenia podziału.



Reguły decyzyjne

Reguły decyzyjne otrzymujemy, przechodząc po drzewie dowolną ścieżką. Mają one postać „jeżeli . . . , to . . . ”.

Definicja

- Pokrycie reguły decyzyjnej to procent rekordów w zbiorze danych przypisanych do danego liścia.
- Ufność (ang. *confidence*) to procent rekordów w liściu, dla których reguła jest prawdziwa.
- Wsparcie (ang. *support*), inaczej pokrycie reguły to procent rekordów w zbiorze danych przypisanych do danego liścia i spełniających regułę.

Rodzaje klasyfikacji

- Prawdziwe klasyfikacje negatywne (ang. *true negatives*, TN) to rekordy, które zostały sklasyfikowane jako negatywne i w rzeczywistości są negatywne (zły jako zły).
- Prawdziwe klasyfikacje pozytywne (ang. *true positives*, TP) to rekordy, które zostały sklasyfikowane jako pozytywne i w rzeczywistości są pozytywne (dobry jako dobry).
- Fałszywe klasyfikacje negatywne (ang. *false negatives*, FN) to rekordy, które zostały sklasyfikowane jako negatywne, gdy w rzeczywistości są pozytywne (dobry jako zły).
- Fałszywe klasyfikacje pozytywne (ang. *false positives*, FP) to rekordy, które zostały sklasyfikowane jako pozytywne, gdy w rzeczywistości są negatywne (zły jako dobry).

Macierz pomyłek

Macierz pomyłek (ang. *confusion matrix*) – podsumowanie klasyfikacji wykonywane na zbiorze testowym.

		przewidywane $\widehat{Y}$	
		0	**1**
faktyczne Y	**0**	TN	FP
	1	FN	TP

Wskaźniki klasyfikacji binarnej

		przewidywane $\widehat{Y}$	
		0	1
faktyczne Y	0	40	10
	1	20	30

		przewidywane $\widehat{Y}$	
		0	1
faktyczne Y	0	40	10
	1	20	30

Trafność (ang. *accuracy*) – odsetek poprawnych klasyfikacji

$$\frac{\#TP + \#TN}{N} = \frac{40 + 30}{100} = 0,7.$$

Całkowity współczynnik błędu (ang. *overall error rate*) – odsetek błędnych klasyfikacji $= 1-$ trafność

$$\frac{\#FN + \#FP}{N} = \frac{20 + 10}{100} = 0,3.$$

Wskaźniki klasyfikacji binarnej c.d.

		przewidywane $\widehat{Y}$	
		0	1
faktyczne Y	0	40	10
	1	20	30

		przewidywane $\widehat{Y}$	
		0	1
faktyczne Y	0	40	10
	1	20	30

Czułość (ang. *sensitivity*, *recall*) – odsetek poprawnie sklasyfikowanych przypadków pozytywnych

$$\frac{\#TP}{\#FN + \#TP} = \frac{30}{20 + 30} = 0,6.$$

Specyficzność, swoistość (ang. *specifity*) – odsetek poprawnie sklasyfikowanych przypadków negatywnych

$$\frac{\#TN}{\#TN + \#FP} = \frac{40}{40 + 10} = 0,8.$$

Wskaźniki klasyfikacji binarnej c.d.

		przewidywane $\widehat{Y}$	
		0	1
faktyczne Y	0	40	10
	1	20	30

Proporcja prawdziwie pozytywnych, precyzja (ang. *proportion of true positive*, *precision*) – odsetek poprawnych klasyfikacji wśród wszystkich klasyfikacji pozytywnych

$$\frac{\#TP}{\#FP + \#TP} = \frac{30}{10 + 30} = 0,75.$$

Proporcja prawdziwie negatywnych (ang. *proportion of true negative*) – odsetek poprawnych klasyfikacji wśród wszystkich klasyfikacji negatywnych

$$\frac{\#TN}{\#TN + \#FN} = \frac{40}{40 + 20} \approx 0,67.$$

Wskaźniki klasyfikacji binarnej c.d.

Wynik $F1$ (ang. $F1$ *score*) jest średnią harmoniczną czułości i precyzji:

$$F1 = \frac{2}{\frac{1}{\text{czułość}} + \frac{1}{\text{precyzja}}} = \frac{2 \cdot \text{czułość} \cdot \text{precyzja}}{\text{czułość} + \text{precyzja}}$$

Jest stosowany zamiast trafności, zwłaszcza, gdy dane są niezbalansowane.
Dla danych z przykładu

$$F1 = \frac{2 \cdot 60\% \cdot 75\%}{60\% + 75\%} = 66\frac{2}{3}\%.$$

Krzywa ROC

Krzywa ROC – krzywa operacyjno-charakterystyczna odbiornika (ang. *Receiver Operating Characteristic*).
Kroki, które należy wykonać, żeby narysować krzywą:

1. Sortujemy rekordy malejąco ze względu na szacowaną przez klasyfikator wartość prawdopodobieństwa, że wartością zmiennej celu jest 1.
2. Każde z tych prawdopodobieństw kolejno staje się progiem powyżej którego klasyfikujemy zmienną jako 1, a w przeciwnym wypadku jako 0.
3. Dla każdego progu wyznaczamy czułość oraz swoistość.
4. Zaznaczamy na wykresie punkt o współrzędnych (1−swoistość, czułość).

Przykład

Wyznaczmy punkt krzywej ROC dla poniższych danych i progu 0,7.

Nr rekordu	Faktyczne Y	P-stwo, że $\widehat{Y} = 1$	Przewidywane $\widehat{Y}$
1	1	0,9	1
2	0	0,8	1
3	1	0,7	0
4	0	0,6	0
5	1	0,4	0
6	0	0,2	0

Macierz pomyłek dla przykładu

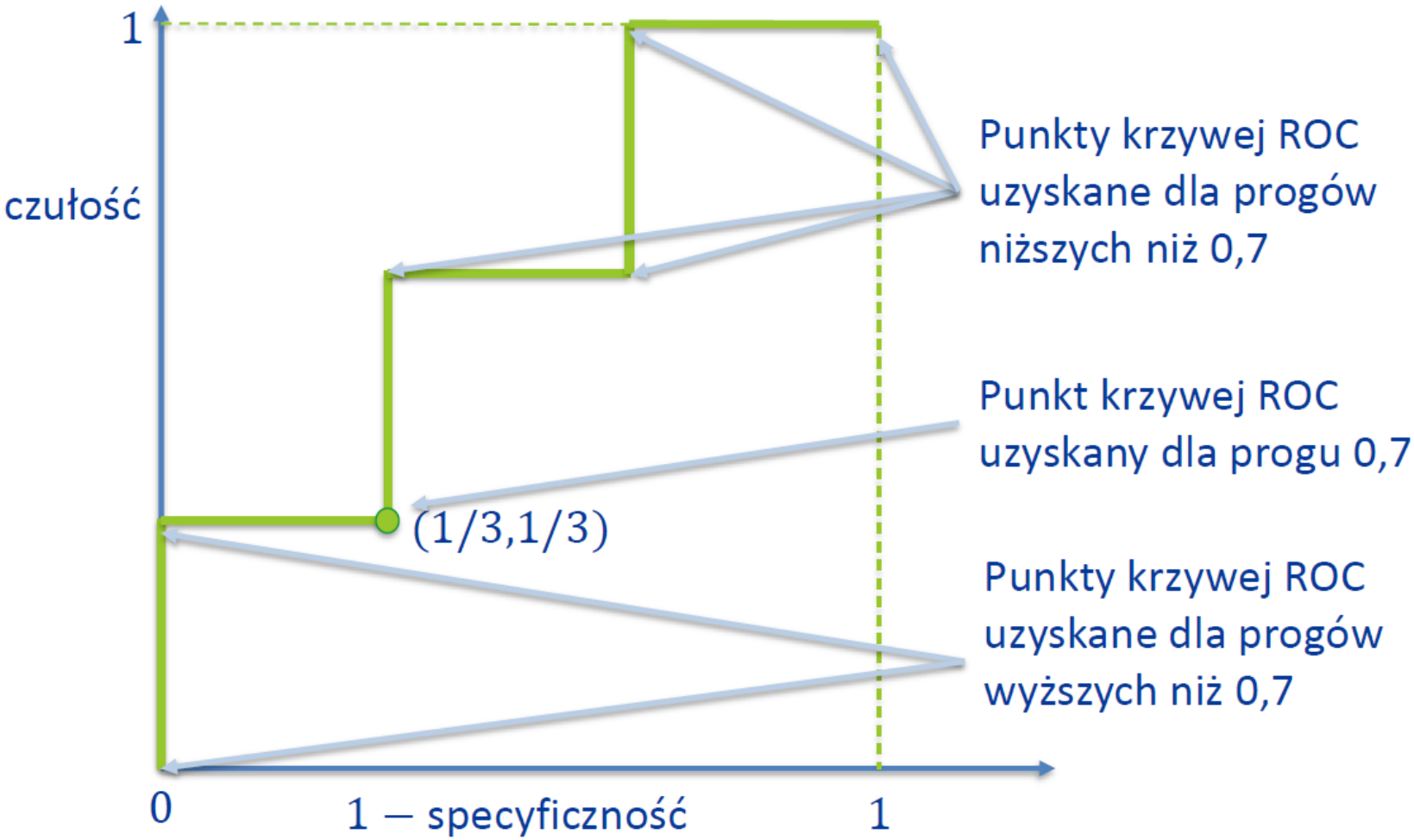
		przewidywane $\widehat{Y}$	
		0	**1**
faktyczne Y	**0**	2	1
	1	2	1

Stąd czułość wynosi 1/3, a specyficzność wynosi 2/3. Na krzywej zaznaczamy punkt o współrzędnych

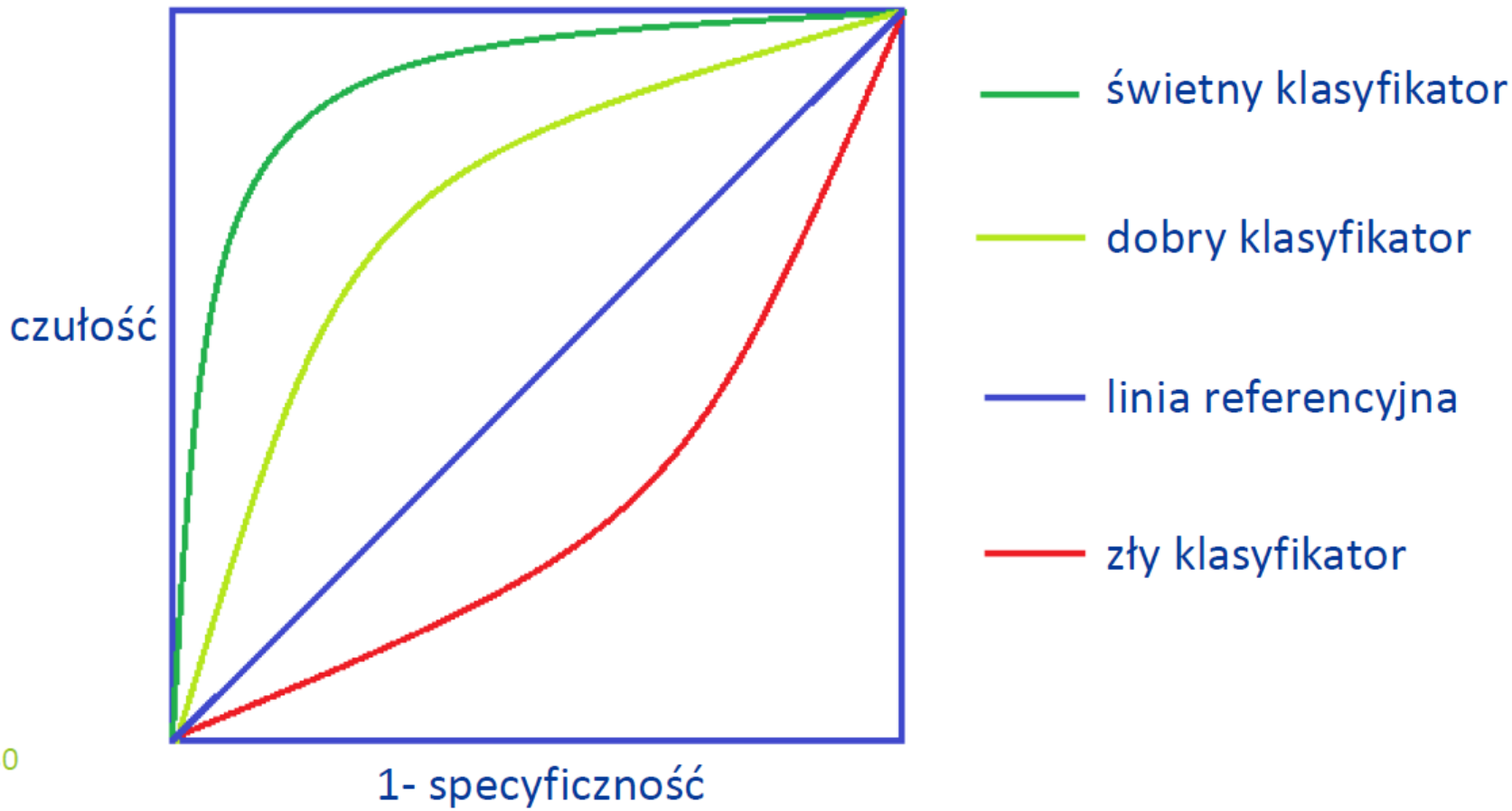
$$(1- \text{ specyficzność, czułość}) = (1/3, 1/3).$$

Analogicznie zaznaczamy punkty uzyskane dla innych progów i łączymy je ze sobą.

Krzywa ROC dla danych z przykładu



Kształt krzywej ROC



30

AUC

Pole pod krzywą ROC (ang. *Area Under Curve*, AUC) może być traktowane jako miara sukcesu klasyfikatora i może służyć do porównywania jakości klasyfikatorów.

Wartości AUC a jakość klasyfikatora (umownie):

- $AUC \in (0,6; 0,7]$ – słaba,
- $AUC \in (0,7; 0,8]$ – akceptowalna,
- $AUC \in (0,8; 0,9]$ – dobra,
- $AUC \in (0,9; 1]$ – doskonała.

Źródło: http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture22.htm

Strategie obliczania AUC w przypadku walidacji krzyżowej

1. Uśrednianie (ang. *averaging*) wyznaczamy krzywą ROC i obliczamy AUC na każdym ze zbiorów testowych z osobna, a następnie uśredniamy wynik.
2. Łączenie (ang. *pooling*) zbieramy wyniki działania klasyfikatora na każdym ze zbiorów testowych i łączymy je w jeden duży zbiór Dla tego zbioru rysujemy krzywą ROC i obliczamy AUC. Ta metoda jest łatwiejsza w implementacji i jest jedyną możliwą do zastosowania w przypadku N-krotnej walidacji krzyżowej (ang. *leave one out*).

Optymalny punkt odcięcia

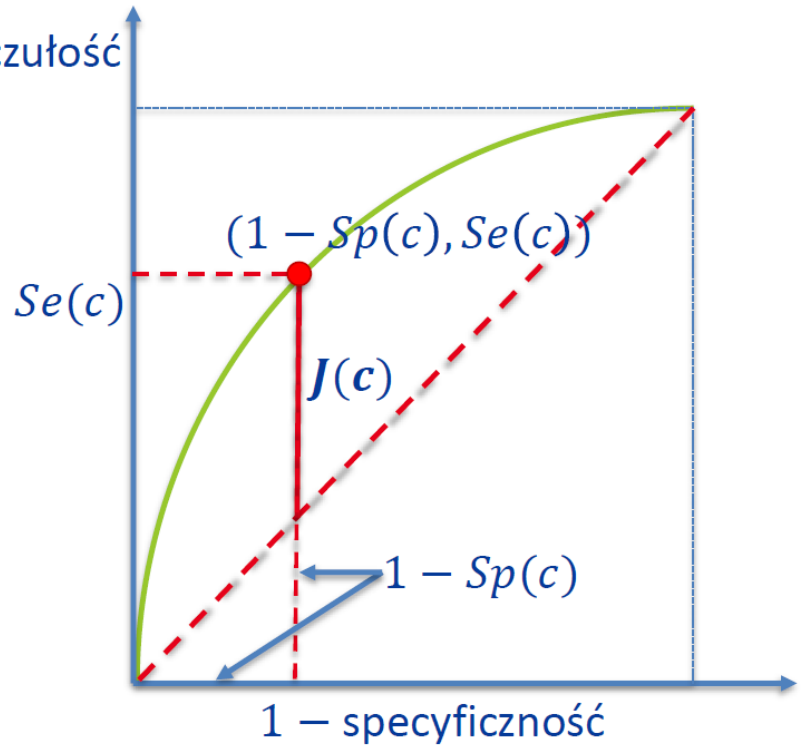
Krzywa ROC jest także używana do określania przydatności diagnostycznej biomarkera. Tak więc dla każdego poziomu biomarkera uznajemy za przypadki pozytywne wartości powyżej progu, a pozostałe jako negatywne (ewentualnie odwrotnie). Porównujemy je z informacją o wartościach zmiennej celu i wyznaczamy czułość i specyficzność. Procedura ta pozwala na uzyskanie krzywej ROC. Jeśli AUC jest duże, to biomarker jest przydatny w przewidywaniu zmiennej celu. Na krzywej znajdujemy optymalny punkt odcięcia (ang. *optimal cut-off*). Odpowiadający mu próg jest szukanym poziomem biomarkera.

Indeks Youdena

Podany przez Williama J. Youdena (1950). Wybieramy jako punkt odcięcia taką wartość progową c, która maksymalizuje funkcję

$$J(c) = Se(c) - (1 - Sp(c)),$$

gdzie $Se(c)$ jest czułością przy progu c, a $Sp(c)$ jest specyficznością. Pozwala to odnaleźć punkt krzywej o największej odległości w pionie od prostej o równaniu czułość $=$ 1$-$ specyficzność.

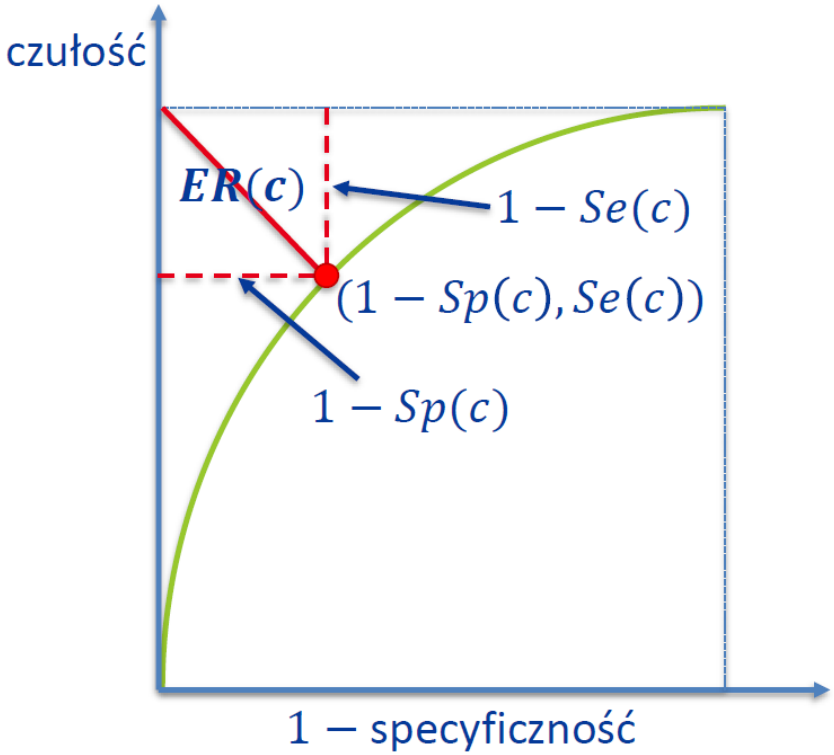


Kryterium bycia najbliżej punktu $(0, 1)$

Wybieramy jako punkt odcięcia taką wartość progową c, która minimalizuje funkcję

$$ER(c) = \sqrt{(1 - Se(c))^2 + (1 - Sp(c))^2}.$$

Pozwala to odnaleźć punkt krzywej leżący najbliżej punktu $(0, 1)$.

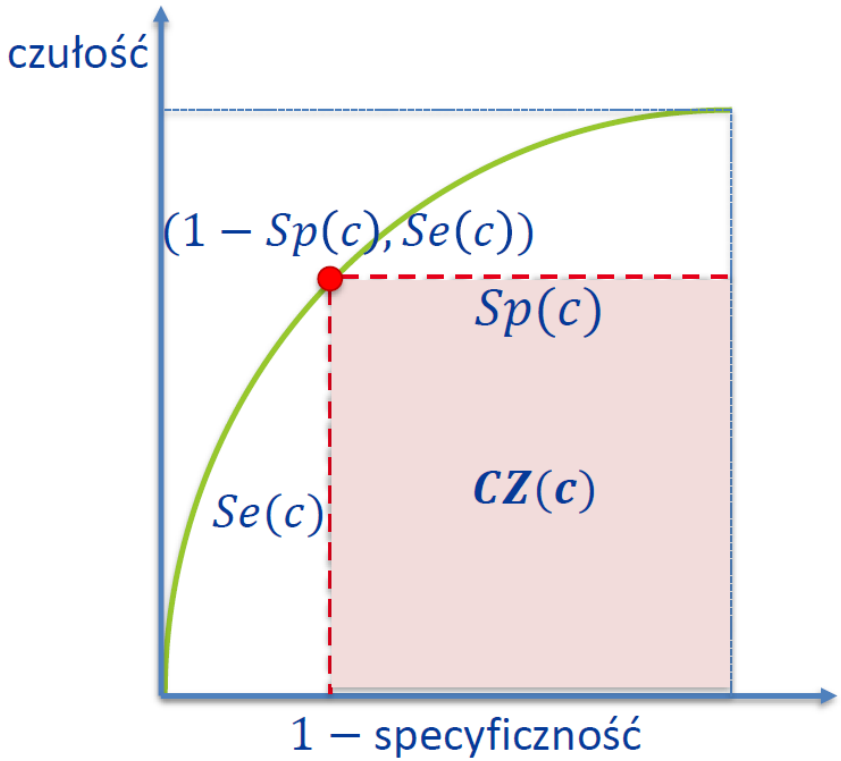


Metoda prawdopodobieństwa zgodności

Ang. *concordance probability method*. Podana przez X. Liu (2012). Wybieramy jako punkt odcięcia taką wartość progową c, która maksymalizuje funkcję

$$CZ(c) = Se(c) \cdot Sp(c).$$

Pozwala to odnaleźć punkt krzywej wyznaczający prostokąt o największym polu.



Klasyfikacja – więcej niż dwie klasy

Przypadek klasyfikacji do więcej niż dwóch klas $k_1, k_2, \ldots, k_I$ można w ocenie jakości klasyfikatora traktować jako I klasyfikacji do dwóch klas:

- k_1 kontra pozostałe klasy razem,
- k_2 kontra pozostałe klasy razem,
- $\vdots$

Wtedy wszystkie miary jakości klasyfikatora wyznacza się I razy.

Ciągła zmienna celu

Dana jest ciągła zmienna celu Y i N rekordów, dla których ma ona wartości $y_1, y_2, \ldots, y_N$, np. zmienna *roczny dochód z klienta*.
W zbiorze uczącym i testowym wartości zmiennej celu są znane i dlatego oceniając jakość szacowania, możemy je wykorzystać i porównać z wartościami $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N$ przewidzianymi przez model.

Miary jakości szacowania

- Średni błąd bezwzględny (ang. *mean absolute error*, MAE)

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}.$$

- Błąd średniokwadratowy (ang. *mean-squared error*, MSE)

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}.$$

Czasami, dla uniknięcia tzw. obciążenia, błąd średniokwadratowy jest liczony wzorem

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N - 1 - p}.$$

Pierwiastek błędu średniokwadratowego (RMSE) jest średnią kwadratową odchyleń wartości przewidywanych od rzeczywistych.

Miary jakości szacowania c.d.

- Średni bezwzględny błąd procentowy (ang. *mean absolute percentage error*, MAPE)

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i}.$$

Nie może być liczony, gdy y_i są równe lub bliskie 0. Mocniej karze błędy niedoszacowania $y_i > \hat{y}_i$, niż przeszacowania.

- Symetryczny średni bezwzględny błąd procentowy (ang. *symmetric mean absolute percentage error*, SMAPE)

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}.$$

Bagging

Bagging (inaczej agregacja bootstrapowa, ang. *bootstrap aggregating*), metoda zaproponowana przez Leo Breimana w 1994 roku, polega na wylosowaniu z oryginalnego zbioru danych o liczebności N próbki o liczebności N metodą losowania ze zwracaniem zgodnie z rozkładem jednostajnym. Powstaje w ten sposób N-elementowy zbiór treningowy, w którym niektóre rekordy wyjściowego zbioru danych się powtarzają. Prawdopodobieństwo wylosowania danego rekordu do tej próbki wynosi

$$1 - \left(1 - \frac{1}{N}\right)^N \rightarrow_{N \rightarrow \infty} 1 - e^{-1} \approx 0,632.$$

Wylosowana próba tworzy zbiór uczący, a niewybrane elementy zbiór testowy, na którym oceniana jest trafność klasyfikatora tr_i.

Bagging – procedura

Dane
1
2
3
4
5
6
7

Próbka 1
1
4
6
1
3
3
4

Próbka 2
7
2
7
5
4
4
2

...

Próbka k
5
5
2
7
3
7
2

Zbiory uczące

Niewybrane 1
2
5
7

Niewybrane 2
1
3
6

...

Niewybrane k
1
4
6

Zbiory testowe

Estymator *.632 bootstrap*

Procedurę losowania próbki powtarzamy k-krotnie, otrzymując k klasyfikatorów. Ogólna trafność szacujemy estymatorem *.632 bootstrap*.

$$tr_{boot} = \frac{1}{k}\sum_{i=1}^{k}(0,632 \cdot tr_i + 0,368 \cdot tr_{opt}),$$

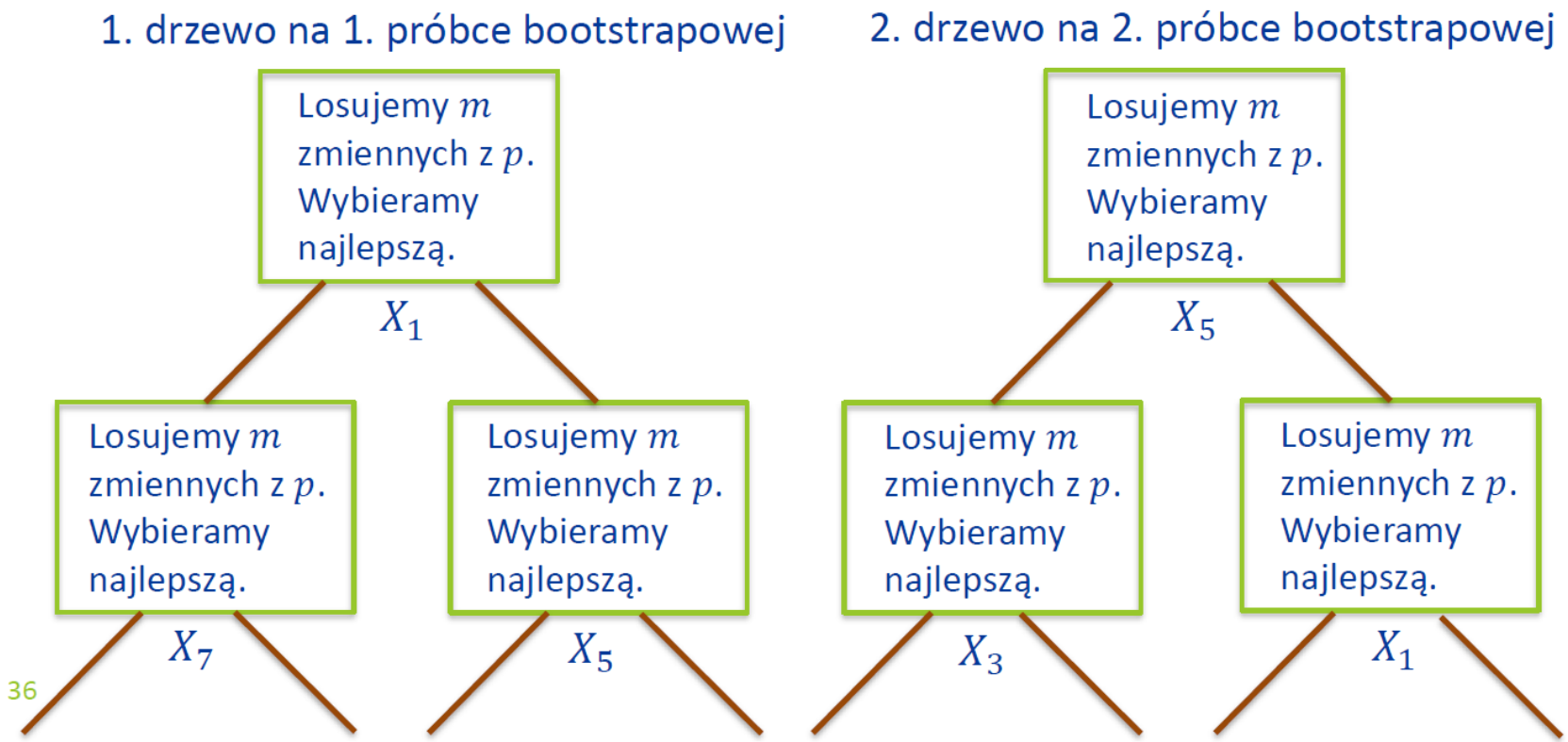
gdzie tr_{opt} oznacza trafność klasyfikatora skonstruowanego dla wszystkich elementów oryginalnego zbioru danych.
Określenie klasy nowego rekordu odbywa się metodą głosowania większościowego poszczególnych klasyfikatorów.

Kroki algorytmu

Las losowy (ang. *random forest*) – Leo Breiman i Adele Cutler (2001).

1. Ze zbioru rekordów $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ losujemy k razy próbę o liczebności N metodą losowania ze zwracaniem zgodnie z rozkładem jednostajnym (próba bootstrapowa).
2. Dla każdej takiej próby budujemy drzewo (CART), ale w każdym wierzchołku spośród p dostępnych zmiennych losujemy tylko m ($m < p$) i tylko one są uwzględniane przy szukaniu zmiennej dającej najlepszy podział).
3. Otrzymujemy w ten sposób k drzew. Klasyfikacja nowego rekordu odbywa się zgodnie z zasadą głosowania większościowego, a szacowanie poprzez uśrednienie wartości otrzymanych z pojedynczych drzew.

Schemat lasu losowego



Uwagi

- Do budowy lasów losowych używa się drzew CART.
- Domyślnie $m = \lfloor \sqrt{p} \rfloor$ dla klasyfikacji, a $m = \lfloor p/3 \rfloor$ dla szacowania. Minimalny rozmiar liści to odpowiednio 1 i 5.
- Błąd klasyfikacji stabilizuje się przy ok. 100 drzewach.
- Trafność klasyfikacji lasu losowego ulega pogorszeniu, gdy liczba zmiennych jest duża, ale liczba istotnych zmiennych jest niewielka.
- Lasy losowe są odporne na przeuczenie.