Regresja logistyczna

Liniowa analiza dyskryminacyjna

Classification

Igor Adamiec

K najbliższych sąsiadów

Czym jest klasyfikacja

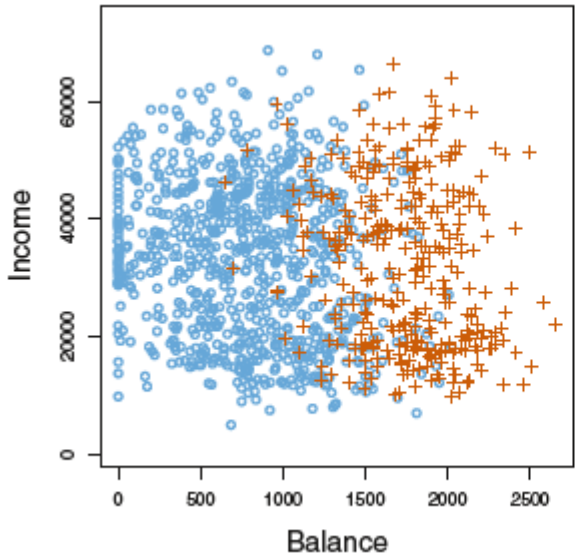
W przeciwieństwie do regresji (która przewidywała wartości liczbowe), klasyfikacja przewiduje zmienne jakościowe (kategoryczne).

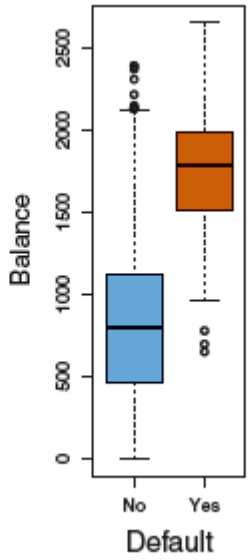
Przykłady: przewidywanie choroby na podstawie symptomów,

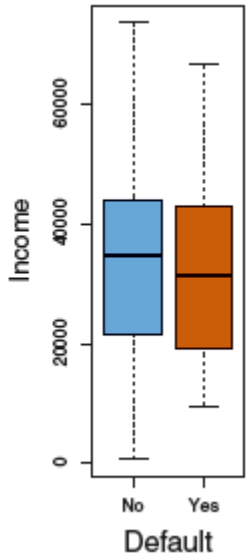
Przewidywanie czy transakcja jest przestępcza,

Przewidywanie, które mutacje DNA powodują daną chorobę, a które nie

Default dataset

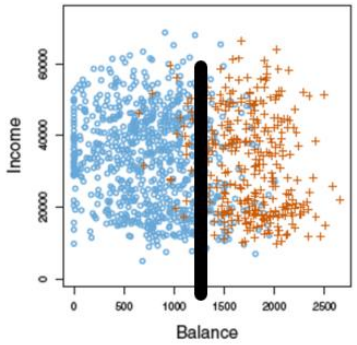






Przewidujemy, czy klient banku nie spłaci kredytu (default) na podstawie jego dochodu i bilansu karty kredytowej

Widać, że zmienna balance jest mocno skorelowana z kategorią default. Dla ludzi, z wartością „Yes”, wartość zmiennej balance średnio jest dużo wyższa.



Widać to na wykresie punktowym. Praktycznie możemy narysować tam pionową kreskę i stworzymy całkiem dobry klasyfikator

Why not linear regression

W przypadku więcej niż 2 rodzajów odpowiedzi musielibyśmy założyć, że są one w jakiś sposób uszeregowane i znajdują się w tej samej odległości od siebie.
O ile jest to możliwe np. w przypadku odpowiedzi w ankiecie (Bardzo źle, źle, średnio, dobrze, bardzo dobrze), tak w przypadku np. choroby jest to niemożliwe

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

W przypadku binarnej odpowiedzi (tylko dwie kategorie) byłoby możliwe wyliczenie tego regresją liniową najmniejszych kwadratów.

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

Odpowiedzią klasyfikacji jest jednak prawdopodobieństwo, że dana obserwacja należy do danej klasy, a w przypadku regresji liniowej moglibyśmy otrzymać prawdopodobieństwo większe od 1 lub nawet ujemne.

$$\Pr(\texttt{drug overdose}|X)$$