Techniki Optymalizacji: Metody klasyfikacji

Wojciech Kotłowski

Instytut Informatyki Politechniki Poznańskiej
email: imię.nazwisko@cs.put.poznan.pl

pok. 2 (CW) tel. (61)665-2936 konsultacje: piątek 15:10-16:50
Slajdy dostępne pod adresem: `http://www.cs.put.poznan.pl/wkotlowski/to/`

13.11.2018

Spis treści

Spis treści

Problem klasyfikacji

- Przewidywania/wyjaśnienie jednej zmiennej dyskretnej (Y) przez inne zmienne ($\boldsymbol{X}$).

Problem klasyfikacji

- Przewidywania/wyjaśnienie jednej zmiennej dyskretnej (Y) przez inne zmienne ($\boldsymbol{X}$).
- Klasyfikacja vs. regresja:
  - Regresja – Y ciągła.
  - Klasyfikacja – Y dyskretna, np. $Y \in \{1, \ldots, K\}$.

Problem klasyfikacji

- Przewidywania/wyjaśnienie jednej zmiennej dyskretnej (Y) przez inne zmienne ($\boldsymbol{X}$).
- Klasyfikacja vs. regresja:
  - Regresja – Y ciągła.
  - Klasyfikacja – Y dyskretna, np. $Y \in \{1, \ldots, K\}$.
- Wartości Y nazywane: klasami, kategoriami, etykietami, ...

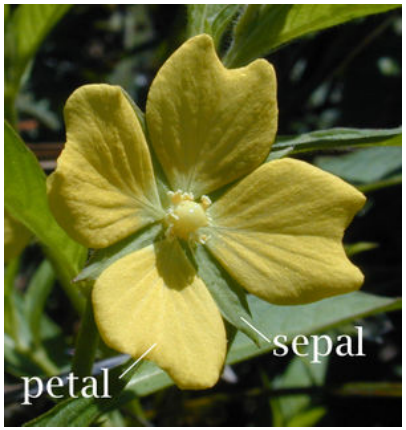
Problem klasyfikacji

- Przewidywania/wyjaśnienie jednej zmiennej dyskretnej (Y) przez inne zmienne ($\boldsymbol{X}$).
- Klasyfikacja vs. regresja:
  - Regresja – Y ciągła.
  - Klasyfikacja – Y dyskretna, np. $Y \in \{1, \ldots, K\}$.
- Wartości Y nazywane: klasami, kategoriami, etykietami, ...

Przykłady

- $\boldsymbol{X}$ – wyniki testów medycznych, $Y \in \{chory, zdrowy\}$.
- $\boldsymbol{X}$ – jasność pikseli w obrazie $Y \in \{0, \ldots, 9\}$ – cyfra na obrazie.
- $\boldsymbol{X}$ – słowa w dokumencie tekstowym, $Y \in \{spam, niespam\}$.
- $\boldsymbol{X}$ – treść strony internetowej + słowa kluczowe zapytania, $Y \in \{odwiedzona, nieodwiedzona\}$ – czy strona została(by) odwiedzona po pokazaniu w wyszukiwarce.

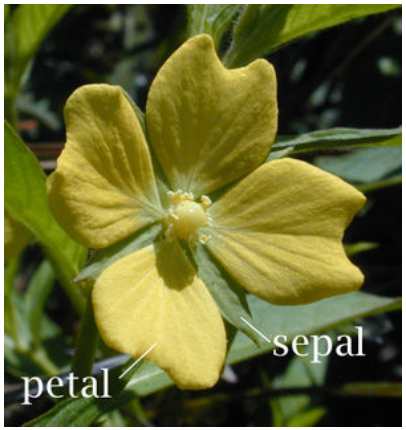
Przykład



- Zbiór uczący: IRIS (Ronald Fisher, 1936)
- Zmienne wejściowe ($\boldsymbol{X}$): *sepal-length*, *sepal-width*, *petal-length*, *petal-width*
- Zmienna wyjściowa (Y): *type*

X_1 **sepal length**	X_2 **sepal width**	X_3 **petal length**	X_4 **petal width**	Y **type**
4.4	2.9	1.4	0.2	setosa
6.8	2.8	4.8	1.4	versicolor
5.1	3.5	1.4	0.2	setosa
7.7	3.0	6.1	2.3	virginica
6.2	2.9	4.3	1.3	versicolor
...	...	...	...	...

Przykład



- Zbiór uczący: IRIS (Ronald Fisher, 1936)
- Zmienne wejściowe ($\boldsymbol{X}$): *sepal-length*, *sepal-width*, *petal-length*, *petal-width*
- Zmienna wyjściowa (Y): *type*

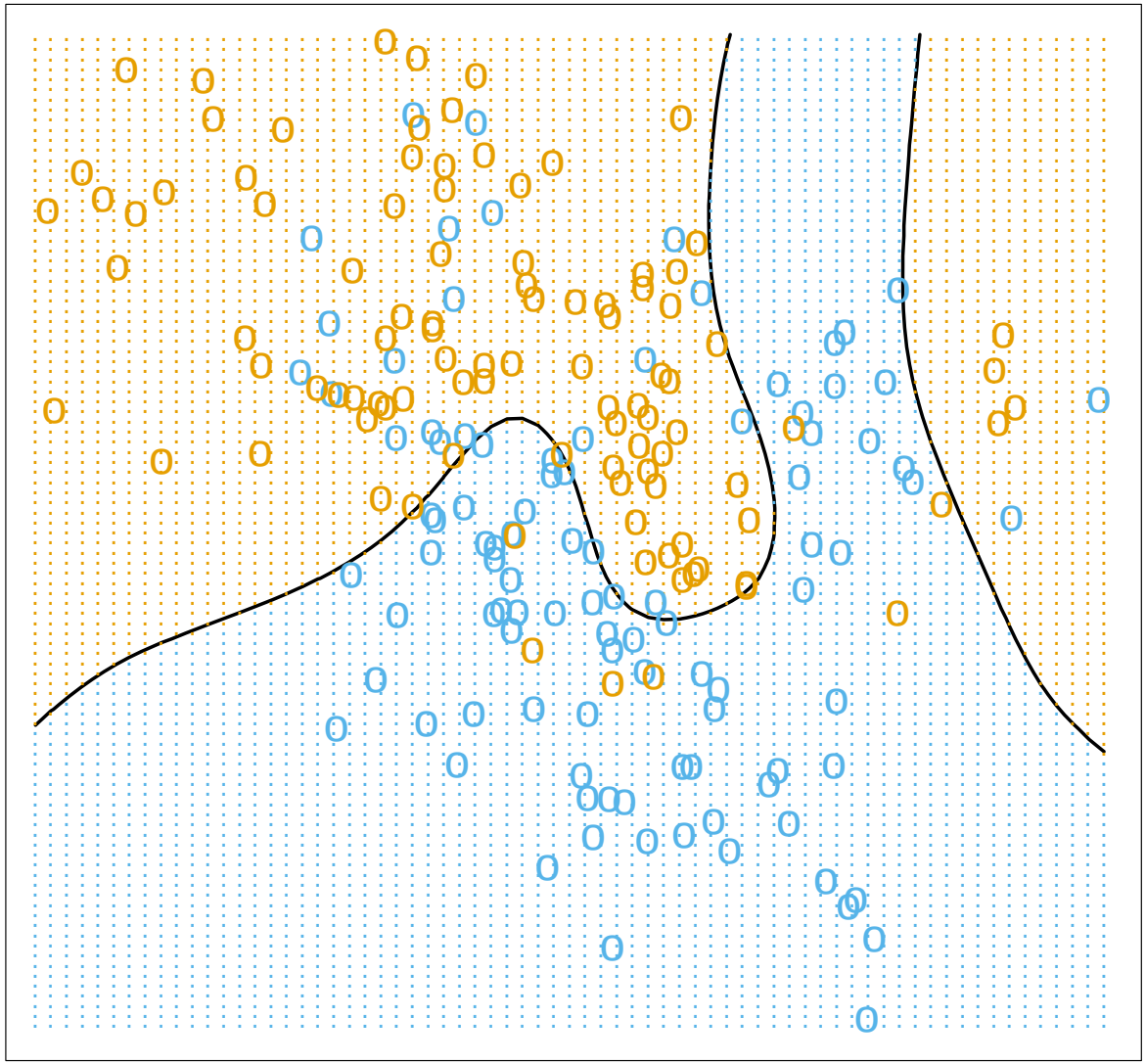
X_1 **sepal length**	X_2 **sepal width**	X_3 **petal length**	X_4 **petal width**	Y **type**
4.4	2.9	1.4	0.2	setosa
6.8	2.8	4.8	1.4	versicolor
5.1	3.5	1.4	0.2	setosa
7.7	3.0	6.1	2.3	virginica
6.2	2.9	4.3	1.3	versicolor
...	...	...	...	...
4.0	2.9	1.9	1.0	?

Klasyfikator

- Klasyfikator to funkcja, która każdemu $\boldsymbol{x}$ przyporządkowuje przewidywaną klasę $\hat{y}$.
- Klasyfikator jest zwykle wyznaczany (uczony) na podstawie zbioru danych (treningowych).
- Klasyfikator dzieli przestrzeń $\boldsymbol{x}$-ów na obszary odpowiadające przewidywanym klasom.

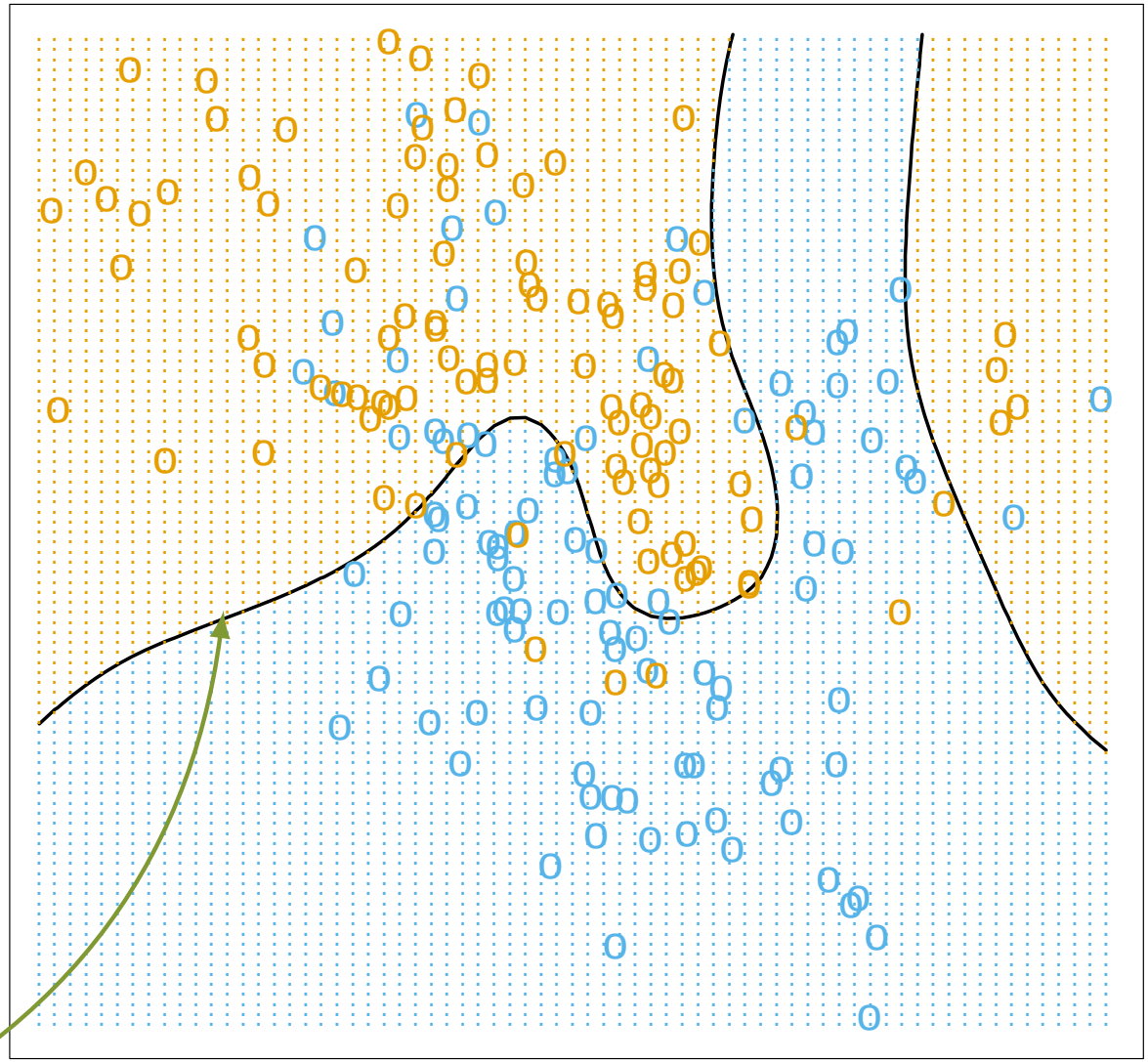
Przykład – 2 klasy

Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Przykład – 2 klasy

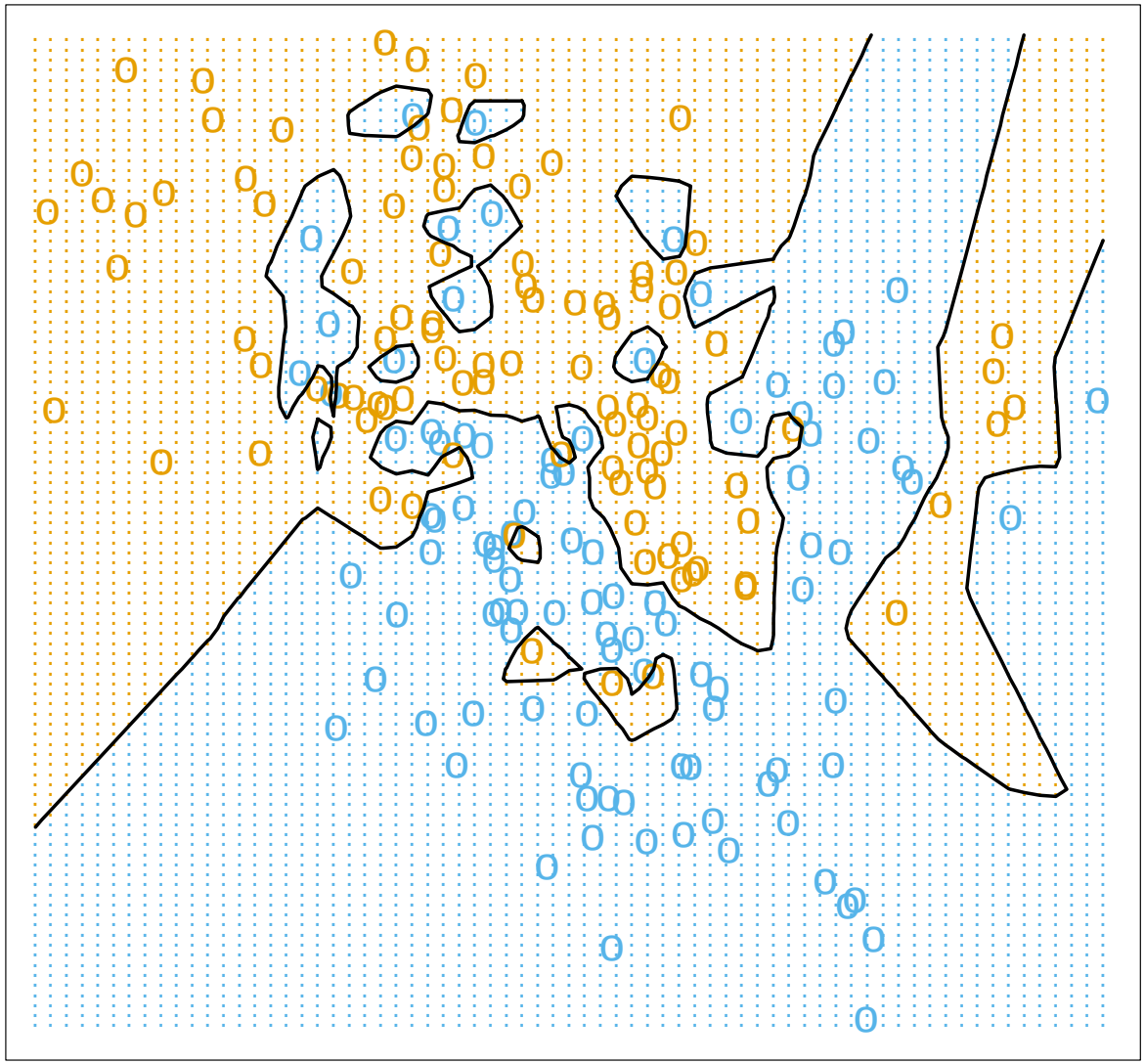
Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Granica między klasami jest nazywana granicą decyzyjną

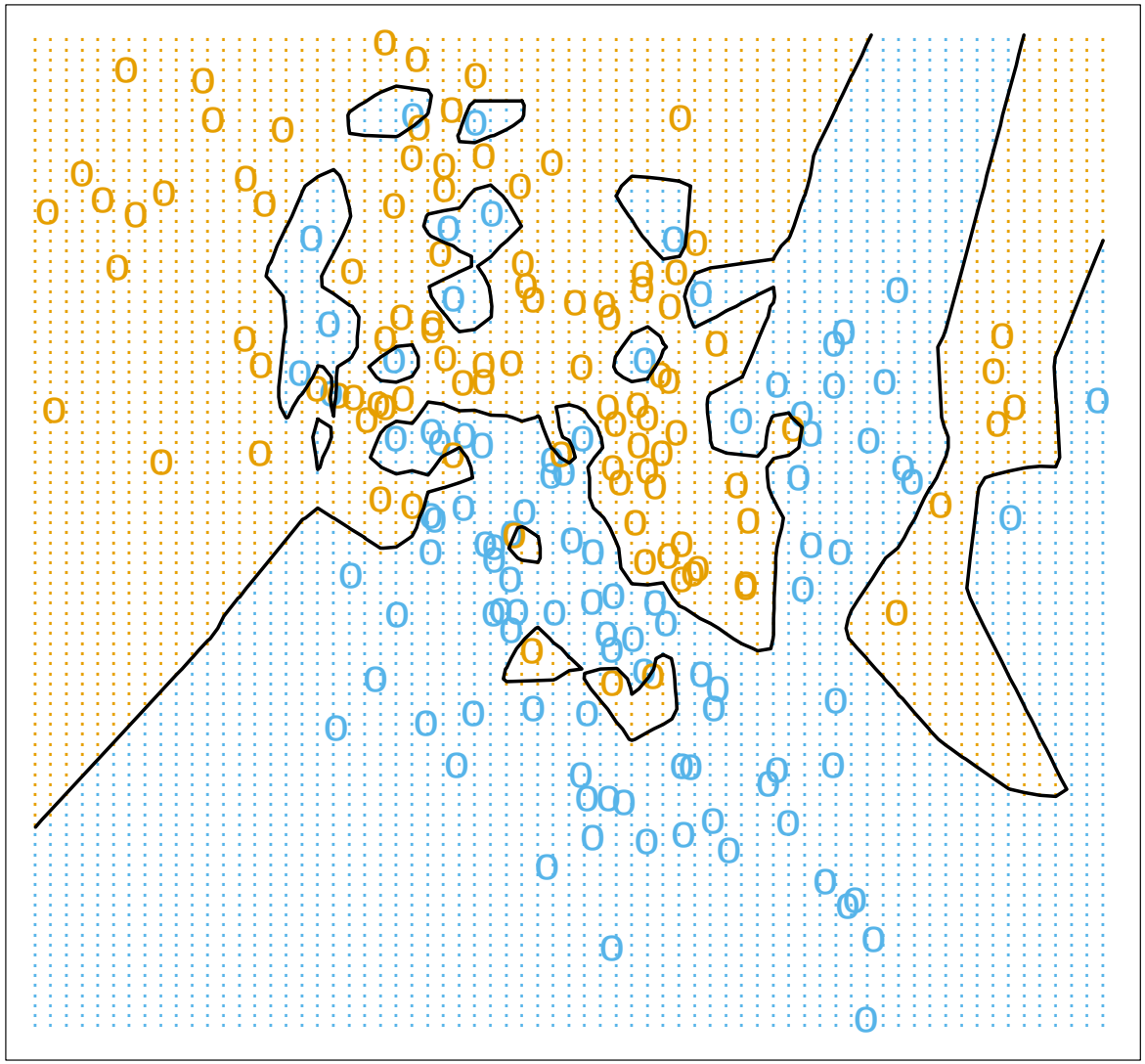
Przykład – 2 klasy

Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Przykład – 2 klasy

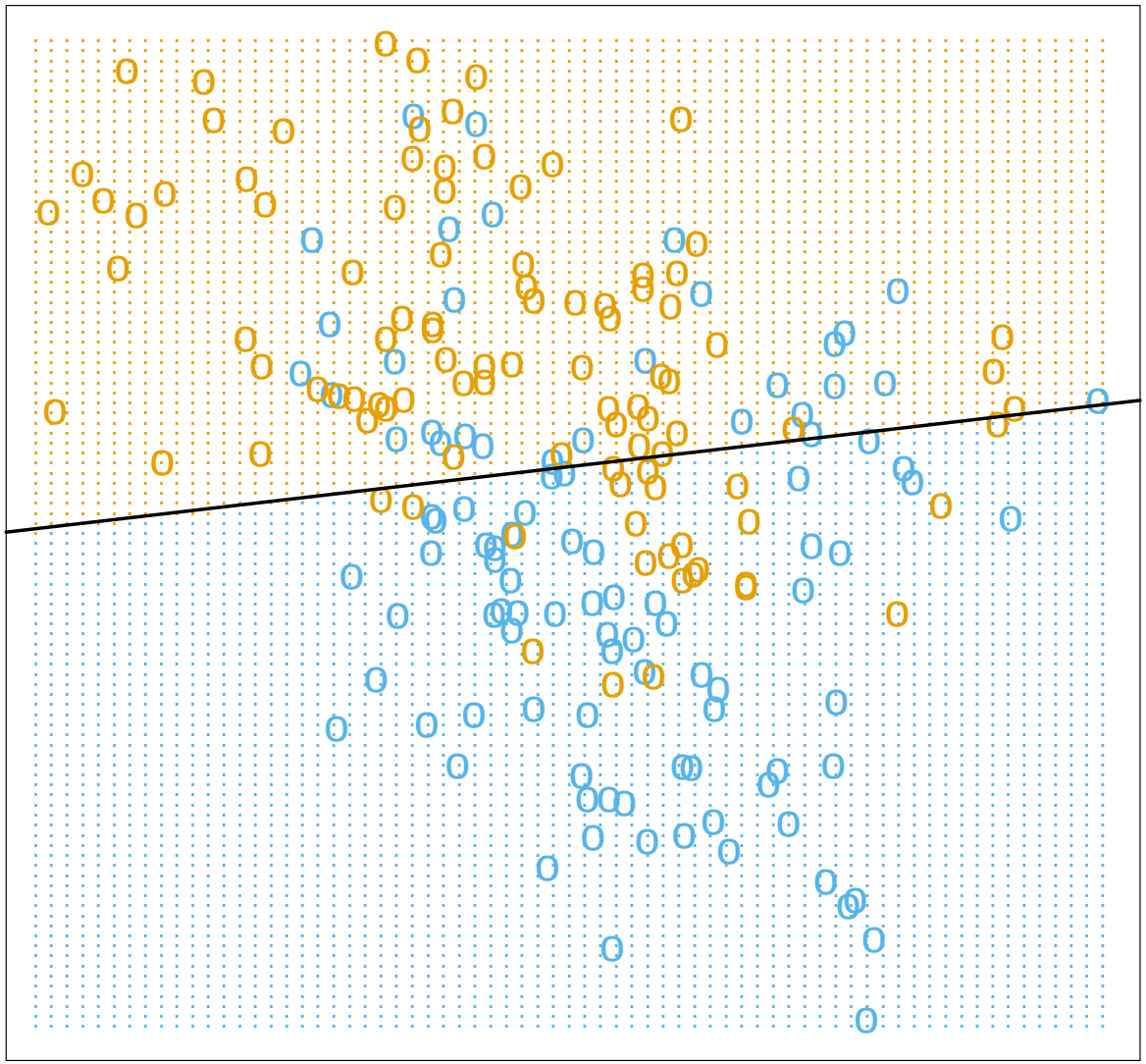
Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Użycie klasyfikatora typu „najbliższy sąsiad” (*nearest neighbour*).

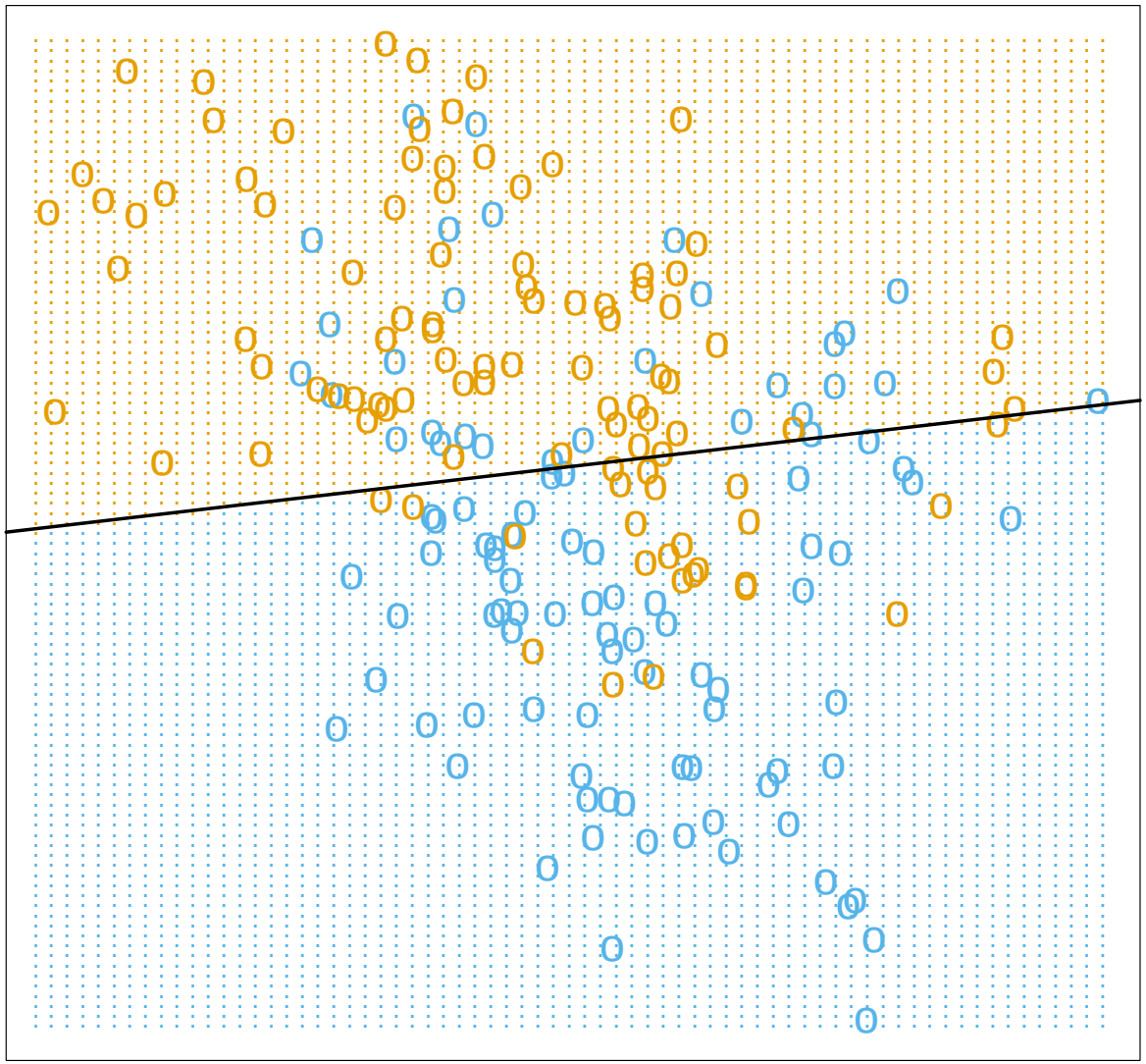
Przykład – 2 klasy

Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Przykład – 2 klasy

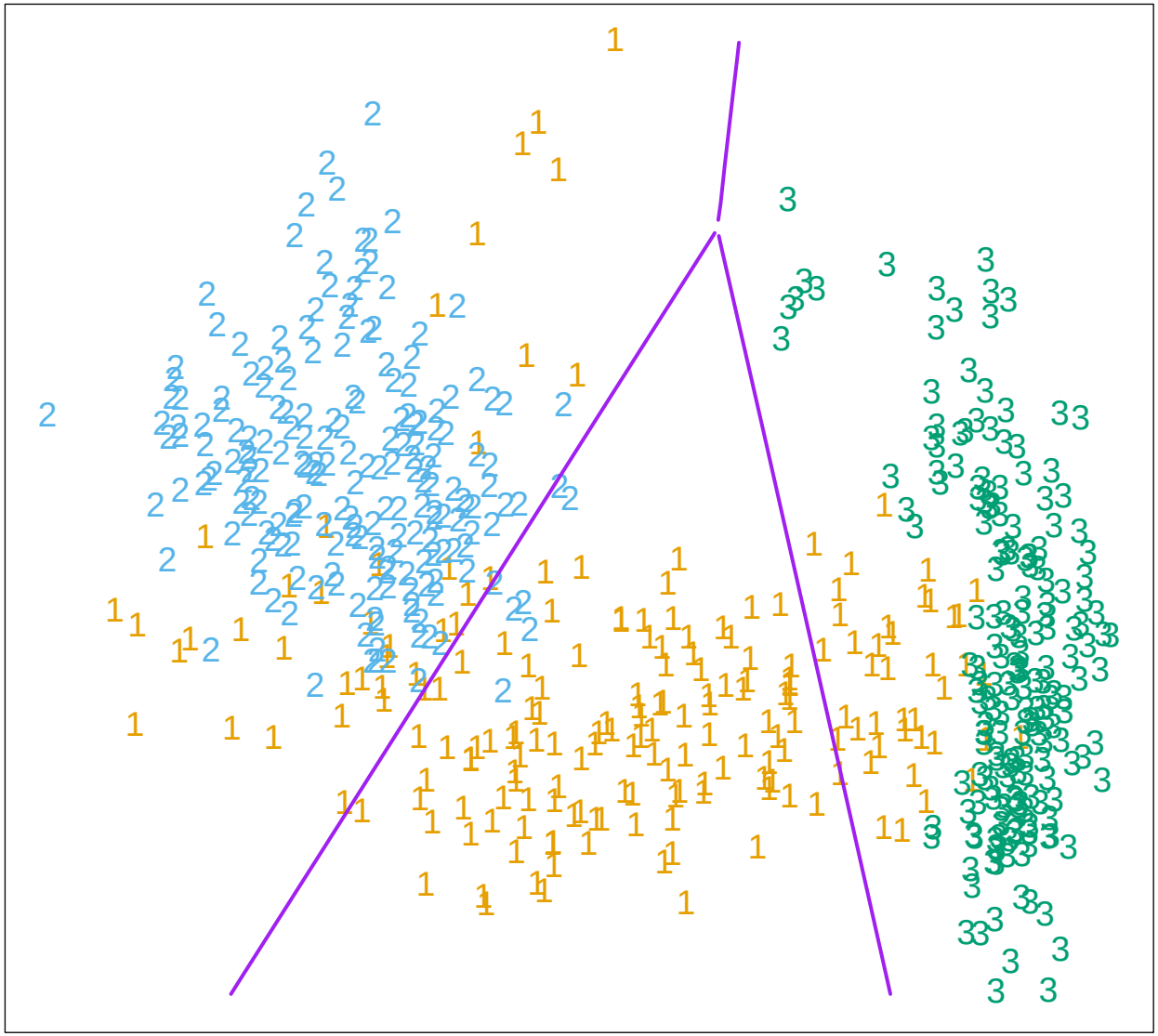
Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Klasyfikator liniowy: granica decyzyjna jest funkcją liniową.

Przykład – 3 klasy (Iris)

Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Klasyfikator liniowy: granica decyzyjna jest funkcją liniową.

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową
- Klasy zwykle kodowane jako $Y \in \{-1, +1\}$.

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową
- Klasy zwykle kodowane jako $Y \in \{-1, +1\}$.
- Klasyfikator złożony z dwóch części:

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową
- Klasy zwykle kodowane jako $Y \in \{-1, +1\}$.
- Klasyfikator złożony z dwóch części:
  - Funkcja liniowa $\boldsymbol{X}$:

$$f(\boldsymbol{X}) = w_0 + \sum_{j=1}^{m} w_j X_j = \boldsymbol{w}^\top \boldsymbol{X}$$

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową
- Klasy zwykle kodowane jako $Y \in \{-1, +1\}$.
- Klasyfikator złożony z dwóch części:
  - Funkcja liniowa $\boldsymbol{X}$:

$$f(\boldsymbol{X}) = w_0 + \sum_{j=1}^{m} w_j X_j = \boldsymbol{w}^\top \boldsymbol{X}$$

  - Klasyfikacja poprzez progowanie w zerze:

$$\hat{Y}(\boldsymbol{X}) = \begin{cases} +1 & \text{jeśli } f(\boldsymbol{X}) \geq 0 \\ -1 & \text{jeśli } f(\boldsymbol{X}) < 0 \end{cases} = \operatorname{sgn}(f(\boldsymbol{X})).$$

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową
- Klasy zwykle kodowane jako $Y \in \{-1, +1\}$.
- Klasyfikator złożony z dwóch części:
  - Funkcja liniowa $\boldsymbol{X}$:

$$f(\boldsymbol{X}) = w_0 + \sum_{j=1}^{m} w_j X_j = \boldsymbol{w}^\top \boldsymbol{X}$$

  - Klasyfikacja poprzez progowanie w zerze:

$$\hat{Y}(\boldsymbol{X}) = \begin{cases} +1 & \text{jeśli } f(\boldsymbol{X}) \geq 0 \\ -1 & \text{jeśli } f(\boldsymbol{X}) < 0 \end{cases} = \operatorname{sgn}(f(\boldsymbol{X})).$$

- Dla większej ilości klas zwykle rozbicie problemu na problemy dwuklasowe (jedna klasa vs. pozostałe klasy).

Klasyfikacja liniowa – 2 klasy

- Granica decyzyjna jest funkcją liniową
- Klasy zwykle kodowane jako $Y \in \{-1, +1\}$.
- Klasyfikator złożony z dwóch części:
  - Funkcja liniowa $\boldsymbol{X}$:

$$f(\boldsymbol{X}) = w_0 + \sum_{j=1}^{m} w_j X_j = \boldsymbol{w}^\top \boldsymbol{X}$$

  - Klasyfikacja poprzez progowanie w zerze:

$$\hat{Y}(\boldsymbol{X}) = \begin{cases} +1 & \text{jeśli } f(\boldsymbol{X}) \geq 0 \\ -1 & \text{jeśli } f(\boldsymbol{X}) < 0 \end{cases} = \operatorname{sgn}(f(\boldsymbol{X})).$$

- Dla większej ilości klas zwykle rozbicie problemu na problemy dwuklasowe (jedna klasa vs. pozostałe klasy).
- Model liniowy jest bardzo ogólny:
  - Przykład: klasyfikacja wielomianowa to klasyfikacja liniowa!
  - Mając X, tworzymy zmienne $X_1 = X, X_2 = X^2, X_3 = X^3, \ldots$

$$f(\boldsymbol{X}) = w_1 X + w_2 X^2 + \ldots + w_0 = w_1 X_1 + w_2 X_2 + \ldots + w_0$$

Spis treści

Trenowanie klasyfikatora liniowego

- Mając zbiór uczący $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, jak znaleźć wektor współczynników $\boldsymbol{w}$?

Trenowanie klasyfikatora liniowego

- Mając zbiór uczący $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, jak znaleźć wektor współczynników $\boldsymbol{w}$?

Podejście I: Minimalizacja błędu klasyfikacji

- Naturalny błąd klasyfikacji (błąd 0/1): liczba niepoprawnie sklasyfikowanych obserwacji.
- Wyznacz współczynniki $\boldsymbol{w}$ jako te, które minimalizują całkowity błąd klasyfikacji na danych.

Minimalizacja błędu klasyfikacji

- Prosty zapis błędu klasyfikacji ($y_i, \hat{y}_i \in \{-1, 1\}$):

$$\mathrm{err}_i = \begin{cases} 1 & \text{jeśli } y_i \hat{y}_i \leq 0 \\ 0 & \text{jeśli } y_i \hat{y}_i > 0 \end{cases} = \mathbb{1}[y_i \hat{y}_i \leq 0],$$

  gdzie $\mathbb{1}[C]$ to funkcja indykatorowa: `C ? 1 : 0`.

Minimalizacja błędu klasyfikacji

- Prosty zapis błędu klasyfikacji ($y_i, \hat{y}_i \in \{-1, 1\}$):

$$\mathrm{err}_i = \begin{cases} 1 & \text{jeśli } y_i \hat{y}_i \leq 0 \\ 0 & \text{jeśli } y_i \hat{y}_i > 0 \end{cases} = \mathbb{1}[y_i \hat{y}_i \leq 0],$$

  gdzie $\mathbb{1}[C]$ to funkcja indykatorowa: `C ? 1 : 0`.

- Minimalizujemy:

$$\begin{aligned} \min : L &= \sum_{i=1}^{n} \mathrm{err}_i \\ &= \sum_{i=1}^{n} \mathbb{1}[y_i \hat{y}_i \leq 0] \\ &= \sum_{i=1}^{n} \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]. \end{aligned}$$

Minimalizacja błędu klasyfikacji

- Prosty zapis błędu klasyfikacji ($y_i, \hat{y}_i \in \{-1, 1\}$):

$$\mathrm{err}_i = \begin{cases} 1 & \text{jeśli } y_i \hat{y}_i \leq 0 \\ 0 & \text{jeśli } y_i \hat{y}_i > 0 \end{cases} = \mathbb{1}[y_i \hat{y}_i \leq 0],$$

  gdzie $\mathbb{1}[C]$ to funkcja indykatorowa: `C ? 1 : 0`.

- Minimalizujemy:

$$\begin{aligned} \min : L &= \sum_{i=1}^{n} \mathrm{err}_i \\ &= \sum_{i=1}^{n} \mathbb{1}[y_i \hat{y}_i \leq 0] \\ &= \sum_{i=1}^{n} \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]. \end{aligned}$$

- Silnie nieliniowy problem z powodu funkcji indykatorowej.

Minimalizacja błędu klasyfikacji

- Prosty zapis błędu klasyfikacji ($y_i, \hat{y}_i \in \{-1, 1\}$):

$$\mathrm{err}_i = \begin{cases} 1 & \text{jeśli } y_i \hat{y}_i \leq 0 \\ 0 & \text{jeśli } y_i \hat{y}_i > 0 \end{cases} = \mathbb{1}[y_i \hat{y}_i \leq 0],$$

  gdzie $\mathbb{1}[C]$ to funkcja indykatorowa: `C ? 1 : 0`.

- Minimalizujemy:

$$\begin{aligned} \min : L &= \sum_{i=1}^{n} \mathrm{err}_i \\ &= \sum_{i=1}^{n} \mathbb{1}[y_i \hat{y}_i \leq 0] \\ &= \sum_{i=1}^{n} \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]. \end{aligned}$$

- Silnie nieliniowy problem z powodu funkcji indykatorowej.
- Można pokazać, że w ogólności problem jest NP-trudny!

Sprowadzenie do problemu liniowego całkowitoliczbowego

- Wprowadzamy zmienne binarne $\mathrm{err}_i \in \{0, 1\}$, które określają błędy klasyfikacji.

Sprowadzenie do problemu liniowego całkowitoliczbowego

- Wprowadzamy zmienne binarne $\mathrm{err}_i \in \{0, 1\}$, które określają błędy klasyfikacji.
- Minimalizujemy liniową funkcję $L = \sum_{i=1}^{n} \mathrm{err}_i$.

Sprowadzenie do problemu liniowego całkowitoliczbowego

- Wprowadzamy zmienne binarne $\mathrm{err}_i \in \{0, 1\}$, które określają błędy klasyfikacji.
- Minimalizujemy liniową funkcję $L = \sum_{i=1}^{n} \mathrm{err}_i$.
- Chcemy wprowadzić ograniczenia liniowe, które gwarantują, że $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$:

Sprowadzenie do problemu liniowego całkowitoliczbowego

- Wprowadzamy zmienne binarne $\mathrm{err}_i \in \{0, 1\}$, które określają błędy klasyfikacji.
- Minimalizujemy liniową funkcję $L = \sum_{i=1}^{n} \mathrm{err}_i$.
- Chcemy wprowadzić ograniczenia liniowe, które gwarantują, że $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$:

$$y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq M(1 - \mathrm{err}_i),$$
$$y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq -M\mathrm{err}_i + \epsilon,$$

gdzie M jest bardzo dużą liczbą, a ϵ bardzo małą liczbą.

Sprowadzenie do problemu liniowego całkowitoliczbowego

$$
\begin{aligned}
\min \quad & L = \sum_{i=1}^{n} \mathrm{err}_i \\
\text{p.o.} \quad & y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq M(1 - \mathrm{err}_i) && i = 1, \ldots, n \\
\text{p.o.} \quad & y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq -M\mathrm{err}_i + \epsilon && i = 1, \ldots, n \\
& \mathrm{err}_i \in \{0, 1\} && i = 1, \ldots, n
\end{aligned}
$$

Sprowadzenie do problemu liniowego całkowitoliczbowego

$$
\begin{aligned}
\min \quad & L = \sum_{i=1}^{n} \mathrm{err}_i \\
\text{p.o.} \quad & y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq M(1 - \mathrm{err}_i) && i = 1, \ldots, n \\
\text{p.o.} \quad & y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq -M\mathrm{err}_i + \epsilon && i = 1, \ldots, n \\
& \mathrm{err}_i \in \{0, 1\} && i = 1, \ldots, n
\end{aligned}
$$

Oczywiście, problem jest nadal NP-trudny ...

Spis treści

Błąd klasyfikacji

- Przypomnijmy, że błąd 0/1 ma postać $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$.

Błąd klasyfikacji

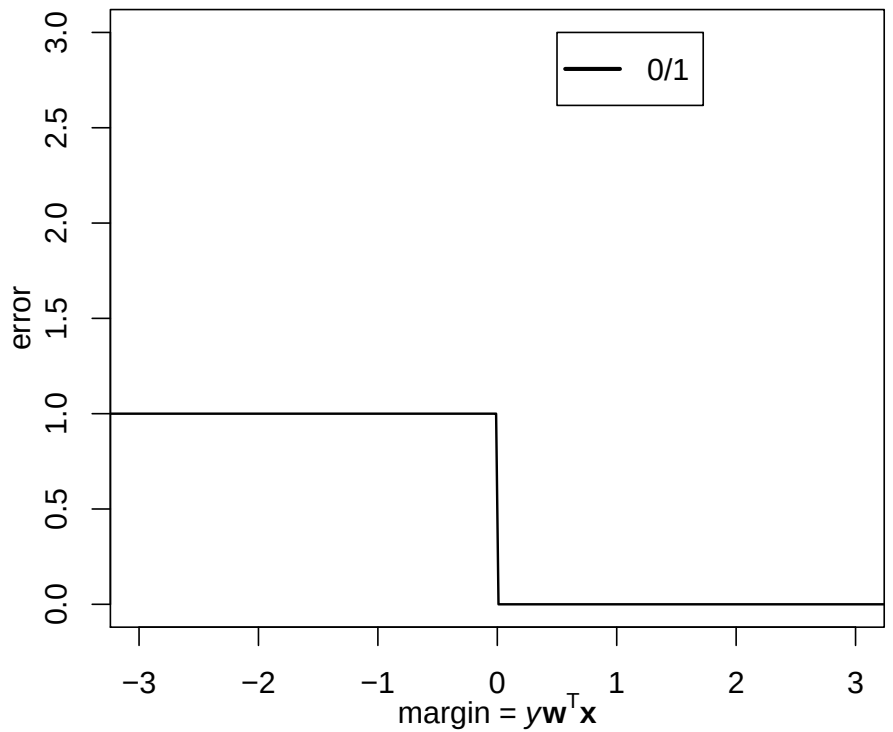
- Przypomnijmy, że błąd 0/1 ma postać $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$.
- Wprowadzamy pojęcie marginesu (*margin*) $\mathrm{marg}_i = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.

Błąd klasyfikacji

- Przypomnijmy, że błąd 0/1 ma postać $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$.
- Wprowadzamy pojęcie marginesu (*margin*) $\mathrm{marg}_i = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.
- Dodatni margines oznacza poprawną klasyfikację (brak błędu), ujemny margines oznacza niepoprawną klasyfikację (błąd).

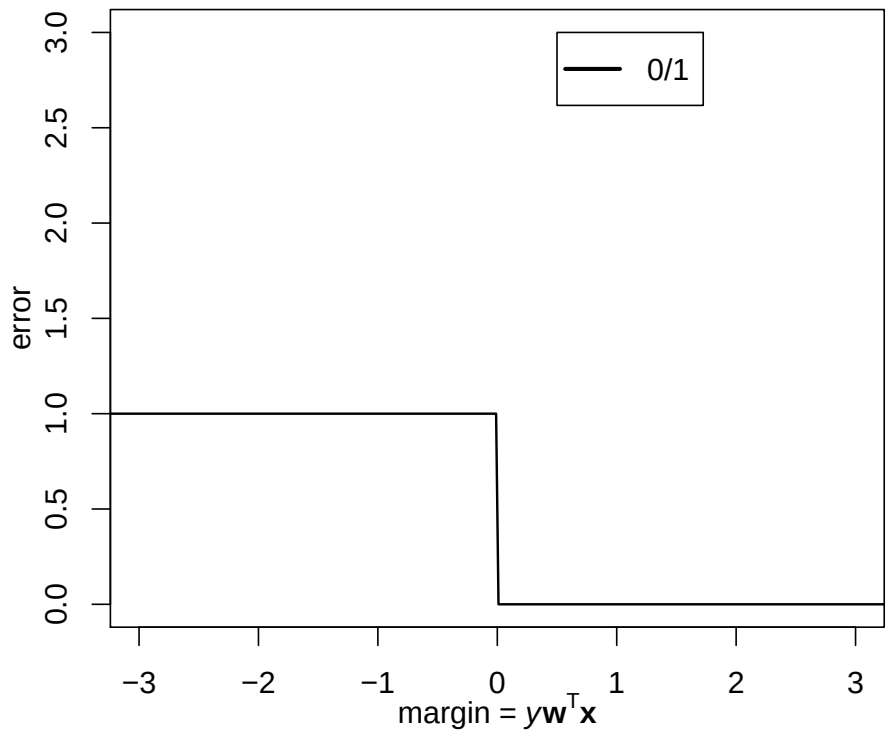
Błąd klasyfikacji

- Przypomnijmy, że błąd 0/1 ma postać $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$.
- Wprowadzamy pojęcie marginesu (*margin*) $\mathrm{marg}_i = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.
- Dodatni margines oznacza poprawną klasyfikację (brak błędu), ujemny margines oznacza niepoprawną klasyfikację (błąd).



Błąd klasyfikacji

- Przypomnijmy, że błąd 0/1 ma postać $\mathrm{err}_i = \mathbb{1}[y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 0]$.
- Wprowadzamy pojęcie marginesu (*margin*) $\mathrm{marg}_i = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.
- Dodatni margines oznacza poprawną klasyfikację (brak błędu), ujemny margines oznacza niepoprawną klasyfikację (błąd).



- Główny problem: nieciągły błąd klasyfikacji.

Relaksacja błędu klasyfikacji

Dokonujemy relaksacji nieciągłego błędu 0/1 do ciągłego błędu będącego nadal funkcją marginesu.

Relaksacja błędu klasyfikacji

Dokonujemy relaksacji nieciągłego błędu 0/1 do ciągłego błędu będącego nadal funkcją marginesu.

- Zrelaksowany błąd powinien być funkcją nieujemną i osiągać zero dla nieskończonego marginesu (chcemy, aby rozwiązanie bezbłędne miało błąd zerowy).

Relaksacja błędu klasyfikacji

Dokonujemy relaksacji nieciągłego błędu 0/1 do ciągłego błędu będącego nadal funkcją marginesu.

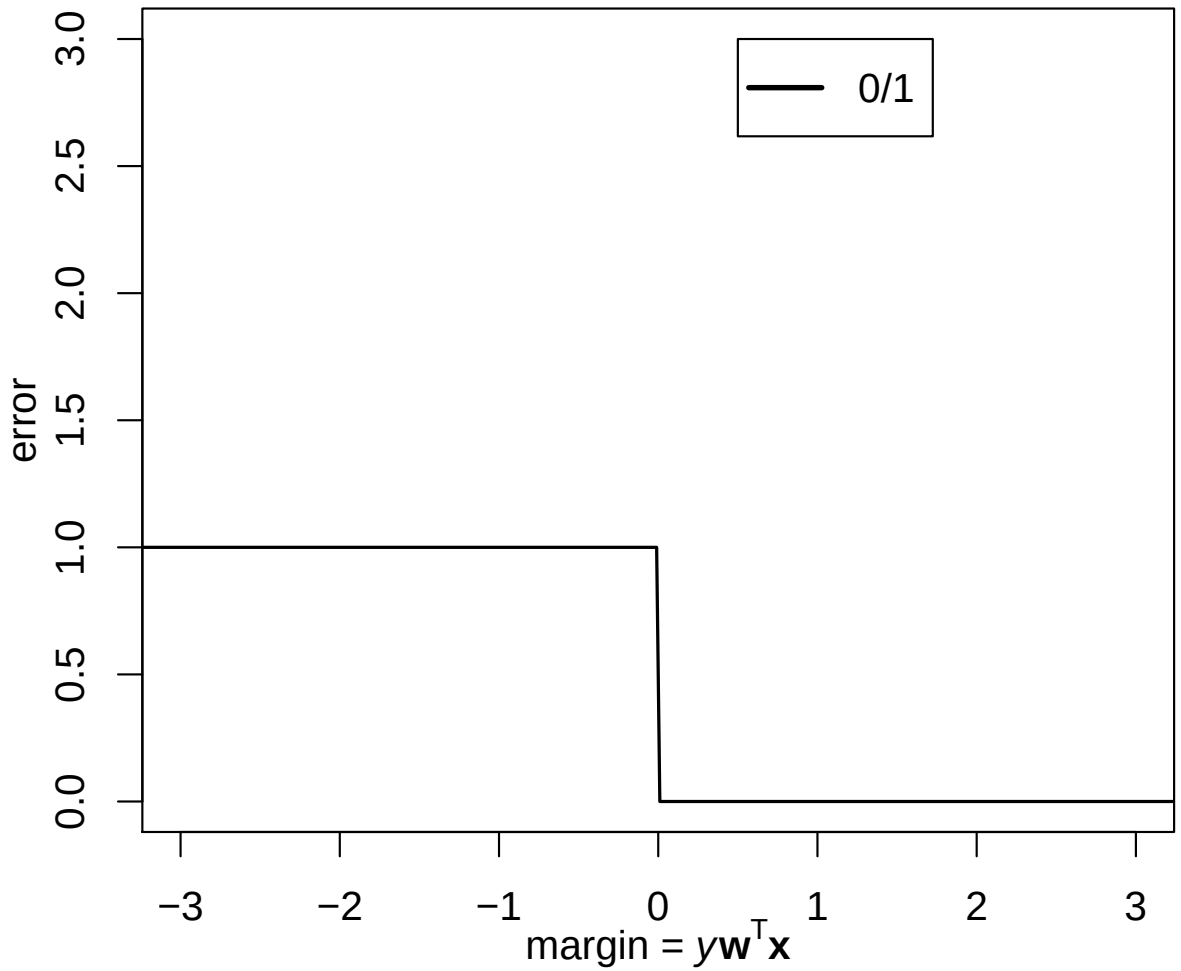
- Zrelaksowany błąd powinien być funkcją nieujemną i osiągać zero dla nieskończonego marginesu (chcemy, aby rozwiązanie bezbłędne miało błąd zerowy).
- Zrelaksowany błąd powinien być malejącą funkcją marginesu (im większy margines, tym mniejszy błąd).

Relaksacja błędu klasyfikacji

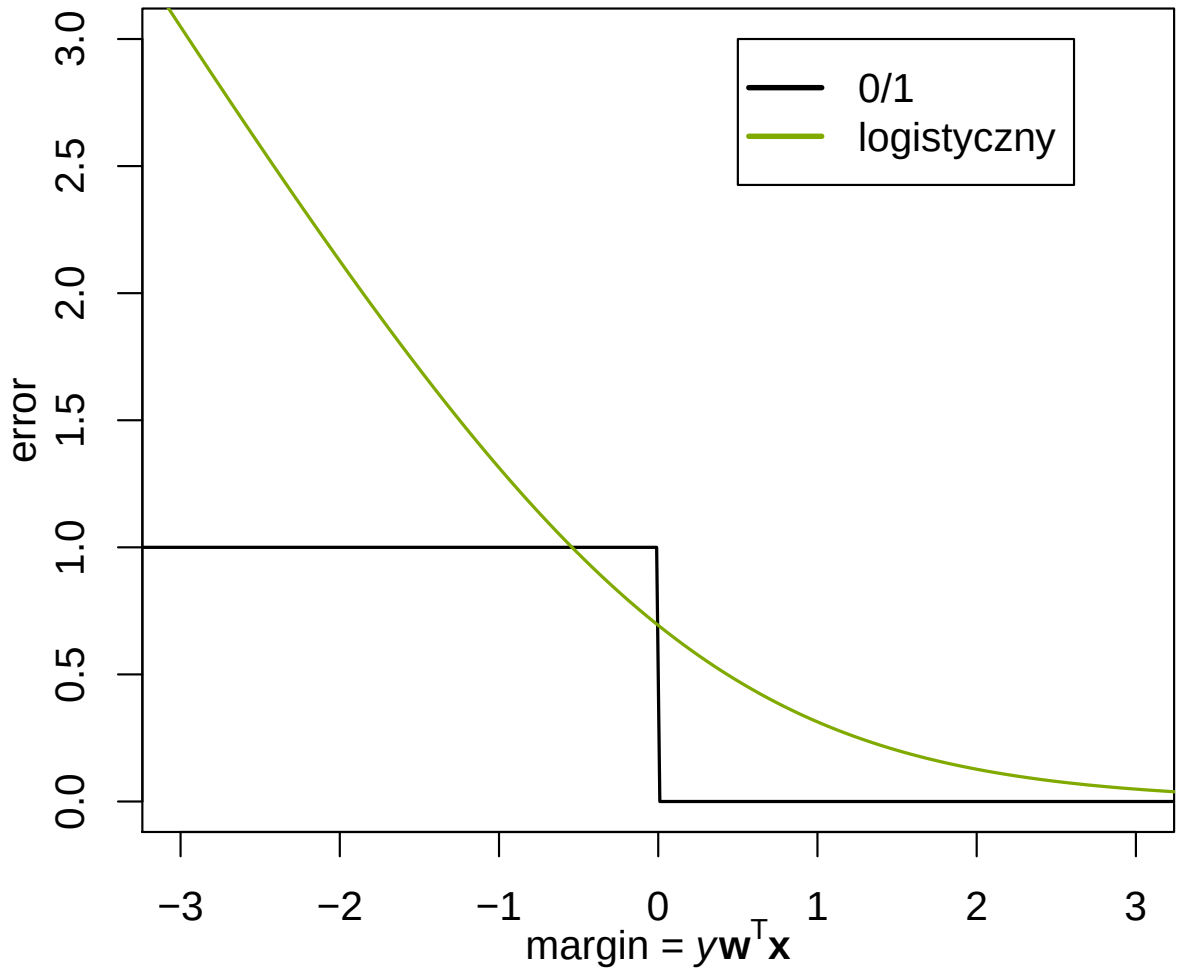
Dokonujemy relaksacji nieciągłego błędu 0/1 do ciągłego błędu będącego nadal funkcją marginesu.

- Zrelaksowany błąd powinien być funkcją nieujemną i osiągać zero dla nieskończonego marginesu (chcemy, aby rozwiązanie bezbłędne miało błąd zerowy).
- Zrelaksowany błąd powinien być malejącą funkcją marginesu (im większy margines, tym mniejszy błąd).
- Zrelaksowany błąd powinien być wypukłą funkcją marginesu (daje to problem o złożoności wielomianowej).

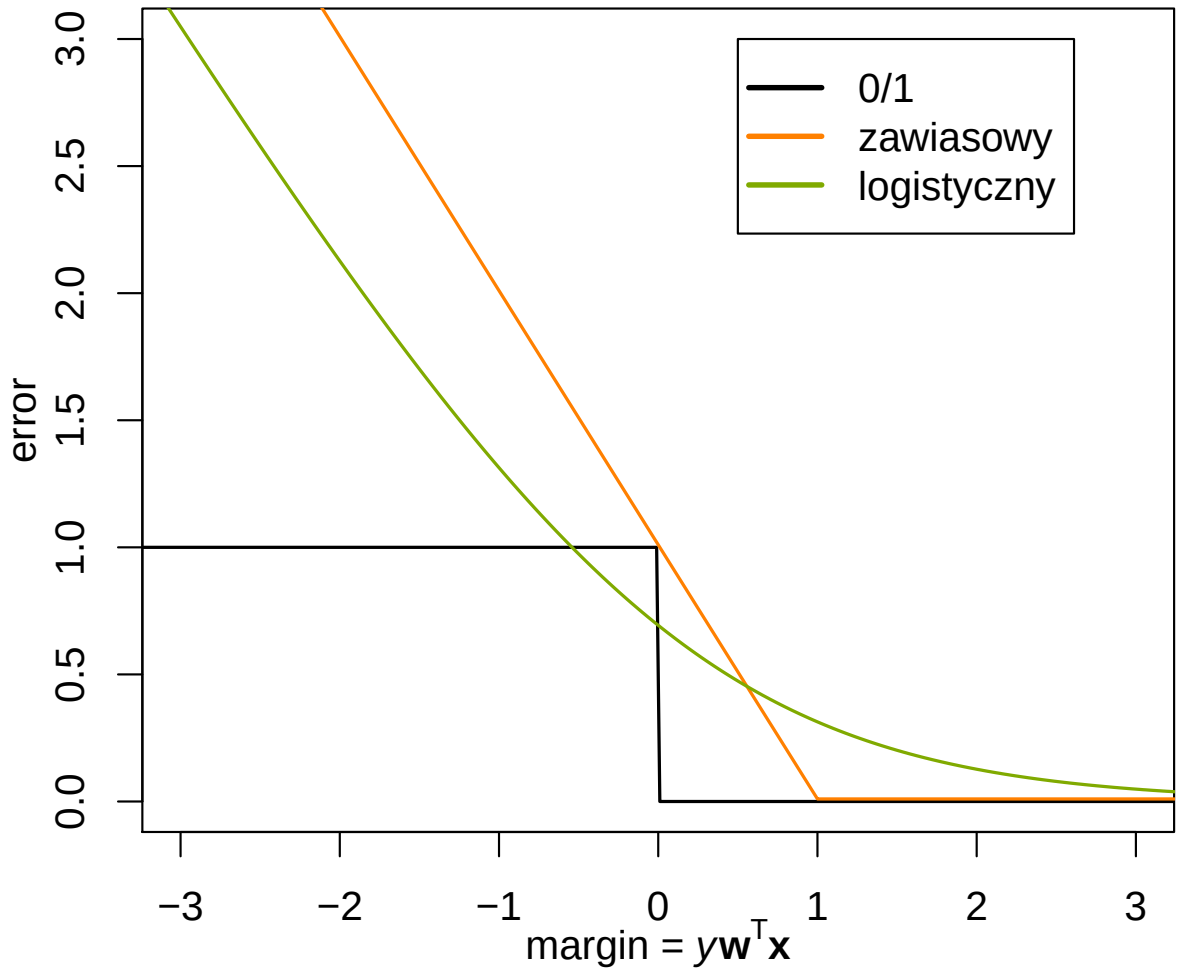
Relaksacja błędu klasyfikacji



Relaksacja błędu klasyfikacji



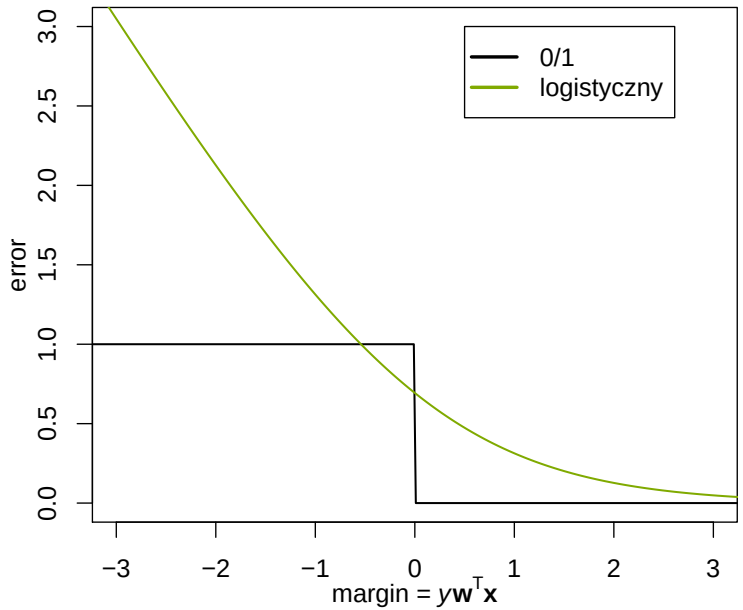
Relaksacja błędu klasyfikacji



Spis treści

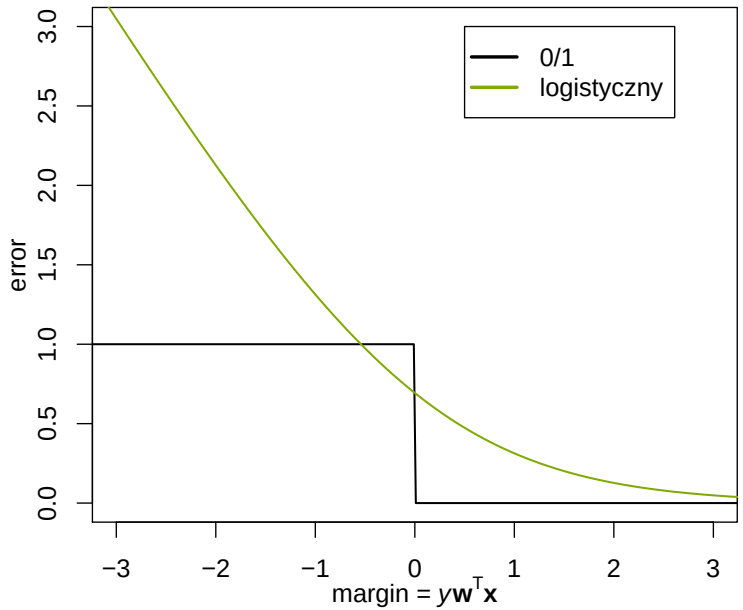
Błąd logistyczny

$$\mathrm{err}(\mathrm{margin}) = \log(1 + \exp(-\mathrm{marg})) = \log(1 + \exp(-y\boldsymbol{w}^\top \boldsymbol{x}))$$



Błąd logistyczny

$$\mathrm{err}(\mathrm{margin}) = \log(1 + \exp(-\mathrm{marg})) = \log(1 + \exp(-y\boldsymbol{w}^\top \boldsymbol{x}))$$



- Błąd logistyczny otrzymujemy poprzez:
  1. wzięcie funkcji liniowej $\boldsymbol{w}^\top \boldsymbol{x}$,
  2. przemnożenie przez prawdziwą etykietę klasy y, otrzymując margines $\mathrm{marg} = y\boldsymbol{w}^\top \boldsymbol{x}$,
  3. przekształcenie przez nieliniową funkcję $\log(1 + \exp(-\mathrm{marg}))$.

Dlaczego błąd logistyczny?

Dlaczego błąd logistyczny?

- Wynika z popularnego modelu statystycznego (szczegóły: wykład dr. Dembczyńskiego).

Dlaczego błąd logistyczny?

- Wynika z popularnego modelu statystycznego (szczegóły: wykład dr. Dembczyńskiego).
- Ma bardzo dobre własności:

Dlaczego błąd logistyczny?

- Wynika z popularnego modelu statystycznego (szczegóły: wykład dr. Dembczyńskiego).
- Ma bardzo dobre własności:
  - Jest funkcją wypukłą (prosta do optymalizacji = jedno minimum globalne).

Dlaczego błąd logistyczny?

- Wynika z popularnego modelu statystycznego (szczegóły: wykład dr. Dembczyńskiego).
- Ma bardzo dobre własności:
  - Jest funkcją wypukłą (prosta do optymalizacji = jedno minimum globalne).
  - Dla dużego dodatniego marginesu marg wykładniczo szybko zbiega do zera:

$$\log(1+\exp(-\mathrm{marg})) \simeq \exp(-\mathrm{marg}),$$

gdzie skorzystaliśmy z rozwinięcia Taylora $\log(1+x) \simeq x$ dla małych x.

Dlaczego błąd logistyczny?

- Wynika z popularnego modelu statystycznego (szczegóły: wykład dr. Dembczyńskiego).
- Ma bardzo dobre własności:
  - Jest funkcją wypukłą (prosta do optymalizacji = jedno minimum globalne).
  - Dla dużego dodatniego marginesu marg wykładniczo szybko zbiega do zera:

$$\log(1+\exp(-\mathrm{marg})) \simeq \exp(-\mathrm{marg}),$$

    gdzie skorzystaliśmy z rozwinięcia Taylora $\log(1+x) \simeq x$ dla małych x.
  - Dla silnie ujemnego marginesu marg rośnie liniowo:

$$\log(1+\exp(-\mathrm{marg})) \simeq \log(\exp(-\mathrm{marg})) = -\mathrm{marg}.$$

    Odporna na wartości odstające!

Regresja logistyczna

Regresja logistyczna

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

$$\min : L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

Regresja logistyczna

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

$$\min : L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Ciągła (i wypukła) funkcja bez ograniczeń.

Regresja logistyczna

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

$$\min : L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Ciągła (i wypukła) funkcja bez ograniczeń.
- Możemy użyć jednej z poznanych metod optymalizacji poznanych na wykładzie.

Regresja logistyczna

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

$$\min : L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Ciągła (i wypukła) funkcja bez ograniczeń.
- Możemy użyć jednej z poznanych metod optymalizacji poznanych na wykładzie.
- Po optymalizacji współczynników, klasyfikacja nowych obserwacji używając funkcji:

$$\hat{Y} = \operatorname{sgn}(\boldsymbol{w}^\top \boldsymbol{X})$$

Spis treści

Metoda Cauchy'ego (spadek wzdłuż gradientu)

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.

Metoda Cauchy'ego (spadek wzdłuż gradientu)

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności

Metoda Cauchy'ego (spadek wzdłuż gradientu)

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności
   - Wyznaczamy gradient w punkcie $\boldsymbol{w}_{k-1}$, $\nabla_L(\boldsymbol{w}_{k-1})$.

Metoda Cauchy'ego (spadek wzdłuż gradientu)

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności
   - Wyznaczamy gradient w punkcie $\boldsymbol{w}_{k-1}$, $\nabla_L(\boldsymbol{w}_{k-1})$.
   - Robimy krok wzdłuż negatywnego gradientu:

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} - \alpha_k \nabla_L(\boldsymbol{w}_{k-1}),$$

   gdzie α_k jest długością kroku ustaloną przez przeszukiwanie liniowe.

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} \frac{-y_i x_{ij} \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} \frac{-y_i x_{ij} \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

- Gradient:

$$\nabla_L(\boldsymbol{w}) = -\sum_{i=1}^{n} \frac{y_i \boldsymbol{x}_i}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

- Krok wzdłuż gradientu:

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \alpha_k \sum_{i=1}^{n} \frac{y_i \boldsymbol{x}_i}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

Interpretacja kroku wzdłuż gradientu

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \alpha_k \sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i, \qquad \beta_i = \frac{1}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$$

Interpretacja kroku wzdłuż gradientu

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \alpha_k \sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i, \qquad \beta_i = \frac{1}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$$

Kierunek kroku determinowany przez sumę „wpływów obserwacji” o postaci $y_i \boldsymbol{x}_i \beta_i$.

Interpretacja kroku wzdłuż gradientu

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \alpha_k \sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i, \qquad \beta_i = \frac{1}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$$

Kierunek kroku determinowany przez sumę „wpływów obserwacji” o postaci $y_i \boldsymbol{x}_i \beta_i$.

- jeśli $y_i = 1$, to kierunek jest w stronę wektora $\boldsymbol{x}_i$

Interpretacja kroku wzdłuż gradientu

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \alpha_k \sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i, \qquad \beta_i = \frac{1}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$$

Kierunek kroku determinowany przez sumę „wpływów obserwacji” o postaci $y_i \boldsymbol{x}_i \beta_i$.

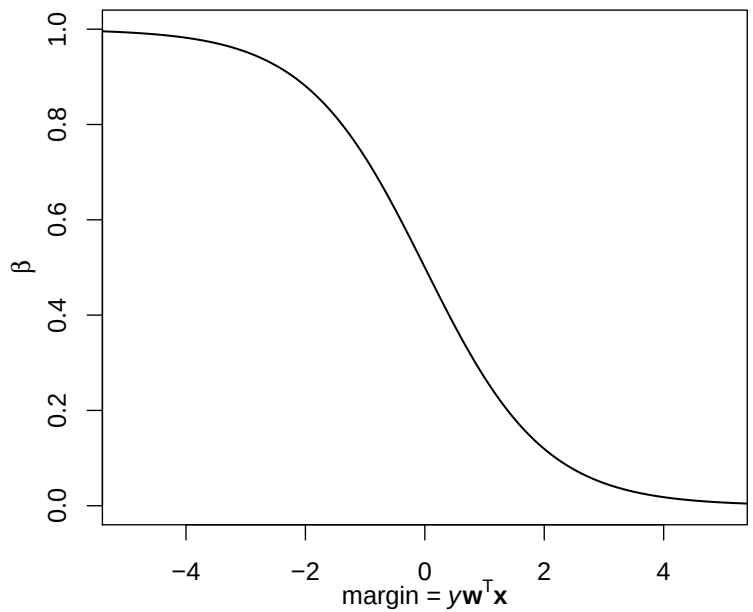
- jeśli $y_i = 1$, to kierunek jest w stronę wektora $\boldsymbol{x}_i$
- jeśli $y_i = -1$, to kierunek jest w stronę przeciwną do $\boldsymbol{x}_i$

Interpretacja kroku wzdłuż gradientu

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \alpha_k \sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i, \qquad \beta_i = \frac{1}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$$

Kierunek kroku determinowany przez sumę „wpływów obserwacji” o postaci $y_i \boldsymbol{x}_i \beta_i$.

- jeśli $y_i = 1$, to kierunek jest w stronę wektora $\boldsymbol{x}_i$
- jeśli $y_i = -1$, to kierunek jest w stronę przeciwną do $\boldsymbol{x}_i$
- Wpływ danego wektora $\boldsymbol{x}_i$ zależy od wartości β_i:



Spis treści

Metoda Newtona

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.

Metoda Newtona

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności

Metoda Newtona

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności
   - Wyznaczamy gradient w punkcie $\boldsymbol{w}_{k-1}$, $\nabla_L(\boldsymbol{w}_{k-1})$.

Metoda Newtona

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności
   - Wyznaczamy gradient w punkcie $\boldsymbol{w}_{k-1}$, $\nabla_L(\boldsymbol{w}_{k-1})$.
   - Wyznaczamy hesjan w punkcie $\boldsymbol{w}_{k-1}$, $\boldsymbol{H}_L(\boldsymbol{w}_{k-1})$.

Metoda Newtona

1. Zaczynamy od rozwiązania $\boldsymbol{w}_0 = \mathbf{0}$.
2. Dla $k = 1, 2, \ldots$ aż do zbieżności
   - Wyznaczamy gradient w punkcie $\boldsymbol{w}_{k-1}$, $\nabla_L(\boldsymbol{w}_{k-1})$.
   - Wyznaczamy hesjan w punkcie $\boldsymbol{w}_{k-1}$, $\boldsymbol{H}_L(\boldsymbol{w}_{k-1})$.
   - Robimy krok:

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} - \left(\boldsymbol{H}_L(\boldsymbol{w}_{k-1})\right)^{-1} \nabla_L(\boldsymbol{w}_{k-1}),$$

Metoda Newtona

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

Metoda Newtona

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} \frac{-y_i x_{ij} \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

Metoda Newtona

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} \frac{-y_i x_{ij} \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

- Drugie pochodne:

$$\frac{\partial^2 L(\boldsymbol{w})}{\partial w_j w_k} = \sum_{i=1}^{n} \frac{y_i y_i x_{ij} x_{ik} \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2}.$$

Metoda Newtona

- Funkcja celu:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)).$$

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} \frac{-y_i x_{ij} \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}.$$

- Drugie pochodne:

$$\frac{\partial^2 L(\boldsymbol{w})}{\partial w_j w_k} = \sum_{i=1}^{n} \frac{y_i y_i x_{ij} x_{ik} \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2}.$$

- Hesjan:

$$\boldsymbol{H}_L(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1 + \exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2}.$$

Interpretacja kroku Newtona

Interpretacja kroku Newtona

Wprowadzając ponownie $\beta_i = \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$, zauważamy, że:

$$\begin{aligned}\frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2} &= \frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))} \cdot \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} \\ &= \beta_i(1-\beta_i)\end{aligned}$$

Interpretacja kroku Newtona

Wprowadzając ponownie $\beta_i = \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$, zauważamy, że:

$$\begin{aligned}\frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2} &= \frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))} \cdot \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} \\ &= \beta_i(1-\beta_i)\end{aligned}$$

- Gradient:

$$\nabla_L(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \boldsymbol{x}_i \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = -\sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i.$$

Interpretacja kroku Newtona

Wprowadzając ponownie $\beta_i = \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$, zauważamy, że:

$$\begin{aligned}\frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2} &= \frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))} \cdot \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} \\ &= \beta_i(1-\beta_i)\end{aligned}$$

- Gradient:

$$\nabla_L(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \boldsymbol{x}_i \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = -\sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i.$$

- Hesjan:

$$\boldsymbol{H}_L(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \frac{\exp(y_i w^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \beta_i(1-\beta_i)$$

Interpretacja kroku Newtona

Wprowadzając ponownie $\beta_i = \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}$, zauważamy, że:

$$\begin{aligned}\frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2} &= \frac{\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))} \cdot \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} \\ &= \beta_i(1-\beta_i)\end{aligned}$$

- Gradient:

$$\nabla_L(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \boldsymbol{x}_i \frac{1}{1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} = -\sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i.$$

- Hesjan:

$$\boldsymbol{H}_L(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \frac{\exp(y_i w^\top \boldsymbol{x}_i)}{(1+\exp(y_i \boldsymbol{w}^\top \boldsymbol{x}_i))^2} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \beta_i(1-\beta_i)$$

- Krok Newtona:

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \beta_i(1-\beta_i)\right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i\right)$$

Interpretacja kroku Newtona

- Regresja liniowa:

$$\boldsymbol{w}^* = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \right)$$

- Regresja logistyczna – krok Newtona.

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \beta_i (1 - \beta_i) \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i \right)$$

Interpretacja kroku Newtona

- Regresja liniowa:

$$\boldsymbol{w}^* = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \right)$$

- Regresja logistyczna – krok Newtona.

$$\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \beta_i (1 - \beta_i) \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \beta_i \right)$$

- Rozwiązanie regresji logistycznej przez metodę Newtona to wielokrotne rozwiązywanie problemu najmniejszych kwadratów z obserwacjami ważonymi przez β_i!

Spis treści

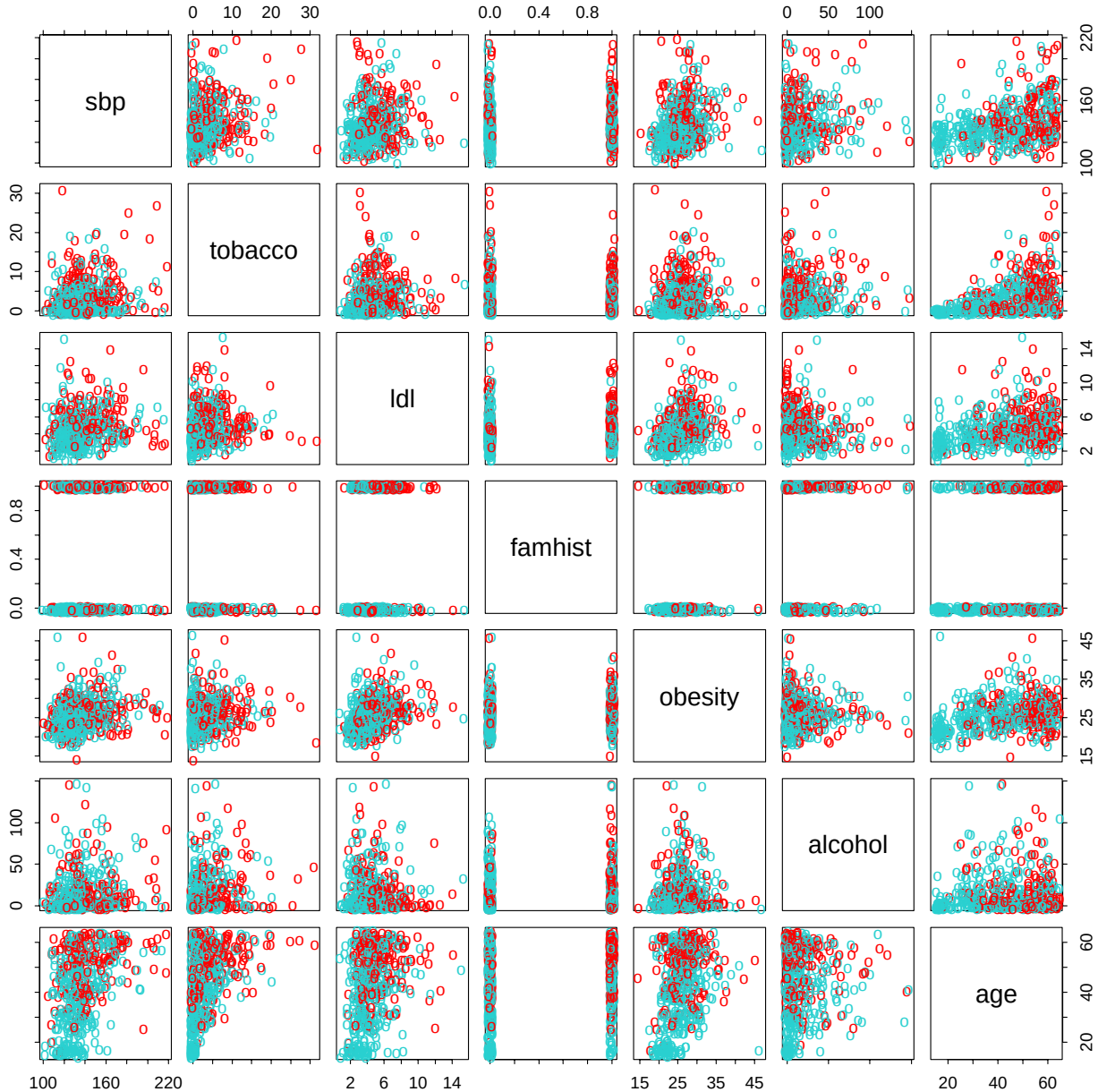
South African heart disease

- 462 obserwacje, 160 chorych i 302 obserwacji kontrolnych.
- 7 cech wejściowych ($\boldsymbol{X}$) i jedna binarna cecha wyjściowa (Y – *zdrowy/chory*).

SBP	*systolic blood pressures*
TOBACCO	*cumulative tobacco (kg)*
LDL	*low density lipoprotein cholesterol*
FAMHIST	*family history of heart disease (present/absent)*
OBESITY	*obesity (bmi)*
ALCOHOL	*alcohol consumption (ltr)*
AGE	*age (years)*
DISEASE	*presence of heart disease (no/yes)*

Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*

South African heart disease

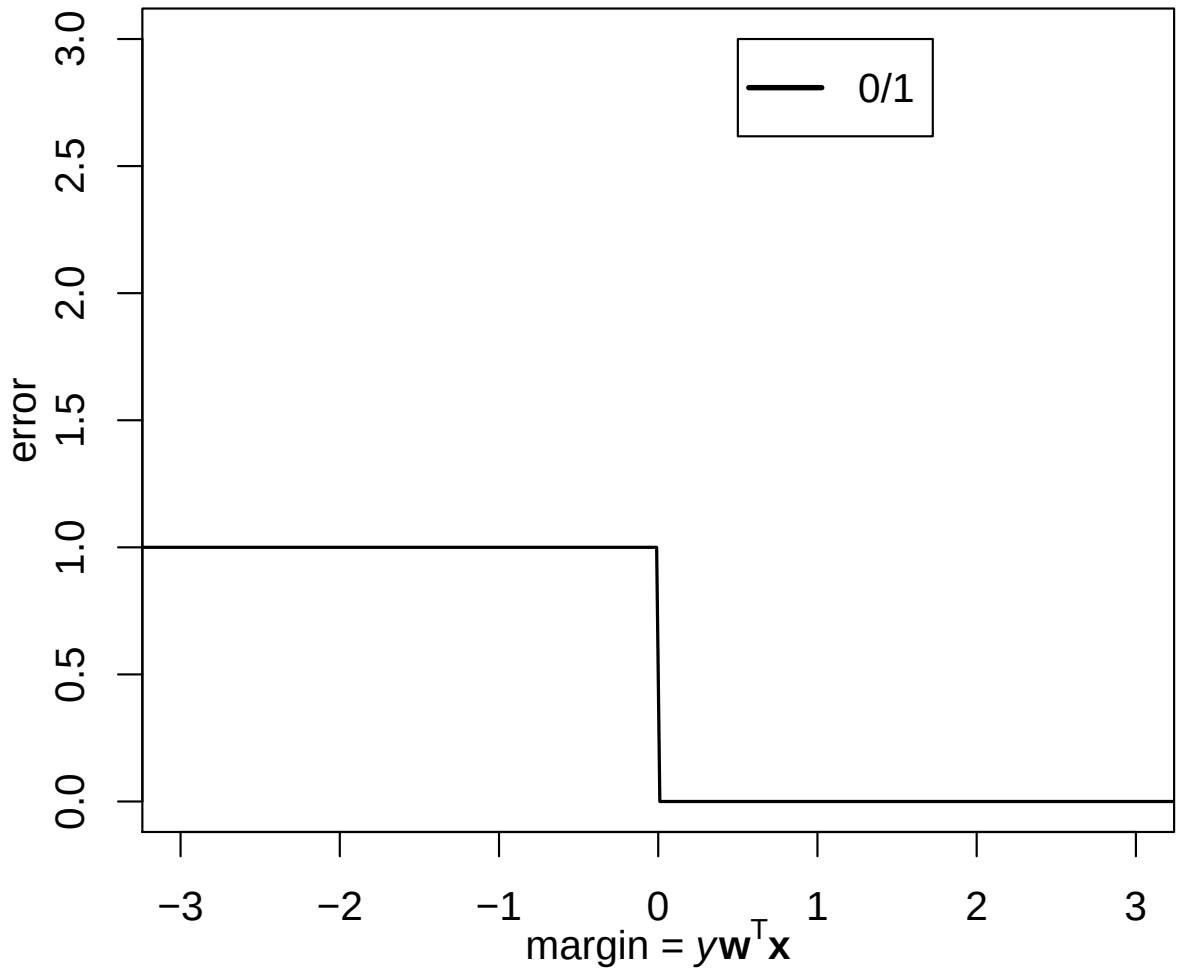


South African heart disease

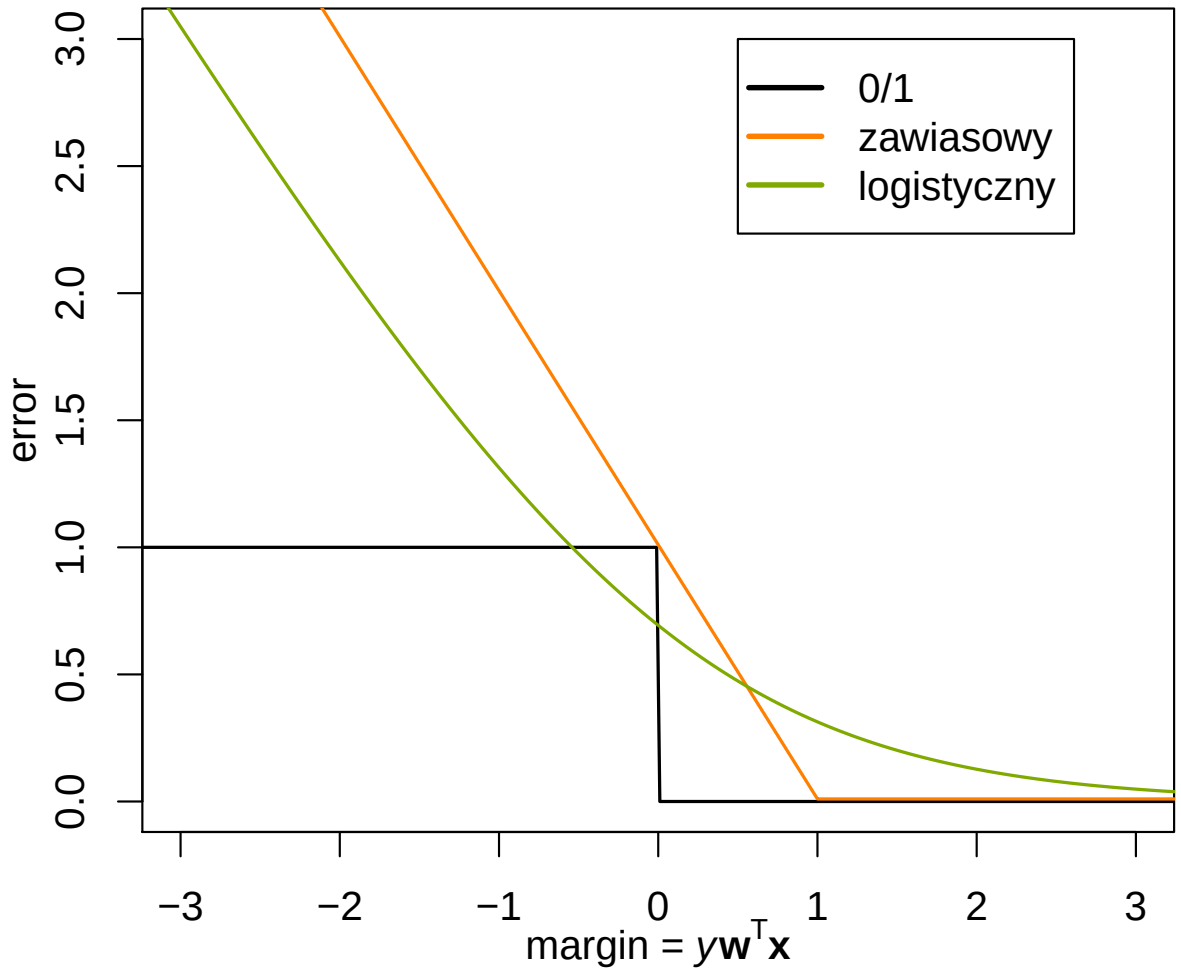
feature	coefficient
(intercept)	-4.130
SBP	0.006
TOBACCO	0.080
LDL	0.185
FAMHIST	0.939
OBESITY	-0.035
ALCOHOL	0.001
AGE	0.043

Spis treści

Relaksacja błędu klasyfikacji



Relaksacja błędu klasyfikacji



Zawiasowa funkcja straty

$$\text{err}(\text{margin}) = \begin{cases} 0 & \text{jeśli} \quad \text{margin} > 1 \\ 1 - \text{margin} & \text{jeśli} \quad \text{margin} \leq 1 \end{cases} = (1-\text{margin})_+$$

gdzie:

$$(a)_+ = \begin{cases} 0 & \text{jeśli} \quad a \leq 0 \\ a & \text{jeśli} \quad a > 0. \end{cases}$$

Zawiasowa funkcja straty

$$\mathrm{err}(\mathrm{margin}) = \begin{cases} 0 & \text{jeśli} \quad \mathrm{margin} > 1 \\ 1 - \mathrm{margin} & \text{jeśli} \quad \mathrm{margin} \leq 1 \end{cases} = (1 - \mathrm{margin})_+$$

gdzie:

$$(a)_+ = \begin{cases} 0 & \text{jeśli} \quad a \leq 0 \\ a & \text{jeśli} \quad a > 0. \end{cases}$$

- Wartość 1 jest tu wzięta arbitralnie – cokolwiek powyżej zera zadziała. Dlaczego?

Zawiasowa funkcja straty

$$\mathrm{err}(\mathrm{margin}) = \begin{cases} 0 & \text{jeśli} \quad \mathrm{margin} > 1 \\ 1 - \mathrm{margin} & \text{jeśli} \quad \mathrm{margin} \leq 1 \end{cases} = (1 - \mathrm{margin})_+$$

gdzie:

$$(a)_+ = \begin{cases} 0 & \text{jeśli} \quad a \leq 0 \\ a & \text{jeśli} \quad a > 0. \end{cases}$$

- Wartość 1 jest tu wzięta arbitralnie – cokolwiek powyżej zera zadziała. Dlaczego?
- Ponieważ wagi można przeskalować. Błąd w zerze musi być jednak niezerowy, inaczej trywialne rozwiązanie $\boldsymbol{w} = \mathbf{0}$ byłoby natychmiast optymalne.

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

$$\min_{\boldsymbol{w}} : \; L(\boldsymbol{w}) = \sum_{i=1}^{n} (1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+$$

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

$$\min_{\boldsymbol{w}} : \; L(\boldsymbol{w}) = \sum_{i=1}^{n} (1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+$$

Sprowadzamy do problemu programowania liniowego:

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

$$\min_{\boldsymbol{w}} : \; L(\boldsymbol{w}) = \sum_{i=1}^{n} (1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego $i = 1, \ldots, n$ wprowadzamy zmienną $\mathrm{err}_i \geq 0$ i chcemy zapewnić, że

$$(1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+ = \mathrm{err}_i$$

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

$$\min_{\boldsymbol{w}} : \; L(\boldsymbol{w}) = \sum_{i=1}^{n} (1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego $i = 1, \ldots, n$ wprowadzamy zmienną $\mathrm{err}_i \geq 0$ i chcemy zapewnić, że

$$(1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+ = \mathrm{err}_i$$

- Wystarczy zapewnić, że

$$1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i \leq \mathrm{err}_i,$$

  ponieważ będziemy minimalizowali błędy err_i.

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

$$\min_{\boldsymbol{w}} : \; L(\boldsymbol{w}) = \sum_{i=1}^{n} (1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego $i = 1, \ldots, n$ wprowadzamy zmienną $\mathrm{err}_i \geq 0$ i chcemy zapewnić, że

$$(1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i)_+ = \mathrm{err}_i$$

- Wystarczy zapewnić, że

$$1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i \leq \mathrm{err}_i,$$

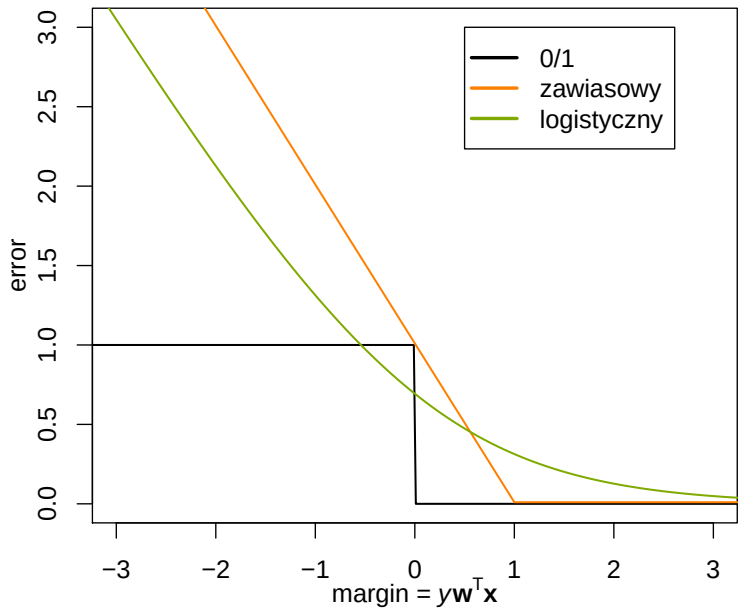
  ponieważ będziemy minimalizowali błędy err_i.
- Dowód:
  - jeśli w rozwiązaniu optymalnym $1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i < \mathrm{err}_i$, to oznacza, że $\mathrm{err}_i = 0$.
  - Inaczej, gdyby $\mathrm{err}_i > 0$, możemy zmniejszyć err_i, co przeczyłoby optymalności.

Optymalizacja

Rozwiązujemy problem:

$$\begin{aligned} \min \quad & L = \sum_{i=1}^{n} \mathrm{err}_i \\ \text{p.o.} \quad & 1 - y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i \leq \mathrm{err}_i && i = 1, \ldots, n \\ & \mathrm{err}_i \geq 0 && i = 1, \ldots, n. \end{aligned}$$

Regresja logistyczna vs. błąd zawiasowy



- Oba błędy są do siebie bardzo podobne.
- Obie funkcje błędu prowadzą do dobrych klasyfikatorów i żaden błąd nie wydaje się istotnie lepszy od drugiego.
- Regresja logistyczna popularniejsza.
- Regresja logistyczna daje oszacowanie niepewności predykcji (wykład dr. Dembczyńskiego).

Koniec na dzisiaj :)