

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: wtorek 15:00-16:30
Slajdy dostępne pod adresem: <http://www.cs.put.poznan.pl/wkotlowski/>

25.11.2013

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

Problem klasyfikacji

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

Problem klasyfikacji

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

Problem klasyfikacji

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

Problem klasyfikacji

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

Przykłady

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

Przykład



- Zbiór uczący: IRIS (Ronald Fisher, 1936)
- Zmienne wejściowe (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



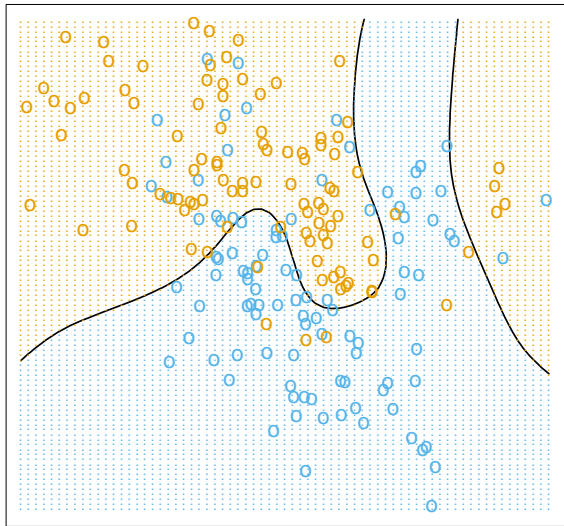
- Zbiór uczący: IRIS (Ronald Fisher, 1936)
- Zmienne wejściowe (X): *sepal-length*, *sepal-width*, *petal-length*, *petal-width*
- Zmienna wyjściowa (Y): *type*

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

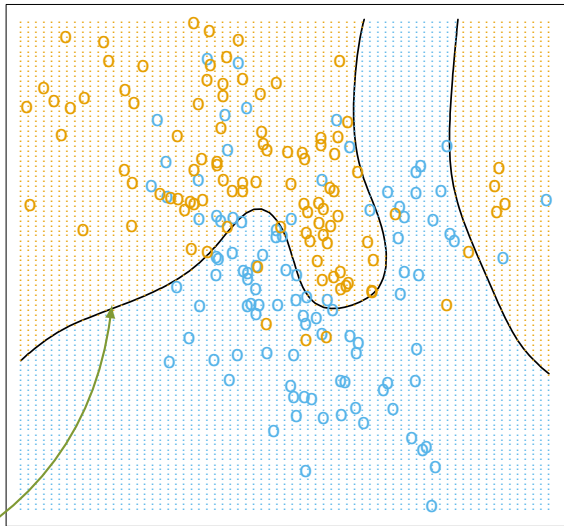
Przykład – 2 klasy

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



Przykład – 2 klasy

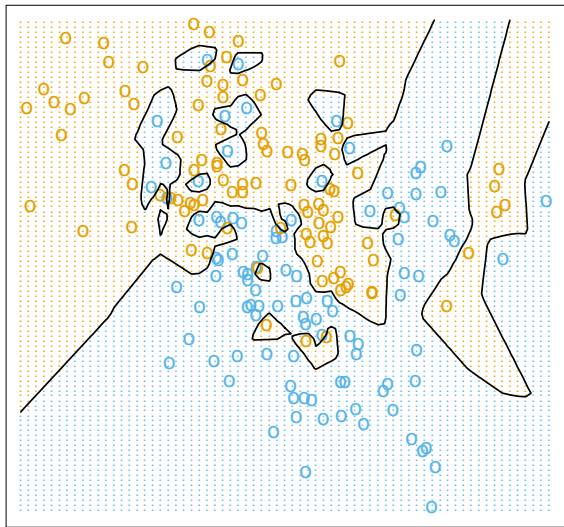
Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



Granica między klasami jest nazywana **granicą decyzyjną**

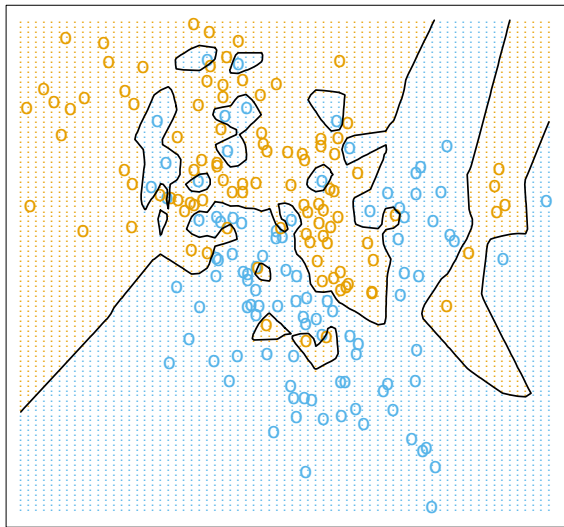
Przykład – 2 klasy

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



Przykład – 2 klasy

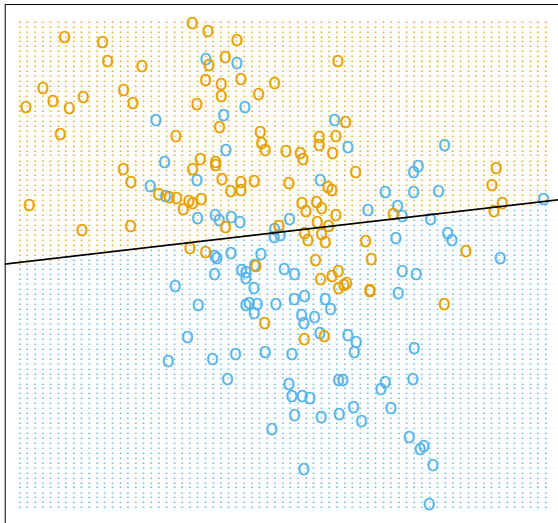
Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



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

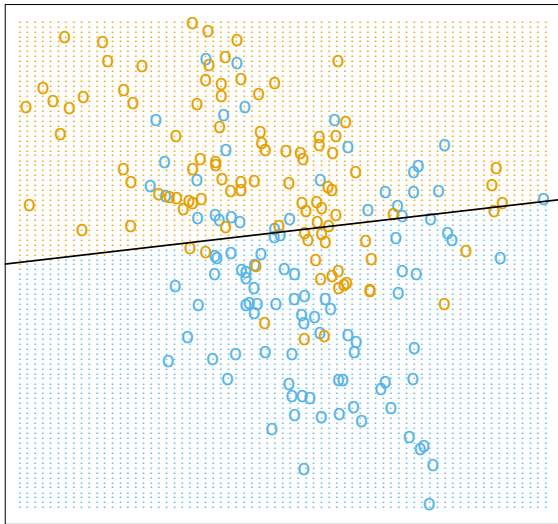
Przykład – 2 klasy

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



Przykład – 2 klasy

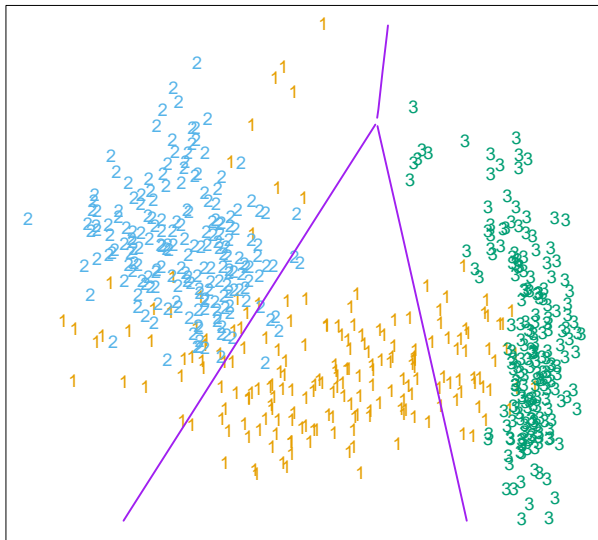
Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*



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 \mathbf{X} :

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

Klasyfikacja liniowa – 2 klasy

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

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

- Klasyfikacja poprzez progowanie w zerze:

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

Klasyfikacja liniowa – 2 klasy

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

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

- Klasyfikacja poprzez progowanie w zerze:

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

- Dla większej ilości klas zwykle rozbić 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 \mathbf{X} :

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

- Klasyfikacja poprzez progowanie w zerze:

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

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

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

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

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

Trenowanie klasyfikatora liniowego

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

Podójście I: Minimalizacja błędu klasyfikacji

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

Minimalizacja błędu klasyfikacji

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

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

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

Minimalizacja błędu klasyfikacji

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

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

gdzie $\mathbb{1}[C]$ to funkcja indykatorowa: $\mathbb{1}[C] = 1$ jeśli C jest prawdziwa, 0 w przeciwnym razie.

- Minimalizujemy:

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

Minimalizacja błędu klasyfikacji

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

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

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

- Minimalizujemy:

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

- Silnie **nieliniowy** problem z powodu funkcji indykatorowej.

Minimalizacja błędu klasyfikacji

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

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

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

- Minimalizujemy:

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

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

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

Srowadzenie do problemu liniowego całkowitoliczbowego

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

Sprowadzenie do problemu liniowego całkowitoliczbowego

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

Sprowadzenie do problemu liniowego całkowitoliczbowego

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

$$y_i \mathbf{w}^\top \mathbf{x}_i \leq M(1 - \text{err}_i),$$

$$y_i \mathbf{w}^\top \mathbf{x}_i \geq -M\text{err}_i + \epsilon,$$

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

Sprowadzenie do problemu liniowego całkowitoliczbowego

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

Sprowadzenie do problemu liniowego całkowitoliczbowego

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

Oczywiście, problem jest nadal NP-trudny ...

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

Błąd klasyfikacji

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

Błąd klasyfikacji

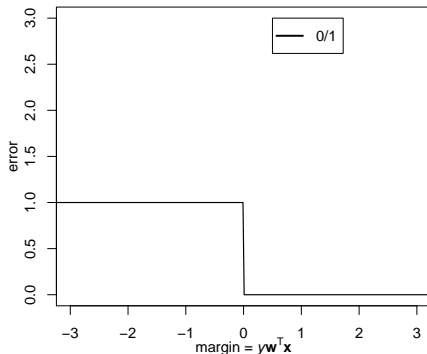
- Przypomnijmy, że błąd 0/1 ma postać $\text{err}_i = \mathbb{1}[y_i \mathbf{w}^\top \mathbf{x}_i \leq 0]$.
- Wprowadzamy pojęcie **marginesu** (*margin*) $\text{marg}_i = y_i \mathbf{w}^\top \mathbf{x}_i$.

Błąd klasyfikacji

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

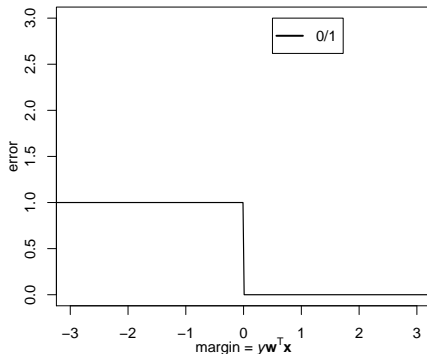
Błąd klasyfikacji

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



Błąd klasyfikacji

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



- Główny problem: **nieciągły** błąd klasyfikacji.

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

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).

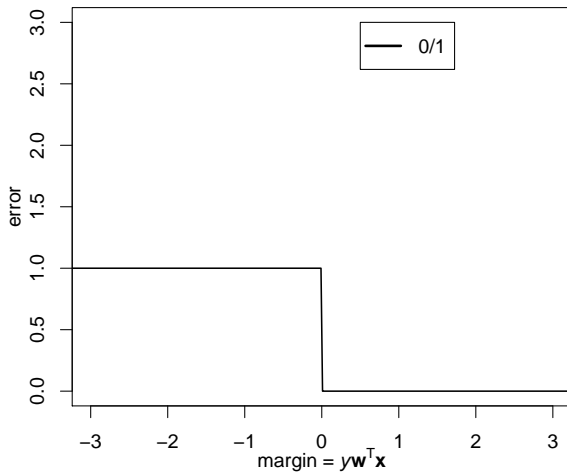
Dokonujemy relaksacji nieciągłego błędu 0/1 do ciągłego błędu będącego nadal funkcją marginesu.

- 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).

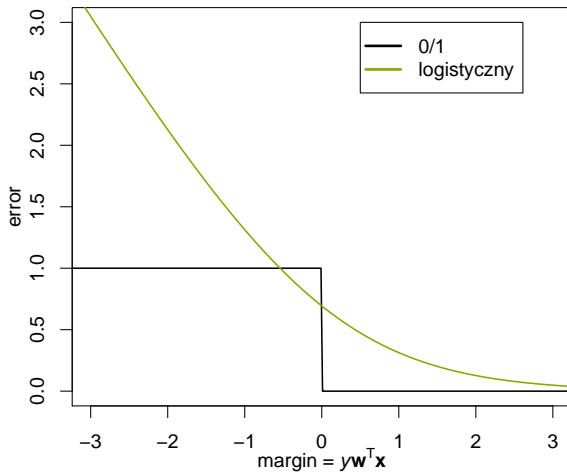
Dokonujemy relaksacji nieciągłego błędu 0/1 do ciągłego błędu będącego nadal funkcją marginesu.

- 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).

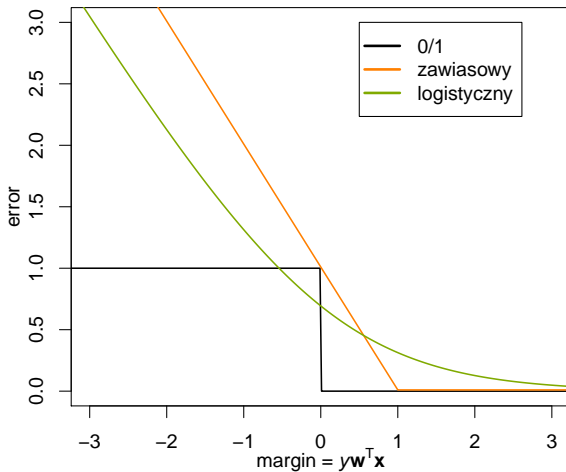
Relaksacja błędu klasyfikacji



Relaksacja błędu klasyfikacji



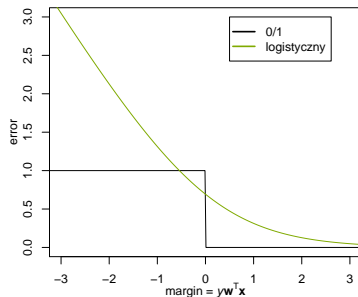
Relaksacja błędu klasyfikacji



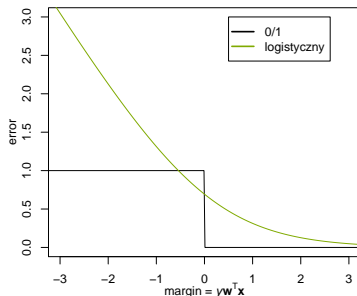
- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna**
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

Błąd logistyczny

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



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



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

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:

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

gdzie skorzystaliśmy z rozwinięcia Taylor $\log(1 + x) \simeq x$ 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:

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

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

- Dla silnie ujemnego marginesu marg rośnie **liniowo**:

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

Odporna na wartości odstające!

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

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

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

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

- Ciągła (i wypukła) funkcja bez ograniczeń.

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

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

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

- Minimalizacja całkowitego błędu logistycznego na zbiorze uczącym.

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

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

$$\hat{Y} = \text{sgn}(\mathbf{w}^\top \mathbf{X})$$

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- 1 Zaczynamy od rozwiązania $w_0 = 0$.

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- 1 Zaczynamy od rozwiązania $w_0 = 0$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności

Metoda Cauchy'ego (spadek wzdłuż gradientu)

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

Metoda Cauchy'ego (spadek wzdłuż gradientu)

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

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

gdzie α_k jest długością kroku ustaloną przez przeszukiwanie liniowe.

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- Funkcja celu:

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

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- Funkcja celu:

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

- Pochodne:

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

Metoda Cauchy'ego (spadek wzdłuż gradientu)

- Funkcja celu:

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

- Pochodne:

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

- Gradient:

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

- Krok wzdłuż gradientu:

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

Interpretacja kroku wzdłuż gradientu

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

Interpretacja kroku wzdłuż gradientu

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

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

Interpretacja kroku wzdłuż gradientu

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

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

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

Interpretacja kroku wzdłuż gradientu

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

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

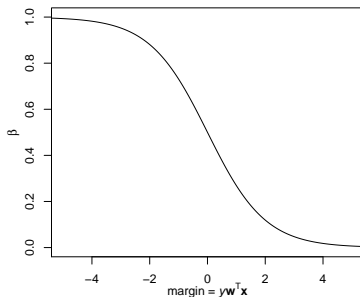
- jeśli $y_i = 1$, to kierunek jest **w stronę** wektora \mathbf{x}_i
- jeśli $y_i = -1$, to kierunek jest **w stronę przeciwną do** \mathbf{x}_i

Interpretacja kroku wzdłuż gradientu

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

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

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



- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

- 1 Zaczynamy od rozwiązania $w_0 = \mathbf{0}$.

- 1 Zaczynamy od rozwiązania $w_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności

- 1 Zaczynamy od rozwiązania $w_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wyznaczamy gradient w punkcie w_{k-1} , $\nabla_L(w_{k-1})$.

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

- 1 Zaczynamy od rozwiązania $w_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wyznaczamy gradient w punkcie w_{k-1} , $\nabla_L(w_{k-1})$.
 - Wyznaczamy hesjan w punkcie w_{k-1} , $H_L(w_{k-1})$.
 - Robimy krok:

$$w_k = w_{k-1} - (H_L(w_{k-1}))^{-1} \nabla_L(w_{k-1}),$$

- Funkcja celu:

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

- Funkcja celu:

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

- Pochodne:

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

- Funkcja celu:

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

- Pochodne:

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

- Drugie pochodne:

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

- Funkcja celu:

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

- Pochodne:

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

- Drugie pochodne:

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

- Hesjan:

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

Interpretacja kroku Newtona

Interpretacja kroku Newtona

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

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

Interpretacja kroku Newtona

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

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

■ Gradient:

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

Interpretacja kroku Newtona

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

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

■ Gradient:

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

■ Hesjan:

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

Interpretacja kroku Newtona

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

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

■ Gradient:

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

■ Hesjan:

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

■ Krok Newtona:

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

- Regresja liniowa:

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

- Regresja logistyczna – krok Newtona.

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

- Regresja liniowa:

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

- Regresja logistyczna – krok Newtona.

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

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

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

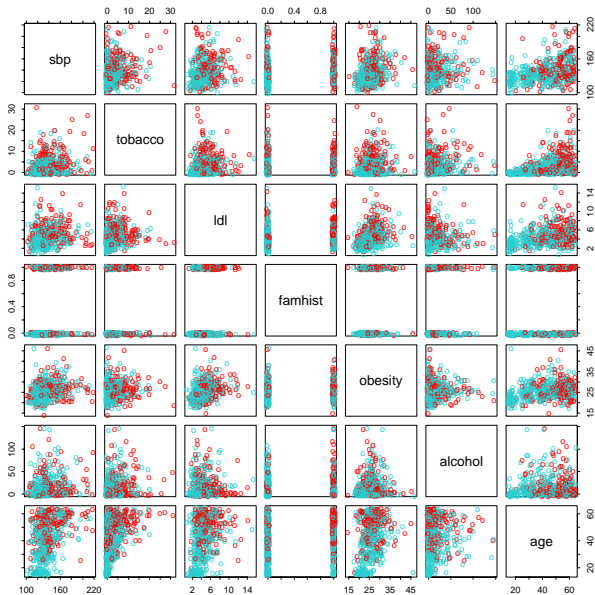
South African heart disease

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

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

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

South African heart disease

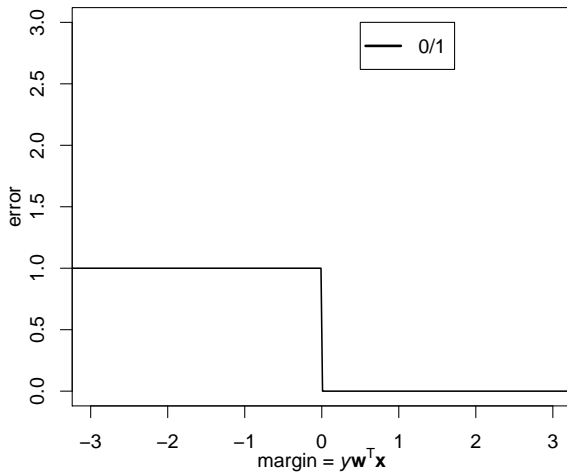


South African heart disease

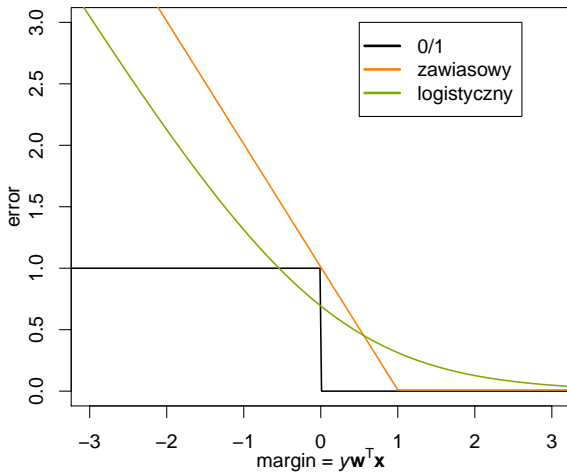
feature	coefficient
(intercept)	-4.130
SBP	0.006
TOBACCO	0.080
LDL	0.185
FAMHIST	0.939
OBESITY	-0.035
ALCOHOL	0.001
AGE	0.043

- 1 Problem klasyfikacji
- 2 Trenowanie klasyfikatora liniowego: błąd 0/1
- 3 Relaksacja błędu 0/1
- 4 Regresja logistyczna
- 5 Optymalizacja regresji logistycznej: metoda Cauchy'ego
- 6 Optymalizacja regresji logistycznej: metoda Newtona
- 7 Przykład
- 8 Zawiasowa funkcja błędu

Relaksacja błędu klasyfikacji



Relaksacja błędu klasyfikacji



Zawiasowa funkcja straty

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

gdzie:

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

Zawiasowa funkcja straty

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

gdzie:

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

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

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

gdzie:

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

- Wartość 1 jest tu wzięta arbitralnie – cokolwiek powyżej zera zadziała. Dlaczego?
- Ponieważ błąd w zerze musi być niezerowy, inaczej trywialne rozwiązanie $w = 0$ byłoby natychmiast optymalne.

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

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

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

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

Sprowadzamy do problemu programowania liniowego:

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

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

Sprowadzamy do problemu programowania liniowego:

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

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

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

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

Sprowadzamy do problemu programowania liniowego:

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

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

- Wystarczy zapewnić, że

$$1 - y_i \mathbf{w}_i^\top \mathbf{x}_i \leq \text{err}_i,$$

ponieważ będziemy minimalizowali błędy err_i .

Zawiasowa funkcja straty – optymalizacja

Minimalizacja sumy błędów zawiasowych na danych:

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

Sprowadzamy do problemu programowania liniowego:

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

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

- Wystarczy zapewnić, że

$$1 - y_i \mathbf{w}_i^\top \mathbf{x}_i \leq \text{err}_i,$$

ponieważ będziemy minimalizowali błędy err_i .

- Dowód:

- jeśli w rozwiązaniu optymalnym $1 - y_i \mathbf{w}_i^\top \mathbf{x}_i < \text{err}_i$, to oznacza, że $\text{err}_i = 0$.
- Inaczej, gdyby $\text{err}_i > 0$, możemy zmniejszyć err_i , co przeczyłoby optymalności.

Rozwiązujemy problem:

$$\min \quad L = \sum_{i=1}^n \text{err}_i$$

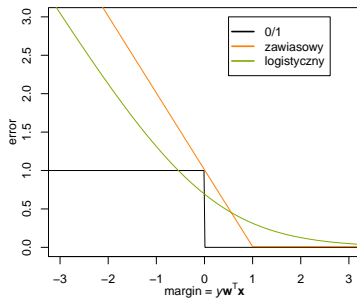
$$\text{p.o.} \quad 1 - y_i \mathbf{w}_i^\top \mathbf{x}_i \leq \text{err}_i$$

$$\text{err}_i \geq 0$$

$$i = 1, \dots, n$$

$$i = 1, \dots, n.$$

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 :)