

Metody Optymalizacji: Stochastyczny spadek wzdłuż gradientu (SGD) II

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/>

09.12.2013

- 1 Stochastyczny gradient (SGD) – postać ogólna
- 2 Adaptacja stochastycznego gradientu
- 3 Przykład

1 Stochastyczny gradient (SGD) – postać ogólna

2 Adaptacja stochastycznego gradientu

3 Przykład

Ogólna postać błędu dla regresji i klasyfikacji liniowej

W większości problemów uczenia maszynowego funkcja celu ma następującą postać:

$$L(\mathbf{w}) = \sum_{i=1}^n \ell_i(\mathbf{w})$$

Ogólna postać błędu dla regresji i klasyfikacji liniowej

W większości problemów uczenia maszynowego funkcja celu ma następującą postać:

$$L(\mathbf{w}) = \sum_{i=1}^n \ell_i(\mathbf{w})$$

- L jest sumarycznym błędem na zbiorze uczącym.

Ogólna postać błędu dla regresji i klasyfikacji liniowej

W większości problemów uczenia maszynowego funkcja celu ma następującą postać:

$$L(\mathbf{w}) = \sum_{i=1}^n \ell_i(\mathbf{w})$$

- L jest sumarycznym błędem na zbiorze uczącym.
- ℓ_i to błędy na poszczególnych obserwacjach.

Ogólna postać błędu dla regresji i klasyfikacji liniowej

W większości problemów uczenia maszynowego funkcja celu ma następującą postać:

$$L(\mathbf{w}) = \sum_{i=1}^n \ell_i(\mathbf{w})$$

- L jest sumarycznym błędem na zbiorze uczącym.
- ℓ_i to błędy na poszczególnych obserwacjach.
- W metodach liniowych, ℓ_i zależy od \mathbf{w} poprzez $\mathbf{w}^\top \mathbf{x}_i$:

$$\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)$$

Specyficzna postać funkcji celu – przykłady

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = (y_i - v)^2$$

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = (y_i - v)^2$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{|y_i - \mathbf{w}^\top \mathbf{x}_i|}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = |y_i - v|$$

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = (y_i - v)^2$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{|y_i - \mathbf{w}^\top \mathbf{x}_i|}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = |y_i - v|$$

- Klasyfikacja liniowa – regresja logistyczna:

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{\log \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) \right)}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = \log(1 + e^{-y_i v})$$

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = (y_i - v)^2$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{|y_i - \mathbf{w}^\top \mathbf{x}_i|}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = |y_i - v|$$

- Klasyfikacja liniowa – regresja logistyczna:

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{\log \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) \right)}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = \log(1 + e^{-y_i v})$$

- Klasyfikacja liniowa – funkcja zawiasowa:

$$L(\mathbf{w}) = \sum_{i=1}^n \underbrace{\left(1 - y_i \mathbf{w}^\top \mathbf{x}_i \right)_+}_{\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)} \quad \ell_i(v) = (1 - y_i v)_+$$

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

Minimalizacja funkcji $L(\boldsymbol{w})$:

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Wyznaczamy gradient funkcji ℓ_i w punkcie \mathbf{w}_{k-1} , $\nabla_{\ell_i}(\mathbf{w}_{k-1})$.

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Wyznaczamy gradient funkcji ℓ_i w punkcie \mathbf{w}_{k-1} , $\nabla_{\ell_i}(\mathbf{w}_{k-1})$.
 - Robimy krok wzdłuż negatywnego gradientu:

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

gdzie α_k jest długością kroku.

Obliczenie gradientu

Obliczenie gradientu

- W metodach liniowych, ℓ_i zależy od \mathbf{w} poprzez $\mathbf{w}^\top \mathbf{x}_i$:
$$\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)$$

Obliczenie gradientu

- W metodach liniowych, ℓ_i zależy od \mathbf{w} poprzez $\mathbf{w}^\top \mathbf{x}_i$:

$$\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)$$

- Wyznaczenie gradientu $\nabla_{\ell_i}(\mathbf{w})$ poprzez pochodną wewnętrzną:

$$\frac{\partial \ell_i(\mathbf{w})}{\partial w_k} = \frac{\partial \ell_i(\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k}$$

Obliczenie gradientu

- W metodach liniowych, ℓ_i zależy od \mathbf{w} poprzez $\mathbf{w}^\top \mathbf{x}_i$:

$$\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)$$

- Wyznaczenie gradientu $\nabla_{\ell_i}(\mathbf{w})$ poprzez pochodną wewnętrzną:

$$\begin{aligned}\frac{\partial \ell_i(\mathbf{w})}{\partial w_k} &= \frac{\partial \ell_i(\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v} \Big|_{v=\mathbf{w}^\top \mathbf{x}_i} \frac{\partial (\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k}\end{aligned}$$

Obliczenie gradientu

- W metodach liniowych, ℓ_i zależy od \mathbf{w} poprzez $\mathbf{w}^\top \mathbf{x}_i$:

$$\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)$$

- Wyznaczenie gradientu $\nabla_{\ell_i}(\mathbf{w})$ poprzez pochodną wewnętrzną:

$$\begin{aligned}\frac{\partial \ell_i(\mathbf{w})}{\partial w_k} &= \frac{\partial \ell_i(\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v} \Big|_{v=\mathbf{w}^\top \mathbf{x}_i} \frac{\partial (\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v} \Big|_{v=\mathbf{w}^\top \mathbf{x}_i} x_{ik}.\end{aligned}$$

Obliczenie gradientu

- W metodach liniowych, ℓ_i zależy od \mathbf{w} poprzez $\mathbf{w}^\top \mathbf{x}_i$:

$$\ell_i(\mathbf{w}) = \ell_i(\mathbf{w}^\top \mathbf{x}_i)$$

- Wyznaczenie gradientu $\nabla_{\ell_i}(\mathbf{w})$ poprzez pochodną wewnętrzną:

$$\begin{aligned}\frac{\partial \ell_i(\mathbf{w})}{\partial w_k} &= \frac{\partial \ell_i(\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v} \Big|_{v=\mathbf{w}^\top \mathbf{x}_i} \frac{\partial (\mathbf{w}^\top \mathbf{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v} \Big|_{v=\mathbf{w}^\top \mathbf{x}_i} x_{ik}.\end{aligned}$$

- Jeśli oznaczymy $\ell'_i(\mathbf{w}^\top \mathbf{x}_i) := \frac{\partial \ell_i(v)}{\partial v} \Big|_{v=\mathbf{w}^\top \mathbf{x}_i}$, to gradient możemy zapisać jako:

$$\nabla_{\ell_i}(\mathbf{w}) = \ell'_i(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i$$

Stochastyczny spadek wzdłuż gradientu

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Wyznaczamy $\ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i)$, pochodną funkcji $\ell_i(v)$ w punkcie $v = \mathbf{w}_{k-1}^\top \mathbf{x}_i$.
 - Robimy krok wzdłuż negatywnego gradientu:

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha_k \ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

gdzie α_k jest długością kroku.

Stochastyczny gradient – długość kroku

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha_k \ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha_k \ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

- Regresja liniowa – metoda najmniejszych kwadratów

$$\ell_i(v) = (y_i - v)^2 \quad \ell'_i(v) = -2(y_i - v)$$

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha_k \ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

- Regresja liniowa – metoda najmniejszych kwadratów

$$\ell_i(v) = (y_i - v)^2 \quad \ell'_i(v) = -2(y_i - v)$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$\ell_i(v) = |y_i - v| \quad \ell'_i(v) = -\text{sgn}(y_i - v)$$

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha_k \ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

- Regresja liniowa – metoda najmniejszych kwadratów

$$\ell_i(v) = (y_i - v)^2 \quad \ell'_i(v) = -2(y_i - v)$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$\ell_i(v) = |y_i - v| \quad \ell'_i(v) = -\text{sgn}(y_i - v)$$

- Klasyfikacja liniowa – regresja logistyczna:

$$\ell_i(v) = \log(1 + e^{-y_i v}) \quad \ell'_i(v) = -\frac{y_i}{1 + e^{y_i v}}$$

Stochastyczny gradient – długość kroku

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha_k \ell'_i(\mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

- Regresja liniowa – metoda najmniejszych kwadratów

$$\ell_i(v) = (y_i - v)^2 \quad \ell'_i(v) = -2(y_i - v)$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$\ell_i(v) = |y_i - v| \quad \ell'_i(v) = -\text{sgn}(y_i - v)$$

- Klasyfikacja liniowa – regresja logistyczna:

$$\ell_i(v) = \log(1 + e^{-y_i v}) \quad \ell'_i(v) = -\frac{y_i}{1 + e^{y_i v}}$$

- Klasyfikacja liniowa – funkcja zawiasowa:

$$\ell_i(v) = (1 - y_i v)_+ \quad \ell'_i(v) = \begin{cases} 0 & \text{jeśli } y_i v > 1 \\ -y_i & \text{jeśli } y_i v \leq 1 \end{cases}$$

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Modyfikuj wagi:

$$\mathbf{w}_k = \mathbf{w}_{k-1} + 2\alpha_k(y_i - \mathbf{w}_{k-1}^\top \mathbf{x}_i)\mathbf{x}_i,$$

gdzie α_k jest długością kroku.

SGD: Błąd absolutny (wartość bezwzględna)

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Modyfikuj wagi:

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \alpha_k \text{sgn}(y_i - \mathbf{w}_{k-1}^\top \mathbf{x}_i) \mathbf{x}_i,$$

gdzie α_k jest długością kroku.

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np. $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Modyfikuj wagi:

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \alpha_k \frac{y_i}{1 + e^{y_i \mathbf{w}_{k-1}^\top \mathbf{x}_i}} \mathbf{x}_i,$$

gdzie α_k jest długością kroku.

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np.
 $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Jeśli $y_i \mathbf{w}_{k-1}^\top \mathbf{x}_i > 1$, nie modyfikujemy wag. W przeciwnym przypadku:

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \alpha_k y_i \mathbf{x}_i,$$

gdzie α_k jest długością kroku.

Minimalizacja funkcji $L(\mathbf{w})$:

- 1 Zaczynamy od wybranego rozwiązania startowego, np.
 $\mathbf{w}_0 = \mathbf{0}$.
- 2 Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$.
 - Jeśli $y_i \mathbf{w}_{k-1}^\top \mathbf{x}_i > 1$, nie modyfikujemy wag. W przeciwnym przypadku:

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \alpha_k y_i \mathbf{x}_i,$$

gdzie α_k jest długością kroku.

⇒ Perceptron!

- 1 Stochastyczny gradient (SGD) – postać ogólna
- 2 Adaptacja stochastycznego gradientu
- 3 Przykład

- Teoria mówi, że długość kroku powinna wynosić:

$$\alpha_k = O\left(\frac{1}{\sqrt{k}}\right),$$

aby algorytm zbiegł do rozwiązania optymalnego.

- Teoria mówi, że długość kroku powinna wynosić:

$$\alpha_k = O\left(\frac{1}{\sqrt{k}}\right),$$

aby algorytm zbiegł do rozwiązania optymalnego.

- Stała przy $O(\cdot)$ może zostać wyznaczona teoretycznie z analizy **najgorszego przypadku**: jeśli optymalne rozwiązanie \mathbf{w}^* będzie ograniczone w sensie normy przez stałą W , tj. $\|\mathbf{w}^*\| \leq W$, oraz mamy ograniczenie górne na normę wszystkich gradientów, $\|\nabla_{\ell_i}(\mathbf{w})\| \leq G$, to

$$\alpha_k = \frac{W}{G} \frac{1}{\sqrt{k}}.$$

- Teoria mówi, że długość kroku powinna wynosić:

$$\alpha_k = O\left(\frac{1}{\sqrt{k}}\right),$$

aby algorytm zbiegł do rozwiązania optymalnego.

- Stała przy $O(\cdot)$ może zostać wyznaczona teoretycznie z analizy **najgorszego przypadku**: jeśli optymalne rozwiązanie \mathbf{w}^* będzie ograniczone w sensie normy przez stałą W , tj. $\|\mathbf{w}^*\| \leq W$, oraz mamy ograniczenie górne na normę wszystkich gradientów, $\|\nabla_{\ell_i}(\mathbf{w})\| \leq G$, to

$$\alpha_k = \frac{W}{G} \frac{1}{\sqrt{k}}.$$

- W praktyce dobiera się stałą sprawdzając jak algorytm zbiega na posiadanych danych.

- Teoria mówi, że długość kroku powinna wynosić:

$$\alpha_k = O\left(\frac{1}{\sqrt{k}}\right),$$

aby algorytm zbiegł do rozwiązania optymalnego.

- Stała przy $O(\cdot)$ może zostać wyznaczona teoretycznie z analizy **najgorszego przypadku**: jeśli optymalne rozwiązanie \mathbf{w}^* będzie ograniczone w sensie normy przez stałą W , tj. $\|\mathbf{w}^*\| \leq W$, oraz mamy ograniczenie górne na normę wszystkich gradientów, $\|\nabla_{\ell_i}(\mathbf{w})\| \leq G$, to

$$\alpha_k = \frac{W}{G} \frac{1}{\sqrt{k}}.$$

- W praktyce dobiera się stałą sprawdzając jak algorytm zbiega na posiadanych danych.
- W praktyce mała i stała szybkość uczenia, $\alpha_k = \text{const}$, działa również bardzo dobrze.

Zmiana skali zmiennych wejściowych

Przykład: samochody

- Dwie zmienne wejściowe: X_1 – cena (w PLN), X_2 – spalanie (w l/100km)

Przykład: samochody

- Dwie zmienne wejściowe: X_1 – cena (w PLN), X_2 – spalanie (w $l/100km$)
- X_1 ma skalę rzędu $10^4 - 10^5$, X_2 rzędu $10^0 - 10^1$.

Przykład: samochody

- Dwie zmienne wejściowe: X_1 – cena (w PLN), X_2 – spalanie (w l/100km)
- X_1 ma skalę rzędu $10^4 - 10^5$, X_2 rzędu $10^0 - 10^1$.
- Krok wzdłuż gradientu $\mathbf{w}_1 = \mathbf{w}_0 - \alpha_1 \ell'_i(\mathbf{w}_0^\top \mathbf{x}_i) \mathbf{x}_i$ wyznaczy każdą z wag \mathbf{w}_{1j} proporcjonalnie do x_{ij} .

Przykład: samochody

- Dwie zmienne wejściowe: X_1 – cena (w PLN), X_2 – spalanie (w l/100km)
- X_1 ma skalę rzędu $10^4 - 10^5$, X_2 rzędu $10^0 - 10^1$.
- Krok wzdłuż gradientu $\mathbf{w}_1 = \mathbf{w}_0 - \alpha_1 \ell'_i(\mathbf{w}_0^\top \mathbf{x}_i) \mathbf{x}_i$ wyznaczy każdą z wag \mathbf{w}_{1j} proporcjonalnie do x_{ij} .
- Czyli waga dla ceny będzie rzędu $10^4 - 10^5$, a waga dla spalania – rzędu $10^0 - 10^1$.

Zmiana skali zmiennych wejściowych

Przykład: samochody

- Dwie zmienne wejściowe: X_1 – cena (w PLN), X_2 – spalanie (w l/100km)
- X_1 ma skalę rzędu $10^4 - 10^5$, X_2 rzędu $10^0 - 10^1$.
- Krok wzdłuż gradientu $\mathbf{w}_1 = \mathbf{w}_0 - \alpha_1 \ell'_i(\mathbf{w}_0^\top \mathbf{x}_i) \mathbf{x}_i$ wyznaczy każdą z wag \mathbf{w}_{1j} proporcjonalnie do x_{ij} .
- Czyli waga dla ceny będzie rzędu $10^4 - 10^5$, a waga dla spalania – rzędu $10^0 - 10^1$.
- Przy kolejnej obserwacji, przemnożenie $\mathbf{w}_1^\top \mathbf{x}_i$ da sumę 2 składników o wielkości $10^8 - 10^{10}$ i $10^0 - 10^2$.

Zmiana skali zmiennych wejściowych

Przykład: samochody

- Dwie zmienne wejściowe: X_1 – cena (w PLN), X_2 – spalanie (w l/100km)
- X_1 ma skalę rzędu $10^4 - 10^5$, X_2 rzędu $10^0 - 10^1$.
- Krok wzdłuż gradientu $\mathbf{w}_1 = \mathbf{w}_0 - \alpha_1 \ell'_i(\mathbf{w}_0^\top \mathbf{x}_i) \mathbf{x}_i$ wyznaczy każdą z wag \mathbf{w}_{1j} proporcjonalnie do x_{ij} .
- Czyli waga dla ceny będzie rzędu $10^4 - 10^5$, a waga dla spalania – rzędu $10^0 - 10^1$.
- Przy kolejnej obserwacji, przemnożenie $\mathbf{w}_1^\top \mathbf{x}_i$ da sumę 2 składników o wielkości $10^8 - 10^{10}$ i $10^0 - 10^2$.
- **Wnioski:**
 - Zmienna o większej skali zdominuje zmienną o mniejszej skali,
 - Algorytm może się rozbiec.

Zmiana skali zmiennych wejściowych

W praktyce można znacznie polepszyć zbieżność algorytmu, sprowadzając zmienne wejściowe **do tej samej skali**.

Zmiana skali zmiennych wejściowych

W praktyce można znacznie polepszyć zbieżność algorytmu, sprowadzając zmienne wejściowe **do tej samej skali**.

Dwie metody sprowadzania do tej samej skali:

- **Normalizacja** zmiennych.
- **Standaryzacja** zmiennych.

Pomysł

Dla każdej zmiennej wejściowej X_j , najmniejszą wartość $x_{j,\min}$ zamień na 0, największą $x_{j,\max}$ na 1, a pozostałe wyznacz proporcjonalnie:

$$x_{ij} \mapsto \frac{x_{ij} - x_{j,\min}}{x_{j,\max} - x_{j,\min}}$$

Normalizacja zmiennych

Pomysł

Dla każdej zmiennej wejściowej X_j , najmniejszą wartość $x_{j,\min}$ zamień na 0, największą $x_{j,\max}$ na 1, a pozostałe wyznacz proporcjonalnie:

$$x_{ij} \mapsto \frac{x_{ij} - x_{j,\min}}{x_{j,\max} - x_{j,\min}}$$

Przykład

Wartości zmiennej: (1, 2, 3, 4, 8, 9, 11).

Pomysł

Dla każdej zmiennej wejściowej X_j , najmniejszą wartość $x_{j,\min}$ zamień na 0, największą $x_{j,\max}$ na 1, a pozostałe wyznacz proporcjonalnie:

$$x_{ij} \mapsto \frac{x_{ij} - x_{j,\min}}{x_{j,\max} - x_{j,\min}}$$

Przykład

Wartości zmiennej: (1, 2, 3, 4, 8, 9, 11).

Nowe wartości: (0, 0.1, 0.2, 0.3, 0.7, 0.8, 1).

Zalety

- Wszystkie zmienne sprowadzone do wartości w skali od 0 do 1

Zalety

- Wszystkie zmienne sprowadzone do wartości w skali od 0 do 1

Wady

- Metoda nie odporna na wartości odstające.

Normalizacja zmiennych

Zalety

- Wszystkie zmienne sprowadzone do wartości w skali od 0 do 1

Wady

- Metoda nie odporna na wartości odstające.

Przykład

Wartości zmiennej: (0, 1, 2, 3, 100000).

Nowe wartości: (0, 0.00001, 0.00002, 0.00003, 1).

Pomysł

Dla każdej zmiennej wejściowej X_j , wyznacz wartość średnią i odchylenie standardowe

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2},$$

a następnie zamień:

$$x_{ij} \mapsto \frac{x_{ij} - \bar{x}_j}{s_j}$$

Standaryzacja zmiennych

Pomysł

Dla każdej zmiennej wejściowej X_j , wyznacz wartość średnią i odchylenie standardowe

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2},$$

a następnie zamień:

$$x_{ij} \mapsto \frac{x_{ij} - \bar{x}_j}{s_j}$$

Przykład

Wartości zmiennej: (1, 2, 3, 4, 5).

Standaryzacja zmiennych

Pomysł

Dla każdej zmiennej wejściowej X_j , wyznacz wartość średnią i odchylenie standardowe

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2},$$

a następnie zamień:

$$x_{ij} \mapsto \frac{x_{ij} - \bar{x}_j}{s_j}$$

Przykład

Wartości zmiennej: (1, 2, 3, 4, 5).

Średnia: $\bar{x} = 3$

Odchylenie: $s = 1.581$

Nowe wartości: (-1.265, -0.632, 0, 0.632, 1.265).

Zalety

- Wszystkie zmienne mają ten sam rozrzut (1) i średnią (0)
- Metoda znacznie odporniejsza na wartości odstające

Zalety

- Wszystkie zmienne mają ten sam rozrzut (1) i średnią (0)
- Metoda znacznie odporniejsza na wartości odstające

Wady

- Problemy mogą się pojawić, gdy wszystkie wartości zmiennej są takie same (zerowe odchylenie standardowe)

Zalety

- Wszystkie zmienne mają ten sam rozrzut (1) i średnią (0)
- Metoda znacznie odporniejsza na wartości odstające

Wady

- Problemy mogą się pojawić, gdy wszystkie wartości zmiennej są takie same (zerowe odchylenie standardowe)

W ogólności standaryzacja preferowana nad normalizację.

- Algorytm stochastycznego spadku wzdłuż gradientu może się w pewnych specyficznych przypadkach rozbiegać (wektor wag zaczyna niekontrolowanie wzrastać).
- Potrzebna jest wtedy tzw. *renormalizacja wag*.

- Algorytm stochastycznego spadku wzdłuż gradientu może się w pewnych specyficznych przypadkach rozbiegać (wektor wag zaczyna niekontrolowanie wzrastać).
- Potrzebna jest wtedy tzw. *renormalizacja wag*.
- Przyjmujemy pewną maksymalną normę wag W (np. $W = 10$ powinno wystarczyć dla znormalizowanych danych).

- Algorytm stochastycznego spadku wzdłuż gradientu może się w pewnych specyficznych przypadkach rozbiegać (wektor wag zaczyna niekontrolowanie wzrastać).
- Potrzebna jest wtedy tzw. *renormalizacja wag*.
- Przyjmujemy pewną maksymalną normę wag W (np. $W = 10$ powinno wystarczyć dla znormalizowanych danych).
- Jeśli kiedykolwiek $\|\mathbf{w}_i\| > W$, to *renormalizujemy wagi* do wartości W :

$$\mathbf{w}_i := \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} W.$$

- Algorytm stochastycznego spadku wzdłuż gradientu może się w pewnych specyficznych przypadkach rozbiegać (wektor wag zaczyna niekontrolowanie wzrastać).
- Potrzebna jest wtedy tzw. *renormalizacja wag*.
- Przyjmujemy pewną maksymalną normę wag W (np. $W = 10$ powinno wystarczyć dla znormalizowanych danych).
- Jeśli kiedykolwiek $\|w_i\| > W$, to *renormalizujemy wagi* do wartości W :

$$w_i := \frac{w_i}{\|w_i\|} W.$$

- Jeśli dobrze dobierzemy W , ta procedura gwarantuje szybszą zbieżność, ma też charakter regularyzacji.

- 1 Stochastyczny gradient (SGD) – postać ogólna
- 2 Adaptacja stochastycznego gradientu
- 3 Przykład

Zbiór danych: Reuters RCV1:

- Zbiór 810 000 dokumentów z Reuters News opublikowanych w latach 1996-1997
- Dla każdego dokumentu, przypisane kategorie tematyczne.

Tworzymy prosty problem binarnej klasyfikacji próbując przewidzieć czy dokument należy do kategorii *CCAT (Corporate/Industrial)*.

Przykład dokumentu

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="2330" id="root" date="1996-08-20" xml:lang="en">
<title>USA: Tylan stock jumps; weighs sale of company.</title>
<headline>Tylan stock jumps; weighs sale of company.</headline>
<dateline>SAN DIEGO</dateline>
<text>
<p>The stock of Tylan General Inc. jumped Tuesday after the maker of
process-management equipment said it is exploring the sale of the
company and added that it has already received some inquiries from
potential buyers.</p>
<p>Tylan was up $2.50 to $12.75 in early trading on the Nasdaq market.</p>
<p>The company said it has set up a committee of directors to oversee
the sale and that Goldman, Sachs & Co. has been retained as its
financial adviser.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
<code code="USA"> </code>
</codes>
<codes class="bip:industries:1.0">
<code code="I34420"> </code>
</codes>
<codes class="bip:topics:1.0">
<code code="C15"> </code>
<code code="C152"> </code>
<code code="C18"> </code>
<code code="C181"> </code>
<code code="CCAT"> </code>
</codes>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
<dc element="dc.date.published" value="1996-08-20"/>
<dc element="dc.source" value="Reuters"/>
<dc element="dc.creator.location" value="SAN DIEGO"/>
```

Zamiana dokumentów tekstowych na wektory zmiennych wejściowych:

- Każde słowo daje nową zmienną wejściową.
- Wartość zmiennej wejściowej j w dokumencie i jest niezerowa, jeśli słowo j pojawiło się w dokumencie i .
- Dodatkowe ważenie zmiennych (TF-IDF), odrzucanie tzw. *stopwords*, użycie słownika wyrazów bliskoznacznych, stemmery, lematyzatora, itp. (patrz wykład dr. Dembczyńskiego).

Zamiana dokumentów tekstowych na wektory zmiennych wejściowych:

- Każde słowo daje nową zmienną wejściową.
- Wartość zmiennej wejściowej j w dokumencie i jest niezerowa, jeśli słowo j pojawiło się w dokumencie i .
- Dodatkowe ważenie zmiennych (TF-IDF), odrzucanie tzw. *stopwords*, użycie słownika wyrazów bliskoznacznych, stemmera, lematyzatora, itp. (patrz wykład dr. Dembczyńskiego).

W naszym przypadku otrzymujemy łącznie prawie 50 000 cech wejściowych.

Użycie dwóch rodzajów klasyfikatorów:

- Regresji logistycznej,
- Minimalizacja błędu 'zawiasowego' (nazywane też *linear SVM*).

Użycie dwóch rodzajów klasyfikatorów:

- Regresji logistycznej,
- Minimalizacja błędu 'zawiasowego' (nazywane też *linear SVM*).

Porównanie różnych metod optymalizacji:

- Standardowa – odmiana metody Newtona-Rapshona lub inna, działająca na całym zbiorze danych.
- Stochastyczny gradient.

Użycie dwóch rodzajów klasyfikatorów:

- Regresji logistycznej,
- Minimalizacja błędu 'zawiasowego' (nazywane też *linear SVM*).

Porównanie różnych metod optymalizacji:

- Standardowa – odmiana metody Newtona-Rapshona lub inna, działająca na całym zbiorze danych.
- Stochastyczny gradient.

Źródło: <http://leon.bottou.org/projects/sgd>

Błąd 'zawiasowy'

metoda	czas obliczeń	funkcja celu	błąd testowy
SVMLight (standard.)	23 642 sek	0.2275	6.02%
SVMPerf (standard.)	66 sek	0.2278	6.03%
SGD	1.4 sek	0.2275	6.02%

Błąd logistyczny

metoda	czas obliczeń	funkcja celu	błąd testowy
LibLinear (standard.)	30 sek	0.18907	5.68%
SGD	2.3 sek	0.18893	5.66%

Dalsze przykłady – błąd zawiasowy

Zbiór danych	#obserwacji	#cech	Czas LIBSVM	Czas SGD
Reuters	781 000	47 000	2.5 dnia	7sek
Translation	1 000 000	274 000	wiele dni	7sek
SuperTag	950 000	46 000	8h	1sek
Voicetone	579 000	88 000	10h	1sek

- Stochastyczny gradient jest jedną z najszybszych metod optymalizacji w analizie danych.
- Bardzo dobrze się skaluje, nie wymaga ładowania do pamięci całego zbioru danych.
- Wymaga ostrożności przy ustalaniu schematu zmian długości kroku α_k .

Koniec na dzisiaj :)