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: piątek 15:10-16:40
Slajdy dostępne pod adresem: `http://www.cs.put.poznan.pl/wkotlowski/to/`

11.12.2018

Spis treści

Spis treści

Ogólna postać błędu dla regresji i klasyfikacji liniowej

W większości problemów uczenia maszynowego funkcja celu ma następującą postać:

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

Ogólna postać błędu dla regresji i klasyfikacji liniowej

W większości problemów uczenia maszynowego funkcja celu ma następującą postać:

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

- 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:

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

- 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:

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

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

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

Specyficzna postać funkcji celu – przykłady

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

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

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

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

- Regresja liniowa – min. wartości bezwzględnych błędów

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

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

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

- Regresja liniowa – min. wartości bezwzględnych błędów

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

- Klasyfikacja liniowa – regresja logistyczna:

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

Specyficzna postać funkcji celu – przykłady

- Regresja liniowa – metoda najmniejszych kwadratów

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

- Regresja liniowa – min. wartości bezwzględnych błędów

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

- Klasyfikacja liniowa – regresja logistyczna:

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

- Klasyfikacja liniowa – funkcja zawiasowa:

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

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

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

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

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

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

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

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

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

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

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

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

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

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

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

Stochastyczny spadek wzdłuż gradientu (SGD) – przypomnienie

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

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

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

gdzie α_k jest długością kroku.

Obliczenie gradientu

Obliczenie gradientu

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

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

Obliczenie gradientu

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

- Wyznaczenie gradientu $\nabla_{\ell_i}(\boldsymbol{w})$ poprzez pochodną wewnętrzną:
$$\frac{\partial \ell_i(\boldsymbol{w})}{\partial w_k} = \frac{\partial \ell_i(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k}$$

Obliczenie gradientu

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

- Wyznaczenie gradientu $\nabla \ell_i(\boldsymbol{w})$ poprzez pochodną wewnętrzną:
$$\begin{aligned}\frac{\partial \ell_i(\boldsymbol{w})}{\partial w_k} &= \frac{\partial \ell_i(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v}\Big|_{v=\boldsymbol{w}^\top \boldsymbol{x}_i} \frac{\partial(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k}\end{aligned}$$

Obliczenie gradientu

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

- Wyznaczenie gradientu $\nabla_{\ell_i}(\boldsymbol{w})$ poprzez pochodną wewnętrzną:
$$\begin{aligned}\frac{\partial \ell_i(\boldsymbol{w})}{\partial w_k} &= \frac{\partial \ell_i(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v}\Big|_{v=\boldsymbol{w}^\top \boldsymbol{x}_i} \frac{\partial(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v}\Big|_{v=\boldsymbol{w}^\top \boldsymbol{x}_i} x_{ik}.\end{aligned}$$

Obliczenie gradientu

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

- Wyznaczenie gradientu $\nabla \ell_i(\boldsymbol{w})$ poprzez pochodną wewnętrzną:
$$\begin{aligned}\frac{\partial \ell_i(\boldsymbol{w})}{\partial w_k} &= \frac{\partial \ell_i(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v}\Big|_{v=\boldsymbol{w}^\top \boldsymbol{x}_i} \frac{\partial(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial w_k} \\ &= \frac{\partial \ell_i(v)}{\partial v}\Big|_{v=\boldsymbol{w}^\top \boldsymbol{x}_i} x_{ik}.\end{aligned}$$

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

Stochastyczny spadek wzdłuż gradientu

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

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

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

   gdzie α_k jest długością kroku.

Stochastyczny gradient – długość kroku

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

Stochastyczny gradient – długość kroku

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

- Regresja liniowa – metoda najmniejszych kwadratów

$$\ell_i(v) = (y_i - v)^2 \qquad \ell_i'(v) = -2(y_i - v)$$

Stochastyczny gradient – długość kroku

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

- Regresja liniowa – metoda najmniejszych kwadratów

$$\ell_i(v) = (y_i - v)^2 \qquad \ell_i'(v) = -2(y_i - v)$$

- Regresja liniowa – min. wartości bezwzględnych błędów

$$\ell_i(v) = |y_i - v| \qquad \ell_i'(v) = -\operatorname{sgn}(y_i - v)$$

Stochastyczny gradient – długość kroku

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

- Regresja liniowa – metoda najmniejszych kwadratów
$$\ell_i(v) = (y_i - v)^2 \qquad \ell_i'(v) = -2(y_i - v)$$
- Regresja liniowa – min. wartości bezwzględnych błędów
$$\ell_i(v) = |y_i - v| \qquad \ell_i'(v) = -\operatorname{sgn}(y_i - v)$$
- Klasyfikacja liniowa – regresja logistyczna:
$$\ell_i(v) = \log\left(1 + e^{-y_i v}\right) \qquad \ell_i'(v) = -\frac{y_i}{1 + e^{y_i v}}$$

Stochastyczny gradient – długość kroku

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

- Regresja liniowa – metoda najmniejszych kwadratów
$$\ell_i(v) = (y_i - v)^2 \qquad \ell_i'(v) = -2(y_i - v)$$

- Regresja liniowa – min. wartości bezwzględnych błędów
$$\ell_i(v) = |y_i - v| \qquad \ell_i'(v) = -\operatorname{sgn}(y_i - v)$$

- Klasyfikacja liniowa – regresja logistyczna:
$$\ell_i(v) = \log\left(1 + e^{-y_i v}\right) \qquad \ell_i'(v) = -\frac{y_i}{1 + e^{y_i v}}$$

- Klasyfikacja liniowa – funkcja zawiasowa:
$$\ell_i(v) = (1 - y_i v)_+ \qquad \ell_i'(v) = \begin{cases} 0 & \text{jeśli} \quad y_i v > 1 \\ -y_i & \text{jeśli} \quad y_i v \leq 1 \end{cases}$$

SGD: Błąd kwadratowy

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

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

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

gdzie α_k jest długością kroku.

SGD: Błąd absolutny (wartość bezwzględna)

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

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

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

gdzie α_k jest długością kroku.

SGD: Błąd logistyczny

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

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

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

gdzie α_k jest długością kroku.

SGD: Błąd zawiasowy

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

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

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

     gdzie α_k jest długością kroku.

SGD: Błąd zawiasowy

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

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

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

     gdzie α_k jest długością kroku.

$\Longrightarrow$ Perceptron!

Spis treści

Długość kroku

Długość kroku

- Teoria mówi, że długość kroku powinna wynosić:

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

aby algorytm zbiegł do rozwiązania optymalnego.

Długość kroku

- 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 $\boldsymbol{w}^*$ będzie ograniczone w sensie normy przez stałą W, tj. $\|\boldsymbol{w}^*\| \leq W$, oraz mamy ograniczenie górne na normę wszystkich gradientów, $\|\nabla_{\ell_i}(\boldsymbol{w})\| \leq G$, to

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

Długość kroku

- 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 $\boldsymbol{w}^*$ będzie ograniczone w sensie normy przez stałą W, tj. $\|\boldsymbol{w}^*\| \leq W$, oraz mamy ograniczenie górne na normę wszystkich gradientów, $\|\nabla_{\ell_i}(\boldsymbol{w})\| \leq G$, to

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

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

Długość kroku

- 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 $\boldsymbol{w}^*$ będzie ograniczone w sensie normy przez stałą W, tj. $\|\boldsymbol{w}^*\| \leq W$, oraz mamy ograniczenie górne na normę wszystkich gradientów, $\|\nabla_{\ell_i}(\boldsymbol{w})\| \leq G$, 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, $\alpha_k = \mathrm{const}$, działa również bardzo dobrze.

Zmiana skali zmiennych wejściowych

Przykład: samochody

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

Zmiana skali zmiennych wejściowych

Przykład: samochody

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

Zmiana skali zmiennych wejściowych

Przykład: samochody

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

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 $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \alpha_1 \ell_i'(\boldsymbol{w}_0^\top \boldsymbol{x}_i)\boldsymbol{x}_i$ wyznaczy każdą z wag $\boldsymbol{w}_{1j}$ 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 $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \alpha_1 \ell_i'(\boldsymbol{w}_0^\top \boldsymbol{x}_i)\boldsymbol{x}_i$ wyznaczy każdą z wag $\boldsymbol{w}_{1j}$ 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 $\boldsymbol{w}_1^\top \boldsymbol{x}_i$ 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 $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \alpha_1 \ell_i'(\boldsymbol{w}_0^\top \boldsymbol{x}_i)\boldsymbol{x}_i$ wyznaczy każdą z wag $\boldsymbol{w}_{1j}$ 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 $\boldsymbol{w}_1^\top \boldsymbol{x}_i$ 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.

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:

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

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

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: $(1, 2, 3, 4, 8, 9, 11)$.
Nowe wartości: $(0, 0.1, 0.2, 0.3, 0.7, 0.8, 1)$.

Normalizacja zmiennych

Zalety

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

Normalizacja zmiennych

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: $(0, 0.00001, 0.00002, 0.00003, 1)$.

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:

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

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: $(-1.265, -0.632, 0, 0.632, 1.265)$.

Standaryzacja zmiennych

Zalety

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

Standaryzacja zmiennych

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)

Standaryzacja zmiennych

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.

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*.

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).

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).
- Jeśli kiedykolwiek $\|\boldsymbol{w}_i\| > W$, to *renormalizujemy wagi* do wartości W:

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

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).
- Jeśli kiedykolwiek $\|\boldsymbol{w}_i\| > W$, to *renormalizujemy wagi* do wartości W:

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

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

Spis treści

Klasyfikacja dokumentów tekstowych

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 &amp;amp; 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"/>
```

Reprezentacja *bag-of-words*

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).

Reprezentacja *bag-of-words*

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.

Eksperyment

Użycie dwóch rodzajów klasyfikatorów:

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

Eksperyment

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.

Eksperyment

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`

Wyniki

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

Wnioski

- 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 :)