Klasyfikacja za pomocą regresji logistycznej

Techniki Optymalizacji

02.12.2013

1 Problem klasyfikacji

Przewidywania/wyjaśnienie zmian jednej zmiennej *dyskretnej* (Y) pod wpływem zmian innych zmiennych ($\boldsymbol{X} = (X_1, \ldots, X_m)$). Zwykle Y przyjmuje tylko kilka możliwych wartości, które nazywa się *klasami, etykietami* lub *kategoriami.* Często $Y \in \{-1, +1\}$, mamy wtedy do czynienia z *klasyfikacją binarną.* Zajmiemy się tylko tą ostatnią, ponieważ każdy problem z więcej niż dwoma klasami da się sprowadzić do klasyfikacji binarnej (np. metodą „jeden przeciwko wszystkim"). Przykłady:

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

Dzisiaj zajmiemy się *klasyfikacją liniową.* Kodując klasy jako $\{-1, +1\}$, klasyfikator liniowy składa się z dwóch części:

- Funkcji liniowej od $\boldsymbol{X}$:

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

- Klasyfikacji poprzez progowanie w zerze:

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

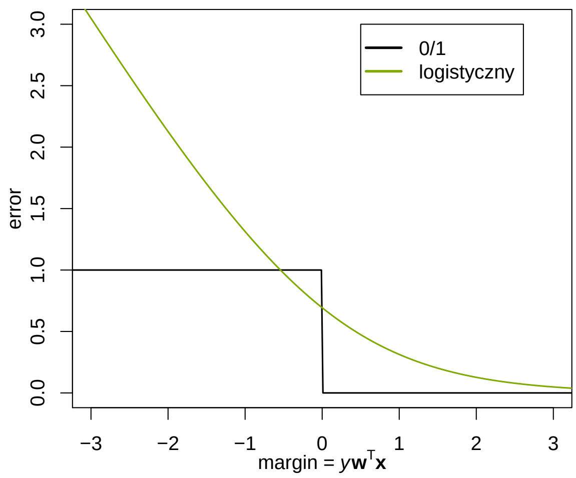
Naturalnym błędem klasyfikacji jest błąd 0/1 („zero-jedynkowy"), zapisany jako:

$$\ell(Y, \hat{Y}) = \mathbb{1}[Y \neq \hat{Y}],$$

gdzie $\mathbb{1}[C]$ to *funkcja indykatorowa*: `C ? 1 : 0`. Chcielibyśmy więc wytrenować klasyfikator minimalizując jego błąd na zbiorze uczącym:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \mathbb{1}[y_i \neq \hat{y}_i] = \sum_{i=1}^{n} \mathbb{1}[y_i \boldsymbol{w}_i^\top \boldsymbol{x}_i < 0].$$

Niestety, jest to problem NP-trudny, stąd przybliżamy błąd 0/1 innym typem błędu. Dzisiaj zajmiemy się błędem logistycznym.



Rysunek 1: Błąd logistyczny w porównaniu do błędu 0/1 jako funkcja *marginesu* $Y\boldsymbol{w}^\top\boldsymbol{X}$

2 Regresja logistyczna

Uwaga: regresja logistyczna jest procedurą klasyfikacji! Nazwa „regresja" ma tu tylko znaczenie historyczne. Zastępujemy błąd 0/1 błędem logistycznym:

$$\ell(Y,\boldsymbol{w}) = \log\left(1 + e^{-Y\boldsymbol{w}^\top\boldsymbol{X}}\right).$$

Wyrażenie $Y\boldsymbol{w}^\top\boldsymbol{X}$ nazywamy często *marginesem*.

Naszą nową funkcją do minimalizacji jest więc sumaryczny błąd logistyczny:

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

W celu uniknięcia osobliwości hesjanu dodaje się, podobnie jak w przypadku regresji liniowej, dodatkowy człon z *regularyzacją*:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n}\log\left(1 + e^{-y_i\boldsymbol{w}^\top\boldsymbol{x}_i}\right) + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2.$$

W celu minimalizacji $L(\boldsymbol{w})$, wyznaczamy gradient i hesjan:

$$\nabla L(\boldsymbol{w}) = \lambda\boldsymbol{w} - \sum_{i=1}^{n} y_i\boldsymbol{x}_i\beta_i, \qquad \boldsymbol{H}(\boldsymbol{w}) = \lambda\boldsymbol{I} + \sum_{i=1}^{n}\beta_i(1-\beta_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

gdzie:

$$\beta_i = \frac{1}{1 + e^{y_i\boldsymbol{w}^\top\boldsymbol{x}_i}}.$$

Dokładne wyprowadzenie znajduje się na slajdach z wykładu. Metoda Newtona-Raphsona prowadzi do następującego algorytmu:

1. Zaczynamy od $\boldsymbol{w}_0 = \boldsymbol{0}$,
2. W kolejnych iteracjach $t = 1, 2, \ldots$ aż do zbieżności:
   (a) Wyznaczamy współczynniki $\beta_i = \frac{1}{1+e^{y_i\boldsymbol{w}_{t-1}^\top\boldsymbol{x}_i}}$.
   (b) Wyznaczamy gradient $\nabla L(\boldsymbol{w}_{t-1}) = \lambda\boldsymbol{w}_{t-1} - \sum_{i=1}^{n} y_i\boldsymbol{x}_i\beta_i$.

(c) Wyznaczamy hesjan $\boldsymbol{H}(\boldsymbol{w}_{t-1}) = \lambda \boldsymbol{I} + \sum_{i=1}^{n} \beta_i(1-\beta_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top$.

(d) Robimy krok Newtona-Rapshona uaktualniając wektor wag:

$$\boldsymbol{w}_t := \boldsymbol{w}_{t-1} - \boldsymbol{H}(\boldsymbol{w}_{t-1})^{-1}\nabla L(\boldsymbol{w}_{t-1}).$$

Jak wiele iteracji robimy? Najlepiej pod koniec każdej iteracji wyznaczać błąd logistyczny i zatrzymać algorytm, gdy błąd zacznie spadać w niewielkim stopniu. Należy zauważyć, że problem bardzo przypomina problem regresji liniowej, z wyjątkiem tego, że pojawiają się współczynniki β_i, oraz tego, że optymalizację trzeba powtórzyć wielokrotnie.

3 Uwagi o implementacji

Kolejne zadania laboratorium będą wykorzystywały napisaną przez Was implementację metody regresji logistycznej. Przy implementacji należy zwrócić uwagę na następujące rzeczy:

- Należy napisać program tak, aby przyjmował na wejściu dowolny zbiór danych w określonym formacie (można użyć skryptu do wczytywania danych z poprzednich zajęć).
- Tak jak w przypadku regresji liniowej, naturalne jest wczytanie danych jako macierz $\boldsymbol{X}$ o rozmiarze $n \times m$, której poszczególne wiersze to obserwacje $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (tak wczytuje to powyższy skrypt).
- Warto również zdefiniować macierz diagonalną $\boldsymbol{B}$ or rozmiarze $n \times n$, gdzie $\boldsymbol{B}_{ii} = \beta_i$ (poza diagonalą są same zera).
- Wtedy hesjan można zapisać jako:

$$\boldsymbol{H}(\boldsymbol{w}_{t-1}) = \lambda \boldsymbol{I} + \boldsymbol{X}^\top \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{X} \ldots$$

- ... ale uwaga: do macierzy $\boldsymbol{X}$ po wczytaniu należy dodać kolumnę jedynek aby uwzględnić wyraz wolny!
- Podobnie, naturalne jest wczytanie wartości zmiennej wyjściowej jako wektora $\boldsymbol{y}$ (tak wczytuje to powyższy skrypt). Wtedy gradient można zapisać jako jako:

$$\nabla L(\boldsymbol{w}_{t-1}) = \lambda \boldsymbol{w}_{t-1} - \boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{y}.$$

4 Sztuczne zbiory danych

Celem zadania jest implementacja metody regresji logistycznej zgodnie z uwagami wymienionymi powyżej, a następnie użycie jej na dwóch małych, sztucznych zbiorach danych: `pierwszy zbiór` oraz `drugi zbiór`. Format obu plików jest taki sam: pierwszy wiersz zawiera nazwy zmiennych oddzielone spacją, a każdy kolejny wiersz to opis jednej obserwacji z listą wartości zmiennych (również oddzielone spacją). *Wskazówka:* Dla pierwszego zbioru danych powinniście dostać współczynniki (przy $\lambda = 0$): $w_0 = 0.5873, w_1 = -1.1723, w_2 = -0.7065$.

Odpowiedz na następujące pytania:

1. Jakie są wartości współczynników $\boldsymbol{w}$ bez regularyzacji $\lambda = 0$ i z regularyzacją?
2. Czy współczynniki zależą od małej regularyzacji w sposób istotny? Tzn. zmniejszając regularyzację do zera, czy współczynniki zmieniają się mocno?