Regresja liniowa – metoda najmniejszych kwadratów

Techniki Optymalizacji

29-30.10.2018

1 Problem regresji

Przewidywania/wyjaśnienie zmian jednej zmiennej (Y) pod wpływem zmian innych zmiennych ($\boldsymbol{X} = (X_1, \ldots, X_m)$). W problemie regresji zmienna Y jest *ciągła*. Po co nam regresja? Zmienne $\boldsymbol{X}$ zwykle łatwe do pozyskania, Y – trudne lub niemożliwe do pozyskania. Przykładowo: $\boldsymbol{X}$ – ceny akcji w ostatnim tygodniu, Y – cena akcji jutro; lub: $\boldsymbol{X}$ – wyniki testów medycznych, Y – poziom zaawansowania choroby.

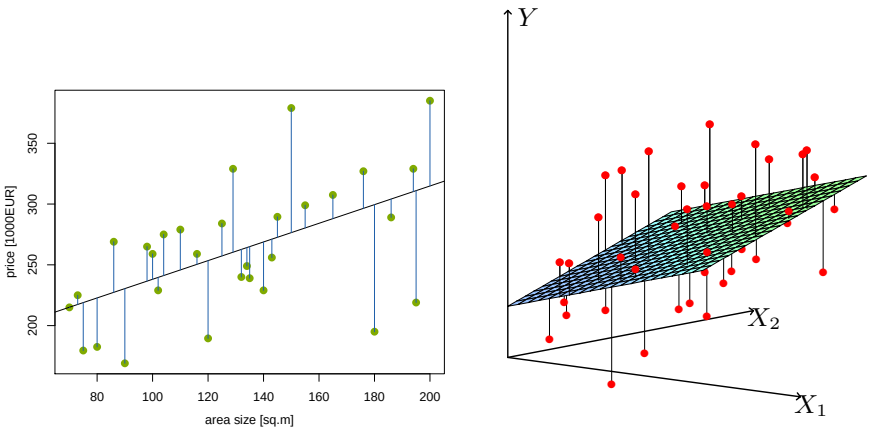
Dzisiaj zajmiemy się *regresją liniową*, kiedy modelujemy zmienną Y jako funkcję liniową $\boldsymbol{X}$:

$$\hat{Y} = w_0 + w_1 X_1 + w_2 X_2 + \ldots w_m X_m = \boldsymbol{w}^\top \boldsymbol{X}, \qquad \text{gdzie} \quad \boldsymbol{X} = (1, X_1, \ldots, X_m).$$

Zwykle otrzymujemy zbiór danych składający się n obserwacji, dla których znamy wartość zmiennej Y:

$$\begin{array}{lcl} (x_{11}, x_{12}, \ldots, x_{1m}, y_1) & & (\boldsymbol{x}_1, y_1) \\ (x_{21}, x_{22}, \ldots, x_{2m}, y_2) & & (\boldsymbol{x}_2, y_2) \\ \ldots & \text{lub w skrócie} & \ldots \\ (x_{n1}, x_{n2}, \ldots, x_{nm}, y_n) & & (\boldsymbol{x}_n, y_n) \end{array}$$

Próbujemy dopasować funkcję liniową (czyli wektor $\boldsymbol{w}$) do zbioru uczącego, tak aby wartości przewidywane przez model $\hat{y}_i$ były jak najbliżej zaobserwowanych wartości y_i.



Rysunek 1: Odchylenia, tj. różnice między wartością zaobserwowaną a przewidywaną (zaznaczone liniami pionowymi). Przypadek jednej (lewy obrazek) i dwóch zmiennych wejściowych (prawy obrazek). Prawy obrazek z Hastie, Tibshirani, Friedman: *Elements of Statistical Learning*.

Odchyleń jest n, a my chcemy mieć jedną funkcję celu. Najpopularniejszą funkcją celu jest *suma kwadratów odchyleń*:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2.$$

Chcemy minimalizować $L(\boldsymbol{w})$. W tym celu możemy użyć metody Newtona-Raphsona, która rozwiąże ten problem w jednym kroku (dlaczego? przypomnę potem na wykładzie), rozpoczynając od dowolnego $\boldsymbol{w}_0$. Liczymy gradient i hesjan:

$$\nabla L(\boldsymbol{w}) = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i, \qquad \boldsymbol{H}(\boldsymbol{w}) = \sum_{i=1}^{n} 2\boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

i używamy metody Newtona-Raphsona:

$$\boldsymbol{w} = \boldsymbol{w}_0 - \big(\boldsymbol{H}(\boldsymbol{w}_0)\big)^{-1} \nabla L(\boldsymbol{w}_0),$$

z czego po podstawieniu $\boldsymbol{w}_0 = \boldsymbol{0}$ (dla wygody, dowolne $\boldsymbol{w}_0$ by dało ten sam wynik) otrzymujemy:

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

Osobliwość hesjanu. Metoda Levenberga-Marquada. Jeśli hesjan jest osobliwy, możemy dodać do niego macierz jednostkową przemnożoną przez stałą λ. Jest to metoda Levenberga-Marquada. Tak, jak metoda Newtona-Raphsona odpowiada zwykłej regresji liniowej z minimalizacją sumy kwadratów odchyleń, tak metoda Levenberga-Marquada odpowiada *regresji grzbietowej* (*ridge regression*). W regresji grzbietowej wychodzimy z problemu minimalizacji:

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

Liczymy gradient i hesjan:

$$\nabla L(\boldsymbol{w}) = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i + 2\lambda \boldsymbol{w}, \qquad \boldsymbol{H}(\boldsymbol{w}) = \sum_{i=1}^{n} 2\boldsymbol{x}_i \boldsymbol{x}_i^\top + \lambda \boldsymbol{I},$$

i używając metody Newtona-Raphsona po podstawieniu $\boldsymbol{w}_0 = \boldsymbol{0}$ otrzymujemy:

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

Dostajemy krok metody Levenberga-Marquada. Tak jak poprzednio, rozwiązujemy problem w jednym kroku.

2 Uwagi o implementacji

Kolejne zadania laboratorium będą wykorzystywały napisaną przez Was implementację metody regresji liniowej poprzez minimalizację sumy kwadratów odchyleń, zgodnie z tym co opisano powyżej. Przy implementacji należy zwrócić uwagę na następujące rzeczy:

- Należy napisać program tak, aby przyjmował na wejściu dowolny zbiór danych w określonym formacie.
- Tu znajduje się prosty skrypt napisane w Octave, który można użyć do odczytania dowolnego pliku używanego na tych zajęciach.

- 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). Czyli $(\boldsymbol{X})_{i,j}$ to wartość j-tej zmiennej wejściowej dla i-tej obserwacji. Wtedy macierz $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ (główny element hesjanu) można zapisać jako $\boldsymbol{X}^\top \boldsymbol{X} \ldots$
- ... 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 $\sum_{i=1}^{n} y_i \boldsymbol{x}_i$ można zapisać jako $\boldsymbol{X}^\top \boldsymbol{y}$ (główny element gradientu).
- Trzeba się upewnić, czy hesjan nie jest osobliwy, jeśli jest, należy użyć metody Levenberga-Marquada. Jak sprawdzić osobliwość macierzy?
- Można użyć `tego zbioru danych` do przetestowania swojego programu. Hesjan nie jest tu osobliwy, a wartości współczynników to: $w_0 = 0.33795, w_1 = 0.97572, w_2 = 0.60207$. Z kolei `ten zbiór danych` powinien dać osobliwy hesjan, a wartości współczynników dla $\lambda = 10^{-5}$ to $w_0 = 0.036364, w_1 = -0.020053, w_2 = -0.033422$.

3 Sztuczne zbiory danych

Celem zadania jest implementacja metody regresji liniowej poprzez minimalizację sumy kwadratów odchyleń (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).

Odpowiedz na następujące pytania:

1. Jakie są wartości współczynników $\boldsymbol{w}$?
2. Czy hesjan był osobliwy? Jeśli tak to dlaczego i jak można sobie z tym poradzić?

4 Rzeczywisty zbiór danych

Celem zadania jest użycie metody regresji liniowej zaimplementowanej w poprzednim zadaniu na zbiorze danych wyceny domów w Den Bosch. Zbiór zawiera 119 domów, opisanych za pomocą 10 zmiennych (9 wejściowych i jedna wyjściowa – cena domu). Więcej o problemie, w tym opis zmiennych, można znaleźć na slajdach z wykładu. Zbiór został podzielony na dwie części – treningową (80 domów) i testową (39 domów), które można pobrać z: `zbiór treningowy` oraz `zbiór testowy`.

Format obu plików jest tak sam: pierwszy wiersz zawiera nazwy zmiennych oddzielone spacją, a każdy kolejny wiersz to opis jednego domu, z listą wartości zmiennych (również oddzielone spacją). Zmienna wyjściowa, którą będziemy modelować ("price") znajduje się na samym końcu listy, pozostałe zmienne są zmiennymi wejściowymi.

Należy zastosować metodę regresji liniowej poprzez minimalizację sumy kwadratów odchyleń *do zbioru treningowego*, uzyskując wartości współczynników. Następnie, należy użyć wartości współczynników do *predykcji* wartości cen domów na zbiorze testowym. Odpowiedz na następujące pytania:

1. Jakie są wartości współczynników $\boldsymbol{w}$? Jak można je interpretować?
2. Czy hesjan był osobliwy?
3. Jaka jest średnia wartość błędu na zbiorze treningowym, a jaka na zbiorze testowym? Spróbuj zinterpretować różnicę.

Uwaga: jako średnią wartość błędu na danym zbiorze przyjmij tak zwany *pierwiastkowany średni błąd kwadratowy* (*root mean squared error*), wyznaczony jako:

$$\mathrm{err} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Jeśli obserwacje znajdują się jako wiersze w macierzy $\boldsymbol{X}$, to wektor predykcji można wyznaczyć jako $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}$.

Zadanie dla chętnych: Jak wybór wartości λ wpływa na wynik (sumaryczny błąd) na zbiorze treningowym i testowym? Rozwiąż problem dla wielu wartości λ wyznaczonych jako kolejne potęgi dwójki, np. $\lambda = 2^{-15}, 2^{-14}, 2^{-13}, \ldots$, aż do $\lambda = 2^{20}$. Dla każdej takiej wartości wyznacz błąd na zbiorze treningowym i testowym, a następnie umieść oba błędy na wykresie w zależności od λ (na osi X zastosuj skalę logarytmiczną). Zinterpretuj wyniki.

5 Pytania

Odpowiedz na następujące pytania:

- Jaka będzie wartość współczynnika regresji, gdy mamy tylko jedną cechę wejściową x (i brak wyrazu wolnego)? W szczególności, jak wygląda taki współczynnik, gdy cecha wejściowa jest stale równa jeden?
- Jak będzie wyglądał ten współczynnik w przypadku regresji grzbietowej? Zinterpretuj zmiany.
- Spróbuj wyznaczyć wzór na prostą regresję liniową jednej zmiennej z wyrazem wolnym:

$$w_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \qquad w_0 = \bar{y} - w_1\bar{x}.$$

- Rozważ regresję poprzez minimalizację sumy wartości bezwzględnych. Spróbuj udowodnić, że jeśli mamy tylko wyraz wolny, to przewidywana wartość $\hat{y}$ będzie równanie medianie z danych.