Modelowanie Matematyczne - Lab 2

Adam Przemysław Chojecki

21 maja 2024

Zadanie 1: Jakość predykcji poza zbiorem uczącym

To zadanie jest modyfikacją zadania 1 z poprzedniego tygodnia.
Zadanie zakłada, że mamy dane pochodzące z pewnej kombinacji liniowej znanych funkcji, ale zostały one zaszumione. Chcemy poznać współczynniki, które stoją przy danej funkcji.
Taka sytuacja występuje często w modelowaniu zjawisk fizycznych. Jeśli wiemy jaki rodzaj zależności występuje między zmiennymi i chcemy poznać współczynnik. Np. jeśli chcemy poznać przyspieszenie ziemskie ($g \approx 9.8 \frac{m}{s^2}$) to możemy zmierzyć prędkość spadania obiektów o znanej masie. Pomiar ten będzie obarczony błędem, ale można go wykorzystać to przybliżenia wartości g w miejscu pomiaru.

1. Wygeneruj n = 100 liczb losowych $x \sim U([-20, -10] \cup [10, 20])$.
2. Przekształć te liczby według wzoru:

$$y = f(x) + \epsilon$$

$$f(x) = b_1 + b_2 x + b_3 x^2 + b_4 \sin(x) + b_5 \cos(x) + b_6 |x|$$

   gdzie ϵ to błąd losowy z rozkładu normalnego o średniej 0 i odchyleniu standardowym sd = 10, a wektor współczynników jest dany jako:

$$b = (120.5, 5.1, -0.4, 1.3, 10.1, 0.0)$$

3. Narysuj wykres przedstawiający prawdziwą funkcję $f(x)$, którą chcemy estymować. Jej dziedziną jest przedział $[-40, 40]$. Do wykresu dodaj punkty, których użyjemy do estymacji.
4. Zamodeluj liniowo zakładając, że znamy funkcje, z których $f(x)$ jest "zbudowany":

$$y \sim x + I(x^2) + \sin(x) + \cos(x)$$

   Zwróć uwagę, że z powodów technicznych zmienna $I(x^2)$ musi być obłożona funkcją $I()$, jeśli ma być użyta w funkcji `lm()`.

   Narysuj dodatkową linię pokazującą wyestymowaną $\hat{f}(x)$.

- Czy wyestymowany $\hat{f}(x)$ jest bliski prawdziwemu $f(x)$? Czy jest bliski tylko na zbiorze uczącym, czy może dobrze się uogólnia poza zbiór uczący?

5. Policz SSE dla Twojego modelu.

6. Zamodeluj te dane używając modelu, gdzie zamiast części x^2 będziemy mieli część $abs(x)$. Dorysuj go do obrazka. Policz SSE.

   - Czy model z $abs(x)$ zamiast x^2 jest podobny do modelu z x^2 na przedziale $[-20, -10] \cup [10, 20]$? Jak ma się to do ich wartości SSE?
   - Co się dzieje na pozostałej części dziedziny funkcji $f(x)$, czyli poza zbiorem uczącym? Która funkcja uogólnia się lepiej (jest lepszym estymatorem funkcji $f(x)$)?

7. Zamodeluj te dane używając modelu:

$$y \sim x + I(x^2) + I(x^3) + \ldots + I(x^k)$$

   gdzie k jest jakąś stałą.

   - Co się dzieje z SSE, gdy zwiększa się liczbę k?
   - Co się będzie działo z SSE, gdy k będzie zbliżało się do n?
   - Co się będzie działo z SSE, gdy $k \geqslant n$?
   - Znajdź najmniejsze k, żeby SSE tego modelu było mniejsze niż SSE modelu z punktu 4.

8. Narysuj wykres wyestymowanej funkcji na podstawie k z ostatniej kropki z punktu 7.

   - Co się dzieje na przedziale $[-20, -10] \cup [10, 20]$?
   - Co się dzieje poza przedziałem $[-20, -10] \cup [10, 20]$?

9. Zamodeluj te dane używając modelu:

$$y \sim x + I(x^2) + \sin(x) + \cos(x) + abs(x)$$

   Czyli tak jak na pierwszych naszych zajęciach. Wywołaj na tym modelu `summary()`.

   - Czy analiza modelu poprawnie zinterpretowała nieistotność jednej ze zmiennych?
   - Czy poprawnie zidentyfikowała, która zmienna jest nieistotna?
   - Zinterpretuj wynik.

Uwagi:

1. Możesz wygenerować liczby z przedziału $x \sim U([0, 1])$, a następnie przesunąć je odpowiednio.

2. Tym razem nie zastawiłem na was pułapki. Jest tyle b_i ile trzeba użyć...

3. Stwórz zmienną dla osi ox używając funkcji `seq(..., length.out = duzo)`.

   Rysując wykres funkcji $f(x)$ użyj

   `plot(..., type = "l", ylim = c(-726, 400))`, aby rozszerzyć oś OY.

   Dodając punkty możesz użyć przeźroczystości:

   - `install.packages("scales")`
   - `points(x, y, pch = 16, col = scales::alpha("red", 0.2))`

4. Do funkcji `predict()` można dołożyć nowe dane:

   `predict(model, data.frame(x=ox))`.

5. $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Wartość będzie zależała od losowości, ale u mnie wyszło około 8500 i taki mniej więcej powinien być rząd wielkości.

6. Dorysowując użyj funkcji `lines(..., col = 'najpiekniejszykolor')`.

7. Wartość k z ostatniej kropki będzie zależała od losowości, ale u mnie jest to $k = 8$.

8. Zwróć uwagę na różnicę w wyborze k parzystego od k nieparzystego.

9. `summary()`

Uwaga ogólna:
Taki wykres (z punktami, dziedziną i modelami w przestrzeni) jest bardzo pomocny i widać na nim jak na dłoni chaotyczne zachowania modeli poza zbiorem uczącym. Niestety, takie wykresy są możliwe do narysowania tylko co najwyżej w 2D (w ramach powyższego zadania robiłxś to w 1D). W wyższych wymiarach (gdy mamy więcej niż 2 kolumny, z których będziemy predykować) trzeba posługiwać się innymi metodami wizualizacji.
Nie mniej jednak, gdy mamy tylko 1 albo 2 zmienne zależne (ewentualnie z dodatkowymi przekształceniami), to można znalezioną zależność narysować, aby przekonać się, że jest ona modelem pożądanym przez nas.

Zadanie 2: Analiza istotności zmiennych w modelach liniowych

W tym zadaniu będziemy analizować istotność zmiennych w modelach liniowych, korzystając ze wbudowanego w R zbioru danych `iris`. Zmiennymi `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` są w centymetrach.

1. Wczytaj zbiór danych `iris` używając funkcji `data(iris)`.

2. Wyświetl pierwsze kilka wierszy danych, aby zapoznać się ze strukturą zbioru, używając funkcji `head(iris)`.

3. Zidentyfikuj dostępne zmienne w zbiorze danych oraz ich typy, używając funkcji `str(iris)`.

4. Dopasuj model liniowy, w którym zmienną zależną będzie `Petal.Length`, a zmiennymi niezależnymi będą `Sepal.Length`, `Sepal.Width`, `Sepal.Length * Sepal.Width`, `Petal.Width` oraz `Species`. **Pamiętaj o użyciu funkcji `I()` do obudowania mnożenia zmiennych!**

5. Wyświetl podsumowanie modelu, używając funkcji `summary()`.

6. Przeanalizuj wyniki testów istotności (gwiazdki) dla poszczególnych zmiennych w modelu. Czy jest jakaś zmienna nieistotna? Jak to interpretować?

7. Stwórz nowy model, w którym nie będzie jednej ze zmiennych w poprzednim punkcie uznanej za nieistotną. Czy teraz wszystkie zmienne są istotne? Jeśli wciąż jest jakaś zmienna nieistotna, to stwórz kolejny model, w którym pozbywasz się jej, aż wszystkie zmienne będą istotne.

8. Wróć do modelu z punktu 4. Tam były 2 nieistotne zmienne. Jedną z nich usunąłxś w punkcie 7, teraz usuń drugą z nich i kontynuuj usuwanie tak jak w punkcie 7, aż będziesz mieć/miała model z samymi zmiennymi istotnymi. Czy ostatecznie orzymałxś ten sam model co w punkcie 7?

9. Zmień jednostki zmiennej `Petal.Width` z centymetrów na cale (1 cm = 0.393701 cala), pozostawiając wszystkie pozostałe zmienne w centymetrach. Dopasuj nowy model liniowy i wyświetl jego podsumowanie. Odpowiedz na pytania:

   - Czy zmiana jednostki zmiennej `Petal.Width` wpłynęła na istotność tej zmiennej w modelu?
   - Czy zmiana jednostki zmiennej `Petal.Width` wpłynęła na współczynnik stojący przy tej zmiennej w modelu?

10. Zmodyfikuj zmienną `Petal.Width` w centymetrach, mnożąc ją przez 1,000,000. Dopasuj nowy model liniowy i wyświetl jego podsumowanie. Odpowiedz na pytania:

- Czy zmiana wartości zmiennej `Petal.Width` wpłynęła na istotność tej zmiennej w modelu?
- Czy zmiana wartości zmiennej `Petal.Width` wpłynęła na współczynnik stojący przy tej zmiennej w modelu?

11. Dla modeli z punktu 7 i 8 wykonaj wykres, gdzie na osi OX będą przewidywane wartości: $\hat{y}_i$, a na osi OX rezydua: $e_i = y_i - \hat{y}_i$.

    Przy pomocy funkcji `abline()` narysuj linię, która świadczyłaby o idealnym dopasowaniu.

    - Czy punkty są równomiernie rozłożone na osi OX? O czym to świadczy?
    - Czy z tego wykresu można wyczytać, że jakaś obserwacja nie pasuje do modelu liniowego?
    - Jakie inne wnioski można wyciągnąć?