Analiza Dużych Zbiorów Danych

Błą predykcji i kryteria informacyjne

Wygeneruj macierz planu $X_{n\times 950}$ z $n = 1000$ tak, że jej elementy są niezależnymi zmiennymi losowymi z rozkładu $N\left(0, \sigma = \frac{1}{\sqrt{1000}}\right)$. Następnie wygeneruj wektor Y zgodnie ze wzorem

$$Y = X\beta + \epsilon \ ,$$

gdzie $\beta_1 = \ldots = \beta_{20} = 3.5, \ \ \beta_{21} = \ldots = \beta_{950} = 0$ a $\epsilon \sim N(0, I)$.

1. Wykonaj poniższe analizy dla modeli zbudowanych w oparciu o

   i) 10 pierwszych zmiennych
   ii) 20 pierwszych zmiennych
   iii) 30 pierwszych zmiennych
   iv) 100 pierwszych zmiennych
   v) 500 pierwszych zmiennych
   vi) wszystkich 950 zmienych.

   – Dla każdego z powyższych modeli
   a) Wyestymuj β metodą najmniejszych kwadratów i wyznacz $RSS = ||\hat{Y} - Y||^2$ oraz wylicz oczekiwaną wartość błędu predykcji

   $$PE = E_{\epsilon^\star}||X(\beta - \hat{\beta}) + \epsilon^\star||^2 \ ,$$

   gdzie $\epsilon^\star \sim N(0, I)$ jest wektorem niezależnym od próby treningowej.
   b) Użyj RSS do estymacji PE wykorzystując prawdziwą wartość σ i zastępując ją jej klasycznym nieobciążonym estymatorem.
   c) Wyestymuj PE stosując walidację krzyżową typu "leave-one-out" (wykorzystaj wzór podany na wykładzie).
   – Wybierz optymalny model stosując powyższe estymatory PE.

   – Powtórz powyższe analizy 100 razy i dla każdego z powyższych modeli porównaj wykresy pudełkowe wartości $\hat{PE} - PE$ dla trzech wyżej wymienionych estymatorów PE.

2. Zastosuj BIC, AIC, RIC, mBIC i mBIC2 (możesz użyć biblioteki *bigstep* w R) do identyfikacji istotnych zmiennych w bazach danych składających się z

   i) pierwszych 50 zmiennych
   ii) pierwszych 200 zmiennych
   iii) pierwszych 500 zmiennych
   iv) wszystkich 950 zmiennych.

   a) Podaj liczę prawdziwych i fałszywych odkryć i kwadratowy błąd estymacji wektora $EY = X\beta$:

   $$SE = ||X\hat{\beta} - X\beta||^2 \ .$$

   b) Powtórz punkt a) 100 razy i podaj wyestymowaną moc, FDR i średni błą kwadratowy estymacji $EY = X\beta$ dla wszystkich powyższych kryteriów. Krytycznie omów uzyskane wyniki.

3. Powtórz zadania 1 i 2 w sytuacji gdy $n = 5000$ (X_{ij} są ciągle generowane z $N(0, \frac{1}{\sqrt{1000}})$).

Malgorzata Bogdan