Analiza Dużych Zbiorów Danych

Lista 2 - Wielokrotne testowanie, estymatory Jamesa-Steina

Wygeneruj ortonormalną macierz planu $X_{1000\times1000}$, tzn. taką macierz, że $X^T X = I_{1000\times1000}$.

Następnie wygeneruj wektor współczynników regresji jako ciąg niezależnych zmiennych losowych z rozkładu

$$\beta_i \sim (1-\gamma)\delta_0 + \gamma\varphi(0,\tau^2) \ ,$$

gdzie δ_0 jest rozkładem skupionym w 0 a $\varphi(0,\tau^2)$ jest gęstością rozkładu normalnego $N(0,\tau^2)$.

Rozważ 6 przypadków:

$$\gamma \in \{0.01, 0.05, 0.1\}, \qquad \tau \in \{1.5\sqrt{2\log 1000}, 3\sqrt{2\log 1000}\}.$$

Dla każdego z tych przypadków wygeneruj wektor odpowiedzi

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

gdzie $\epsilon \sim N(0, I_{1000\times1000})$ i przeprowadź poniższe analizy, zakładając, że wariancja błędu jest znana ($\sigma^2 = 1$).

1. **a)** Podaj wzór na estymator najmniejszych kwadratóow $\hat{\beta}^{LS}$ dla wektora β i rozkład tego estymatora.

   b) Skonstruuj oba estymatory Jamesa-Steina dla wektora parametrów β (tzn. estymator ściągający do zera i do wspólnej średniej).

2. Zastosuj następujące procedury do ustalenia które zmienne są istotne:

   a) procedurę Bonferoniego;

   b) procedurę Benjaminiego-Hochberga;

   c) klasyfikator Bayesowski przy założeniu tej samej funkcji straty za błąd pierwszego i drugiego rodzaju.

3. Następnie dla każdej procedury z punktu 2) wyznacz "ucięte" estymatory wektora β

$$\hat{\beta}_i^{uc} = \begin{cases} \hat{\beta}^{LS}, & \text{jeżeli odrzucono } H_{0i} : \mu_i = 0; \\ 0, & \text{w przeciwnym wypadku} \ . \end{cases}$$

4. **a)** Estymatory z punktów $1), 3)$ (6 estymatorów) porównaj pod kątem błędu kwadratowego

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

b) Procedury testowania z punktu 2) porównaj pod kątem sumy liczby błędów pierwszego i drugiego rodzaju.

c) Dla każdej kombinacji ϵ i τ powtórz doświadczenie 1000 razy i porównaj analizowane estymatory pod kątem MSE = E(SE) a analizowane procedury pod kątem wartości oczekiwanej sumy liczby błędów pierwszego i drugiego rodzaju.