dr Wojciech Rejchel
Uniwersytet Mikołaja Kopernika
Wydział Matematyki i Informatyki

Szybka i odporna selekcja cech w modelach wysokowymiarowych

Selekcja cech jest zagadnieniem ważnym, zwłaszcza gdy badamy wysokowymiarowe zbiory danych, w których liczba cech znacząco przekracza liczbę obserwacji. W wielu przypadkach znalezienie małego zbioru złożonego z cech istotnych jest równie ważne, bądź ważniejsze, jak poprawna estymacja czy predykcja.

Rozważamy problem selekcji cech w modelu

$$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n,$$

gdzie $Y_i \in \mathbb{R}$ jest zmienną zależną, $X_i \in \mathbb{R}^p$ wektorem cech, β prawdziwym parametrem oraz ε_i błędem losowym. Zakładamy, że nieznana funkcja g jest rosnąca względem pierwszej zmiennej. Rozkład błędów ε_i jest dowolny, w szczególności nie wymagamy istnienia jego momentów.

Proponujemy prostą i obliczeniowo szybką procedurę selekcji cech, która oparta jest na standardowym algorytmie Lasso ze zmiennymi Y_i zastąpionymi przez ich rangi. Przedstawimy teoretyczne i numeryczne wyniki dotyczące zgodności w wyborze modelu naszych metod.

Wspólna praca z Małgorzatą Bogdan (Uniwersytet Wrocławski).