12 Regresja nieparametryczna CD

Zadanie 1

Dla danych wygenerowanych na poprzednim laboratorium dopasuj krzywe używając funkcji: `locpoly`, `loess`, `ksmooth` (pakiety `KernSmooth`, do wybrania parametru `bandwidth` można użyć funkcji `dpill`).

Zadanie 2 (GAM)

W pakiecie `mlbench` znajduje się zbiór `BostonHousing`. Podziel zbiór na treningowy (70% obserwacji) i testowy.

1. Dopasuj model liniowy zmiennej `medv` od zmiennych `lstat` i `tax`. Na zmienną `lstat` nałóż naturalny spline kubiczny.
2. Narysuj wykres przedstawiający zależność zmiennej `medv` od zmiennej `lstat` wyestymowanej przez model dla mediany wartości zmiennej `tax`.
3. Policz MSE dla predykcji na zbiorze testowym. Porównaj z MSE dla modelu liniowego bez spline'ów.
4. Dopasuj model liniowy zmiennej `medv` od wszystkich zmiennych numerycznych. Na wybrane przez siebie zmienne nałóż spline'y. Dokonaj wyboru zmiennych dowolną metodą eliminacji krokowej. Porównaj MSE dopasowanych modeli.
5. Dopasuj uogólniony model addytywny zmiennej `medv` od zmiennych `lstat` i `tax`, tzn. `medv~s(lstat)+tax` (pakiet `mgcv`, funkcja `gam`). `s(lstat)` oznacza, że na wybraną zmienną nakładamy spline, ale nie podajemy położenia węzłów, bo się same dopasują (poprzez maksymalizację funkcji wiarogodności z karą). Narysuj wykres przedstawiający zależność zmiennej `medv` od zmiennej `lstat` wyestymowanej przez model dla mediany wartości zmiennej `tax`.

Zadanie 3 (MARS)

Przetestuj działanie metody MARS (Multivariate Adaptive Regression Splines) na danych symulacyjnych. Implementacja metody MARS znajduje się w pakiecie `earth`. Rozważamy następujące schematy symulacji:

1. Przykład 1:
   - Liczność próbki $n = 1000$, liczba zmiennych $p = 50$,
   - $X_1 \sim U(0, 4)$, $X_2, \ldots, X_{50} \sim N(0, 1)$, $\epsilon \sim N(0, 0.1)$,
   - Zmienna odpowiedzi $Y = \sqrt{X_1} + \epsilon$.
2. Przykład 2:
   - Liczność próbki $n = 1000$, liczba zmiennych $p = 50$,
   - $X_1 \sim U(0, 4)$, $X_2, \ldots, X_{50} \sim N(0, 1)$, $\epsilon \sim N(0, 0.1)$,
   - Zmienna odpowiedzi $Y = X_1^2 + \epsilon$.

3. Przykład 3:
    - Liczność próbki $n = 1000$, liczba zmiennych $p = 50$,
    - $X_1, \ldots, X_{50} \sim N(0, 1)$, $\epsilon \sim N(0, 0.1)$,
    - Zmienna odpowiedzi $Y = (X_1 - 0)_+ + (X_1 - 1)_+ + \epsilon$.
4. Przykład 4:
    - Liczność próbki $n = 1000$, liczba zmiennych $p = 50$,
    - $X_1 \sim U(0, 4)$, $X_2, \ldots, X_{50} \sim N(0, 1)$, $\epsilon \sim N(0, 0.1)$,
    - Zmienna odpowiedzi $Y = \sin(X_1) + \epsilon$.
5. Przykład 5:
    - Liczność próbki $n = 1000$, liczba zmiennych $p = 50$,
    - $X_1, \ldots, X_{50} \sim N(0, 1)$, $\epsilon \sim N(0, 0.1)$,
    - Zmienna odpowiedzi $Y = I(X_1 < 0)$.

Polecenia:

- Dopasuj model $Y \sim X_1, \ldots, X_{50}$ używając metody MARS.
- Wypisz (korzystając z funkcji `summary()`) równanie modelu.
- W powyższych przykładach, zmienna odpowiedzi zależy tylko od jednej ze zmiennych objaśniających. Przeanalizuj działanie metod wykonując wykresy pokazujące zależność $Y \sim X_1$, $\hat{Y} \sim X_1$, $\hat{Y} \sim Y$ ($\hat{Y}$ to wartości przewidywane przez model).