Selekcja cech (Feature selection)

Zadania:

1. Problem pozornych korelacji przy bardzo dużej liczbie cech.

   Przeprowadź następujący eksperyment:

   (a) Przyjmijmy liczbę obserwacji $n = 100$, $p = 1000$.

   (b) Wygeneruj zmienną odpowiedzi y ze standardowego rozkładu normalnego (funkcja `rnorm`).

   (c) Wygeneruj niezależne cechy $x_1, \ldots, x_p$, niezależnie od y, również ze standardowego rozkładu normalnego.

   (d) Oblicz maksymalną korelację liniową między zmienną odpowiedzi a cechami, t.j. $\max_{j=1,\ldots,p} cor(x_j, y)$.

   (e) Powtórz powyższe punkty $L = 100$ razy i zbadaj rozkład wartości maksymalnych (narysuj boxplot lub histogram).

   (f) Powtórz eksperyment dla $p = 10000, 50000, 100000$.

2. Porównanie filtrów: współczynnik korelacji vs informacja wzajemna.

   Generujemy pary zmiennych x i y w następujący sposób:

   - **Przykład 1:** $x \sim U[0,1]$, $y = 2x + \epsilon$, $\epsilon \sim N(0, \sigma)$.
   - **Przykład 2:** $x \sim U[0,1]$, $y = \sqrt{x} + \epsilon$, $\epsilon \sim N(0, \sigma)$.
   - **Przykład 3:** $x \sim U[-1,1]$, $y = x^2 + \epsilon$, $\epsilon \sim N(0, \sigma)$.
   - **Przykład 4:** $x \sim U[0,6]$, $y = \sin(x) + \epsilon$, $\epsilon \sim N(0, \sigma)$.

   Zadanie:

   - Narysuj wykres który pokazuje jak współczynnik korelacji (funkcja `cor`) i informacja wzajemna (np. funkcja `information.gain` z pakietu `FSelector`) zależy od $\sigma = 0, 0.1, 0.2, \ldots, 5$. Dla każdej wartości σ oblicz uśrednioną po 50 symulacjach wartość współczynnika korelacji i informacji wzajemnej.

3. Selekcja zmiennych dla regresji logistycznej.

   Zadanie:

   - Wygeneruj dane z modelu logistycznego

     $$y_i \sim Bern(p_i),$$

     gdzie

     $$p_i = \frac{1}{1 + \exp[-(\beta_1 x_{i,1} + \ldots + \beta_p x_{i,p})]},$$

     dla $i = 1, \ldots, n$, $x_{1,i}, \ldots, x_{p,i} \sim N(0,1)$, $n = 100$, $p = 1000$. Parametry $\beta_1 = \beta_2 = \beta_3 = 1$ oraz $\beta_j = 0$, dla $j = 4, \ldots, p$.

- Wykonaj selekcję zmiennych przy użyciu trzech metod:
  - Metoda lasso (funkcja `glmnet` z pakietu `glmnet`). Optymalną wartość parametru kary λ wybierz za pomocą kroswalidacji (funkcja `cv.glmnet`).
  - Metoda krokowego dołączania zmiennych (funkcje `glm` i `step`).
  - Metoda multisplit (funkcja `multi.split` z pakietu `hdi`).
- Porównaj wybrany zbiór cech istotnych $\hat{t}$ ze zbiorem prawdziwym t (w naszym przypadku $t = \{1, 2, 3\}$) obliczając czułość i precyzje:

$$Recall(t, \hat{t}) = \frac{|t \cap \hat{t}|}{|t|},$$

$$Prec(t, \hat{t}) = \frac{|t \cap \hat{t}|}{|\hat{t}|}.$$

- Zbadaj jak czułość i precyzja zależą od p oraz n. W tym celu dla ustalonych p i n należy wykonać pewną liczbę symulacji i uwzględnić uśrednione wyniki.

4. Wpływ szumu na zmienność współczynników.

   Zadanie:

   - Zastosuj schemat generacji danych z poprzedniego zadania.
   - Dopasuj model logistyczny i oblicz współczynniki (funkcja `glm`).
   - Powtórz dwa poprzednie punkty 500 razy aby oszacować wariancję współczynnika dla pierwszej zmiennej.
   - Zbadaj jak wariancja zależy od liczby zmiennych nieistotnych.