

Algorytmiczna Analiza Danych 2025/2026

Lista ćwiczeń

Używam następujących skrótów:

1. ISL = [An Introduction to Statistical Learning](#) autorstwa G. Jamesa i in.
2. ESL = [The Elements of Statistical Learning](#) autorstwa T. Hastie i in.
3. ...

Ćwiczenie 1 — (ISL) Wyjaśnij, czy dany scenariusz jest związany z problemem klasyfikacji, czy regresji, oraz wskaż, czy bardziej interesuje nas wnioskowanie (ang. inference), czy przewidywanie (ang. prediction). Podaj rozmiar próby n i liczbę predyktorów p .

- (a) Zbieramy dane na temat 500 największych firm w Polsce. Dla każdej firmy rejestrujemy zysk, liczbę pracowników, branżę i wynagrodzenie prezesa. Interesuje nas, które czynniki wpływają na wynagrodzenie prezesa.
- (b) Rozważamy wprowadzenie nowego produktu i chcemy wiedzieć, czy odniesie sukces, czy porażkę na rynku. Zbieramy dane o 20 podobnych produktach, które zostały wcześniej wprowadzone. Dla każdego produktu zapisaliśmy, czy odniósł sukces, czy porażkę, cenę tego produktu, cenę u bezpośredniej konkurencji oraz budżet marketingowy.
- (c) Interesuje nas prognozowanie procentowej zmiany dolara amerykańskiego w stosunku do tygodniowych zmian na światowych rynkach akcji. Zbieramy dane tygodniowe dotyczące roku 2024. Dla każdego tygodnia rejestrujemy procentową zmianę dolara oraz procentową zmianę na rynku akcji w USA, Wielkiej Brytanii, Niemczech i Japonii.

Ćwiczenie 2 — (ISL) Praktyczne zastosowania uczenia maszynowego: dla każdego z poniższych punktów zaproponuj dwa niesztafpowe praktyczne zastosowania. Opisz jakie dane są potrzebne w wybranych zastosowaniach, np. jakie cechy (ang. features) mogą zostać użyte jako predyktory, a jakie odpowiedzi, czy potrzebna jest duża liczba obserwacji.

- (a) Klasyfikacja.
- (b) Regresja.
- (c) Analiza klastrow.

Ćwiczenie 3 — Opisz różnice między parametrycznym a nieparametrycznym podejściem do uczenia statystycznego. Jakie są wady i zalety podejścia parametrycznego w porównaniu do podejścia nieparametrycznego?

Ćwiczenie 4 — (ISL) Tabela poniżej przedstawia zbiór treningowy zawierający sześć obserwacji, trzy predyktory X_1, X_2, X_3 oraz jedną odpowiedź Y .

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Załóżmy, że chcemy użyć tego zbioru danych do przewidzenia wartości Y , gdy $X_1 = X_2 = X_3 = 0$, korzystając z metody K najbliższych sąsiadów.

- (a) Oblicz odległość euklidesową między każdą obserwacją a punktem testowym $X_1 = X_2 = X_3 = 0$.
- (b) Jaka jest nasza predykcja dla $K = 1$? Dlaczego?
- (c) Jaka jest nasza predykcja dla $K = 3$? Dlaczego?
- (d) Jeśli granica decyzyjna Bayesa w tym problemie jest silnie nieliniowa, to czy spodziewalibyśmy się, że najlepsza wartość K będzie duża czy mała? Dlaczego?

Ćwiczenie 5 — Załóżmy, że zmienne losowe X i Y są niezależne. Udowodnij, że

- (a) $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$,
- (b) $\mathbb{V}\text{ar}(X + Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)$.

Ćwiczenie 6 — Przypomnij, co to jest estymator oraz co to jest obciążenie (ang. bias) i wariancja estymatora. Podaj przykłady estymatorów obciążonych i nieobciążonych.

Ćwiczenie 7 — (ISL) Szukamy odpowiedniego modelu dla problemu związanego z uczeniem nadzorowanym. Dla każdego z punktów od (a) do (d), wskaż, czy ogólnie oczekivalibyśmy, że mało „elastyczny” model będzie lepszy czy gorszy od bardziej „elastycznego”. Uzasadnij swoją odpowiedź.

- (a) Rozmiar próby n jest bardzo duży, a liczba predyktorów (ang. predictor) p jest mała.
- (b) Liczba predyktorów p jest bardzo duża, a liczba obserwacji n jest mała.
- (c) Zależność między predyktorami a odpowiedzią (ang. response) wykazuje wyraźnie nieliniowy charakter.
- (d) Wariancja błędów, tj. $\sigma^2 = \mathbb{V}\text{ar}(\epsilon)$, jest bardzo wysoka.

Sformułuj wnioski dotyczące tego, jakie są zalety i wady bardziej elastycznego modelu, w porównaniu z mniej elastycznym? W jakich okolicznościach bardziej elastyczny model może być preferowany? Kiedy preferowany może być mniej elastyczny model?

Ćwiczenie 8 — (ESL, str. 223) Załóżmy, że mamy zbiór treningowy złożony z punktów $(x_1, y_1), \dots, (x_n, y_n)$. Zakładamy, że istnieje zależność $y = f(x) + \epsilon$, gdzie ϵ reprezentuje szum i jest zmienną losową o zerowej wartości oczekiwanej i wariancji σ_ϵ^2 . Używamy zbioru treningowego, aby znaleźć $\hat{f}(x)$ aproksymującą $f(x)$. Pokaż, że możemy rozłożyć oczekiwany błąd kwadratowy dla nowej wartości x_0 jako:

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \text{Bias}[\hat{f}(x_0)]^2 + \text{Var}[\hat{f}(x_0)] + \sigma_\epsilon^2.$$

Na czym polega [kompromis między obciążeniem a wariancją](#)?

Ćwiczenie 9 — Naskicuj typowy przebieg krzywej kwadratu obciążenia, wariancji, średniokwadratowego błędu treningowego, średniokwadratowego błędu testowego oraz błędu nieusuwalnego na jednym wykresie, gdy przechodzimy od mniej do bardziej elastycznych modeli $\hat{f}$. Oś x powinna reprezentować miarę elastyczności modelu (np. stopień wielomianu), a oś y powinna reprezentować wartości dla każdej z funkcji. Wyjaśnij kształt każdej funkcji.

Ćwiczenie 10 — Załóżmy, że mamy n obserwacji $\{(x_1, y_1), \dots, (x_n, y_n)\}$ i rozważamy model liniowy $y_i = \beta_0 + \beta_1 x_i + e_i$. Szacujemy parametry β_0 i β_1 poprzez minimalizację średniego błędu kwadratowego:

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

Pokaż, że w takim przypadku

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

gdzie $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ i $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Uzasadnij, że uzyskana linia zawsze przechodzi przez punkt $(\bar{x}, \bar{y})$.

Ćwiczenie 11 — Wyprowadź błąd systematyczny, wariancję i błąd standardowy dla estymatorów $\hat{\beta}_0$ i $\hat{\beta}_1$. Zakładamy, że $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \mathcal{N}(0, \sigma^2)$ oraz wszystkie e_i dla $i \in \{1, \dots, n\}$ są niezależne.

Ćwiczenie 12 — Przypomnij, jak udowodnić, że suma dwóch niezależnych zmiennych losowych o rozkładzie normalnym ma rozkład normalny.

Ćwiczenie 13 — Wykorzystując rozwiązania wcześniejszych zadań oraz regułę trzech sigm, wyjaśnij, dlaczego z około 95% prawdopodobieństwem przedział

$$\hat{\beta}_1 \pm 2\sqrt{\text{Var}(\hat{\beta}_1)}$$

zawiera prawdziwą wartość β_1 .

Ćwiczenie 14 — Pokaż, że dla modelu regresji liniowej z $k + 1$ parametrami możemy uzyskać estymatory

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$$

ze wzoru

$$\hat{\beta} = (XX^T)^{-1}X^T\vec{y},$$

gdzie X to macierz danych, a $\vec{y}$ to wektor odpowiedzi (patrz np. [tutaj](#)).

Ćwiczenie 15 — Dla $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ i $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ definiujemy R^2 jako

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Pokaż, że jeśli rozważymy model $Y = \beta_0 + \beta_1 X + \varepsilon$, to mamy

$$R^2 = \text{Corr}(X, Y)^2,$$

gdzie $\text{Corr}(X, Y)$ to współczynnik korelacji.

Ćwiczenie 16 — Wyjaśnij czym jest t-statystyka i w jaki sposób możemy jej użyć w kontekście regresji liniowej. Czym jest p-wartość?

Ćwiczenie 17 — Wyjaśnij czym jest [odległość Mahalanobisa](#) i jaki jest jej związek z *leverage statistic*.

cdn.

J.L.