Statystyka i analiza danych 2025/2026
Lista ćwiczeń

Lista 10

Ćwiczenie 59 — Ślad macierzy kwadratowej $A \in \mathbb{R}^{n\times n}$ definiujemy jako: $\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii}$.

(a) Udowodnij, że dla macierzy $A \in \mathbb{R}^{m\times n}$ oraz $B \in \mathbb{R}^{n\times m}$ zachodzi

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

(b) Korzystając z (a), pokaż własność cykliczności śladu: dla macierzy A, B, C odpowiednich wymiarów zachodzi

$$\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA).$$

Czy $\operatorname{tr}(ABC) = \operatorname{tr}(BAC)$? Jeśli nie, podaj kontrprzykład.

Ćwiczenie 60 — W oparciu o ćwiczenie 53 dla modelu regresji liniowej $\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ wartości $\hat{\mathbf{y}}$ dopasowane przez model możemy przedstawić jako

$$\hat{\mathbf{y}} = X(X^\top X)^{-1}X^\top \mathbf{y} = H\mathbf{y},$$

gdzie $H = X(X^\top X)^{-1}X^\top$ (ang. hat matrix), $X \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$. *Leverage statistic* i-tej obserwacji definiuje się jako $h_{ii} = H_{ii}$ (element na przekątnej H).

(a) W oparciu o to, że H jest symetryczna ($H^\top = H$) i idempotentna ($H^2 = H$), udowodnij, że $0 \le h_{ii} \le 1$.

(b) Korzystając z własności śladu udowodnionych w poprzednim zadaniu, oblicz $\operatorname{tr}(H)$, a następnie pokaż, że średnia wartość *leverage statistic* wynosi

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p}{n}.$$

(c) Podaj intuicyjne uzasadnienie, dlaczego obserwacja z $h_{ii} \gg \overline{h}$ może być uważana za *wpływową* w kontekście dopasowania modelu.

Ćwiczenie 61 — Wyjaśnij jaka jest intuicja związana z odległością Mahalanobisa (zobacz np. tu i tu) i jaki jest związek tej odległości z *leverage statistic*.

Ćwiczenie 62 — Wyjaśnij, czym jest naiwny klasyfikator Bayes'a i kiedy warto go zastosować. Wyjaśnij i udowodnij następujący wzór:

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{p(\mathbf{x})} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \ ,$$

gdzie

$$p(\mathbf{x}) = \sum_k \ p(\mathbf{x} \mid C_k) p(C_k) \quad \text{a} \quad \mathbf{x} = (x_1, \ldots, x_n) \ .$$

Ćwiczenie 63 — Rozważ problem klasyfikacji wiadomości e-mail na dwie klasy: Spam (C_1) lub Nie-spam (C_2). Dysponujesz dwoma cechami binarnymi:

- x_1 opisuje czy e-mail zawiera słowo *promocja*: $x_1 = 1$ jeśli tak, $x_1 = 0$ jeśli nie,
- x_2 opisuje czy e-mail zawiera słowo *oferta*: $x_2 = 1$ jeśli tak, $x_2 = 0$ jeśli nie.

Klasa Spam zawiera 100 e-maili i każdy z nich zawiera oba słowa: $x_1 = 1$ i $x_2 = 1$. Klasa Nie-spam zawiera 900 e-maili:

- 450 e-maili zawiera tylko słowo *promocja*: $x_1 = 1$, $x_2 = 0$,
- 450 e-maili zawiera tylko słowo *oferta*: $x_1 = 0$, $x_2 = 1$,

Korzystając z naiwnego klasyfikatora Bayesa (zakładającego niezależność cech), oblicz prawdopodobieństwa $p(C_1|x)$ i $p(C_2|x)$ dla nowego e-maila, który zawiera oba słowa: $x_1 = 1$, $x_2 = 1$. Czy wynik jest zaskakujący? Wyjaśnij.

Ćwiczenie 64 — Ocena modelu w problemie klasyfikacji.

a) Przedstaw i wyjaśnij metryki często stosowane do oceny modelu w problemie klasyfikacji: *accuracy*, *precision*, *recall* oraz *F1*. Podaj prosty przykład.

b) Czym się różni ich wersja *micro* od wersji *macro* (zobacz np. tutaj)? Podaj prosty przykład.

c) Jak te metryki mają się do bardziej potocznego rozumienia pojęć *accuracy* i *precision*?

cdn.

J.L.