Algorytmiczna Analiza Danych 2024/2025
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 — Załóżmy, że zmienne losowe X i Y są niezależne. Udowodnij, że

(a) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$,

(b) $\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$.

Ćwiczenie 2 — Przypomnij, co to jest obciążenie (ang. bias) i wariancja estymatora. Podaj przykłady estymatorów obciążonych i nieobciążonych.

Ćwiczenie 3 — (ESL, str. 223) Załóżmy, że mamy zbiór treningowy złożony z punktów $(x_1, y_1), \ldots, (x_n, y_n)$. 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:

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

Na czym polega kompromis między obciążeniem a wariancją?

Ćwiczenie 4 — (ISL) Dla każdego z punktów od (a) do (d), wskaż, czy ogólnie oczekiwalibyśmy, że mało „elastyczny" model będzie lepszy czy gorsza od bardziej „elastycznego". Uzasadnij swoją odpowiedź.

(a) Rozmiar próby n jest bardzo duży, a liczba predyktorów (and. predictor) p jest mała.

(b) Liczba predyktorów p jest bardzo duża, a liczba obserwacji n jest mała.

(c) Zależność między predyktorami a odpowiedzią (and. response) wykazuje wyraźnie nieliniowy charakter.

(d) Wariancja błędów, tj. $\sigma^2 = \mathbb{V}\mathrm{ar}(\epsilon)$, 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 5 — Narysuj typowy przebieg krzywej kwadratu obciążenia, wariancji, błędu treningowego, błędu testowego oraz błędu nieusuwalnego na jednym wykresie, gdy przechodzimy od mniej do bardziej elastycznych modeli. 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 6 — Opisz różnice między parametrycznym a nieparametrycznym podejściem do uczenia statystycznego. Jakie są wady i zalety podejścia parametrycznego w porównaniu do podejścia nieparametrycznego?

Ćwiczenie 7 — (ISL) Tabela poniżej przedstawia zbiór treningowy zawierający sześć obserwacji, trzy predyktory $X1, X2, X3$ 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 $X1 = X2 = X3 = 0$, korzystając z metody K najbliższych sąsiadów.

(a) Oblicz odległość euklidesową między każdą obserwacją a punktem testowym $X1 = X2 = X3 = 0$.

(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 8 — (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 9 — (ISL) Praktyczne zastosowania uczenia maszynowego: dla każdego z poniższych punktów zaproponuj dwa niesztampowe 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 klastrów.

Ćwiczenie 10 — Załóżmy, że mamy n obserwacji $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ 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:

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

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 $i \in \{1, \ldots, n\}$ są niezależne.

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

Ćwiczenie 13 — Wyjaśnij, dlaczego z około 95% prawdopodobieństwem przedział

$$\hat{\beta}_1 \pm 2\sqrt{\mathbb{V}\mathrm{ar}\left(\hat{\beta}_1\right)}$$

zawiera prawdziwą wartość β_1.

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

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

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 $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ definiujemy R^2 jako

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

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

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

gdzie $Corr(X, Y)$ to współczynnik korelacji.

Ćwiczenie 16 — Przedstaw formułę na *leverage statistic* i postaraj się podać jej intuicyjne uzasadnienie.

Ćwiczenie 17 — Wyjaśnij czym jest t-statystyka i w jaki sposób możemy jej użyć w kontekście regresji liniowej. Czym jest p-wartość?

Ćwiczenie 18 — Wyjaśnij czym jest odległość Mahalanobisa i jaki jest jej związek z *leverage statistic*.

Ćwiczenie 19 — Przypomnij, czym jest estymator największej wiarygodności. Przedstaw i wyjaśnij własności takiego estymatora (w szczególności zgodność i efektywność).

Ćwiczenie 20 — Załóżmy, że mamy n obserwacji $x_1, \ldots, x_n$ pochodzących z rozkładu normalnego zmiennej losowej $X \sim N(\mu, \sigma^2)$ o nieznanych parametrach μ i σ^2. Wyprowadź estymator największej wiarygodności dla parametru μ.

Ćwiczenie 21 — Rozważmy n niezależnych obserwacji $x_1, x_2, \ldots, x_n$, które pochodzą z rozkładu wykładniczego z nieznanym parametrem λ. Wyprowadź estymator największej wiarygodności dla parametru λ. Wskaż intuicyjnie, dlaczego ten estymator ma sens.

Ćwiczenie 22 — Rozważmy n niezależnych obserwacji $x_1, x_2, \ldots, x_n$, które pochodzą z rozkładu jednostajnego na przedziale $[0, \theta]$, gdzie θ jest nieznanym parametrem. Wyprowadź estymator największej wiarygodności dla parametru θ. Wskaż intuicyjnie, dlaczego ten estymator ma sens. Czy ten estymator jest nieobciążony?

Ćwiczenie 23 — (ISL) Gdy liczba cech p jest duża, zazwyczaj obserwuje się pogorszenie wyników metod takich jak k-najbliższych sąsiadów (KNN) i innych podejść, które dokonują prognoz w oparciu jedynie o obserwacje najbliższe obserwacji testowej, dla której dokonujemy predykcji. Zjawisko to jest znane jako przekleństwo wymiarowości i wiąże się z faktem, że metody nieparametryczne często osiągają słabe wyniki, gdy p jest duże. Przeanalizujemy teraz to zjawisko.

a) Przyjmijmy, że mamy zbiór obserwacji, z których każda posiada jeden pomiar cechy X (tzn. $p = 1$). Zakładamy, że X jest jednorodnie rozłożony na przedziale $[0, 1]$. Każdej obserwacji przypisana jest wartość odpowiedzi. Załóżmy, że chcemy przewidzieć wartość odpowiedzi dla obserwacji testowej, wykorzystując jedynie te obserwacje, które są w odległości 10% zakresu X od tej obserwacji testowej. Na przykład, aby przewidzieć odpowiedź dla obserwacji testowej z $X = 0.6$, użyjemy obserwacji z zakresu $[0.55, 0.65]$. Jaka część dostępnych obserwacji zostanie użyta do dokonania predykcji?

b) Załóżmy teraz, że mamy zbiór obserwacji, z których każda posiada dwa pomiary cech X_1 i X_2 (tzn. $p = 2$). Zakładamy, że (X_1, X_2) są jednorodnie rozłożone na $[0, 1]Ö[0, 1]$. Chcemy przewidzieć wartość odpowiedzi dla obserwacji testowej, wykorzystując tylko te obserwacje, które są w odległości 10% zakresu X_1 i w odległości 10% zakresu X_2 od obserwacji testowej. Na przykład, aby przewidzieć odpowiedź dla obserwacji testowej z $X_1 = 0.6$ i $X_2 = 0.35$, użyjemy obserwacji z zakresu $[0.55, 0.65]$ dla X_1 i z zakresu $[0.3, 0.4]$ dla X_2. Jaka część dostępnych obserwacji zostanie użyta do dokonania predykcji?

c) Załóżmy teraz, że mamy zbiór obserwacji dla $p = 100$ cech. Ponownie zakładamy, że obserwacje są jednorodnie rozłożone dla każdej cechy, a każda cecha przyjmuje wartości z zakresu od 0 do 1. Chcemy przewidzieć wartość odpowiedzi dla obserwacji testowej, używając jedynie tych obserwacji, które są w odległości 10% zakresu każdej cechy od obserwacji testowej. Jaka część dostępnych obserwacji zostanie użyta do dokonania predykcji?

d) Na podstawie odpowiedzi na części (a)–(c), uzasadnij, że jednym z ograniczeń metody KNN, gdy p jest duże, jest to, że istnieje bardzo mało obserwacji treningowych „blisko" danej obserwacji testowej.

e) Załóżmy teraz, że chcemy dokonać predykcji dla obserwacji testowej, tworząc wokół niej p-wymiarową hiperkostkę zawierającą średnio 10% obserwacji treningowych. Dla $p = 1$, 2 oraz 100, jaka jest długość każdego boku hiperkostki? Skomentuj swoją odpowiedź.

Ćwiczenie 24 — 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*?

Ćwiczenie 25 — 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 26 — 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: $x_1 = 1$, $x_2 = 1$.

Ćwiczenie 27 — Załóżmy, że mamy pojedynczy predyktor X, binarną odpowiedź Y i chcielibyśmy stworzyć model parametryczny dla $p(X) = Pr(Y = 1|X)$.

a) Możemy spróbować użyć modelu regresji liniowej. Dlaczego nie jest to dobry pomysł?

b) Jeśli $Pr(Y = 1|x) = 0.75$ jakie sa szanse (ang. odds) że dla x mamy $Y = 1$?

c) W algorytmie regresji logistycznej binarnej modelujemy prawdopodobieństwo $p(X)$ za pomocą funkcji logistycznej

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \ .$$

Udowodnij, że powyższe równanie jest równoważne z równaniem na *log-odds*:

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \ .$$

d) Jaka jest zależność między funkcją logistyczną $\sigma(x) = \frac{e^x}{1+e^x}$ i funkcją logit $l(p) = \ln\left(\frac{p}{1-p}\right)$?

Ćwiczenie 28 — Wyjaśnij, w jaki sposób możemy uzyskać wielomianowy model regresji logistycznej (ang. multinomial logistic regression) dla K klas wykorzystując $K-1$ modeli binarnej regresji logistycznej. Wskazówka: zobacz tutaj.

Ćwiczenie 29 — Opisz elementy wykresu pudełkowego (boxplot).

Ćwiczenie 30 — Podczas wykładu pokazaliśmy, że dla problemów klasyfikacji binarnej funkcja wiarygodności może być wyrażona w oparciu o entropię krzyżową. Pokaż podobny wynik dla problemów klasyfikacji wieloklasowej.

Ćwiczenie 31 — Przypomnij jak przebiega konstrukcja drzewa decyzyjnego. Załóżmy, że próbujemy zastosować drzewo decyzyjne do danych z predyktorem kategorycznym posiadającym q możliwych *nieuporządkowanych* wartości. Jaki może być problem? Wskazówka: pokaż, że istnieje $2^{q-1}-1$ możliwych podziałów q wartości na dwie grupy. Jak można próbować rozwiązać ten problem?

Ćwiczenie 32 — (ESL, strony 309-310) Dla konstrukcji drzew decyzyjnych jako miarę jakości podziału możemy zastosować odsetek błędnych klasyfikacji (ang. misclassification rate), entropię lub indeks Giniego.

a) Wyjaśnij, jak można interpretować indeks Giniego.

b) Załóżmy, że używamy drzewa decyzyjnego do problemu dwuklasowego z 400 obserwacjami w każdej klasie - oznaczmy tą sytuację jako (400, 400) - i że jeden podział utworzył węzły (300, 100) oraz (100, 300), podczas gdy drugi podział utworzył węzły (200, 400) oraz (200, 0). Pokaż, że oba podziały dają odsetek błędnych klasyfikacji równy 0,25. Drugi podział tworzy czysty węzeł i wobec tego jest preferowany. Pokaż, że indeks Giniego jest mniejszy dla drugiego podziału.

Ćwiczenie 33 — Wyjaśnij co to jest współczynnik Giniego (ang. Gini coefficient) i w jaki sposób jest wyliczany. Jaka jest pozycja Polski w rankingu krajów OECD względem współczynnik Giniego przed i po transferach podatkowych (zobacz tutaj.)

Ćwiczenie 34 — Niech X oznacza liczba unikalnych obserwacji w próbce bootstrapowej rozmiaru n. Jaka jest wartość oczekiwana i wariancja X? Jaki jest średnio odsetek unikalnych obserwacji i wariancja odsetka unikalnych obserwacji w próbie bootstrapowej dla dużych n? Jakie jest prawdopodobieństwo, że dana obserwacja nie znajdzie się w żadnej z k próbek bootstrapowych?

J.L.