Algorytmiczna Analiza Danych
Ćwiczenia 2

2025-10-14

Adrian Herda
Informatyka Algorytmiczna
Politechnika Wrocławska

1. Zadanie 1

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

1.1. Podpunkt a) wynagrodzenie prezesa a czynniki firmowe

Typ problemu: problem regresji - wynagrodzenie to zmienna ciągła
Cel: wnioskowanie - chcemy zrozumieć zależności a nie przewidzieć przyszłe zarobki prezesów
Rozmiar próby n: 500 - liczba firm
Liczba predyktorów p: 3 - zysk, liczba pracowników, branża - predyktory do dane które pozwalają na znalezienie korelacji

1.2. Podpunkt b) wprowadzenie nowego produktu na rynek

Typ problemu: problem klasyfikacji - kategoryzujemy produkty na te które odniosły sukces lub porażkę, są dwie klasy
Cel: przewidywanie - chcemy przewidzieć jak produkt poradzi sobie na rynku
Rozmiar próby n: 20 - liczba podobnych produktów
Liczba predyktorów p: 3 - cena, cena bezpośredniej konkurencji, budżet marketingowy

1.3. Podpunkt c) prognozowanie zmian kursu dolara

Typ problemu: problem regresji - procentowe zmiany kursu to wartość liniowa
Cel: przewidywanie - *„prognoza”* to próba przewidzenia danych w przyszłości
Rozmiar próby n: 52 - liczba tygodni w roku
Liczba predyktorów p: 4 - liczba porównywanych rynków akcji: USA, Wielka Brytania, Niemcy, Japonia

2. Zadanie 2

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.

2.1. Klasyfikacja

2.1.1. Przykład 1 - rozpoznawanie rodzajów wina

- Predykory:
  - gęstość / lepkość
  - zapach
  - kolor
  - smak
- Odpowiedzi
  - rodzaje wina

Potrzebna jest duża ilość obserwacji ze względu na ilość rodzajów wina oraz ze względu na podobieństwo niektórych rodzajów.

2.1.2. Przykład 2 - klasyfikacja emocji w rozmowach telefonicznych z klientami

- Predyktory:
  - ton mowy
  - tempo mowy
  - używane wyrażenia
  - głośność
- Odpowiedzi
  - odpowiedzi na pytanie o zadowolenie z obsługi (bardzo pozytywne, pozytywne, neutralne, negatywne, bardzo negatywne)

Potrzebna jest duża ilość danych ponieważ ludzkie emocje są bardzo skomplikowane i często niezrozumiałe nawet dla nas samych. Nawet do setek tysięcy

2.2. Regresja

2.2.1. Przykład 1 - Szacowanie wieku człowieka na podstawie zdjęcia (np takiego do ID)

- Predyktory:
  - kolor włosów
  - zmarszczki
  - proporcje oczu do głowy (szczególnie u dzieci)
  - kolor skóry
  - przebarwienia
- Odpowiedzi:
  - wiek

Potrzebne może być nawet do dziesiątek tysięcy obserwacji może być potrzebne gdyż fizjologia ludzka jest skomplikowana

2.2.2. Przykład 2 - Prognozowanie czasu rozkładu biodegradowalnych materiałów

- Predyktory:
  - skład chemiczny
  - gęstość
  - porowatość

  - temperatura
  - wilgotność
  - rodzaj środowiska (gleba, woda, kompost).
- Odpowiedź:
  - czas rozkładu

Potrzebne będą setki eksperymentów laboratoryjnych i terenowych.

2.3. Analiza klastrów

2.3.1. Przykład 1 - Grupowanie utworów muzyczny według klas np. nastroju czy gatunku (np. Spotify)

- Predyktory:
  - tempo
  - tonacja
  - analiza tekstu - używane słowa i wyrażenia
  - głośność
  - użyte instrumenty
  - użyte narzędzia modyfikowania dźwięku
- Odpowiedzi: brak klastrowanie odbywa się bez nadzoru i tworzy własne klasy / etykiety.
- Przykładowe etykiety nadane przez algorytm: muzyka melancholijna, wesoła, skoczna, buntownicza

Setki tysięcy lub nawet miliony utworów będą potrzebne by znaleźć widocznie wyróżniające się klastry.

2.3.2. Przykład 2 - Podział klientów sklepu na grupy o podobnych zainteresowaniach

- Predyktory:
  - średnia wartość zakupów
  - częstotliwość zakupów
  - najczęściej kupowane i przeglądane oferty
  - ocena satysfakcji klienta
  - lokalizacja klienta (może być nie dokładna, np. region - ciepły czy zimny)
- Odpowiedzi: brak klastrowanie odbywa się bez nadzoru i tworzy własne klasy / etykiety.
- Przykładowe etykiety nadane przez algorytm: klient zainteresowany ciuchami zimowymi / letnimi, wędkarz, sportowiec, kolekcjoner butów, stały klient, klient jednorazowy

Setki tysięcy klientów będą potrzebne by znaleźć wyróżniające się klastry.

3. Zadanie 3 - uczenie statystyczne - parametryczne a nieparametryczne

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

3.1. Parametryczne

Zakłada pewien rozkład danych, za pomocą skończonej (nie ogromnej) ilości parametrów. Umożliwia stosunkowo łatwą analizę modelu ale sprawia że jest ona tylko przybliżeniem. To podejście jest też mniej elastyczne i jest podatne na błędy związane z biasem.

3.2. Nieparametryczna

Nie wymaga założeń co do rozkładu danych . Model uczy się bardziej elastycznie dopasowując się do obserwacji. Dzięki temu otrzymujemy mniejszy bias ale możliwym jest zbytnio dopasować model do danych (ang. overfitting). Podejście to wymaga dużej liczby obserwacji i przez to jest bardziej kosztowne niż podejście parametryczne.

3.3. Zalety i wady

	Uczenie parametryczne	Uczenie nieparamteryczne
Zalety	• mały wariancja • łatwość analizy i interpretacji modelu • niski koszt obliczeń	• mały bias • duża elastyczność i dopasowanie do danych • Nie wymaga założeń o danych
Wady	• duży bias • mała elastyczność • słaba moc predykcyjna • wymaga założeń o danych	• duża wariancja • niezrozumiały sposób działania (black box) • duży koszt obliczeń

4. Zadanie 4 - zbiór treningowy dla metody K najbliższych sąsiadów (ang. K-nearest neighbors)

4.1. a) odległość euklidesowa dla każdego punktu

1. $\sqrt{0^2 + 3^2 + 0^2} = \sqrt{9} = 3$,
2. $\sqrt{2^2 + 0^2 + 0^2} = \sqrt{4} = 2$,
3. $\sqrt{0^2 + 1^2 + 3^2} = \sqrt{10}$,
4. $\sqrt{0^2 + 1^2 + 2^2} = \sqrt{5}$,
5. $\sqrt{(-1)^2 + 0^2 + 1^2} = \sqrt{2}$,
6. $\sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$

4.2. b) predykcja dla $K = 1$

Bierzemy najbliższego sąsiada którym jest obserwacja 5 i bierzemy jego odpowiedź → ***Green***

4.3. c) predykcja dla $K = 3$

Bierzemy 3 najbliższych sąsiadów którymi są obserwacje 5 → Green, 6 → Red oraz 2 → Red i bierzemy najpopularniejszą odpowiedź wśród tych obserwacji → ***Red***

4.4. d) lepsze małe czy duże wartości K dla silnie nieliniowej granicy decyzyjnej Bayesa

Jeśli granica decyzyjna Bayesa jest silnie nieliniowa (czyli prawdziwa granica między klasami jest bardzo złożona), to potrzebujemy bardziej elastycznego modelu, aby tę złożoność uchwycić. Małe wartości K dają klasyfikatorowi KNN większą elastyczność (niższy bias, wyższa wariancja) i pozwalają na odwzorowanie skomplikowanych, lokalnych zależności. Dlatego przy silnie nieliniowej granicy Bayesa zwykle lepiej sprawdzą się małe wartości K. (Przy przeciwnym założeniu — gładkiej granicy — większe K redukuje wariancję i może dać lepsze wyniki).

5. Zadanie 5 - Wartość oczekiwana i wariancja sumy zmiennych losowych

$$\mathrm{Cov}(X, Y) = 0 \tag{1}$$

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

Proof.

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_x \sum_y (x+y)\Pr(X=x, Y=y) \\
&= \sum_x \sum_y x \cdot \Pr(X=x, Y=y) + \sum_x \sum_y y \cdot \Pr(X=x, Y=y) \\
&= \sum_x x \cdot \left(\sum_y \cdot \Pr(X=x, Y=y) \right) + \sum_y y \cdot \left(\sum_x \cdot \Pr(X=x, Y=y) \right) \\
&= \sum_x x \cdot \Pr(X=x) + \sum_y y \cdot \Pr(Y=y) \\
&= \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned} \tag{2}$$

■

5.2. b) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proof.

$$\begin{aligned}
&\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big] \wedge (1) \\
&\Downarrow
\end{aligned} \tag{3}$$

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}\big[(X+Y-\mathbb{E}(X+Y))^2\big] \\
&= \mathbb{E}\big[((X-\mathbb{E}(X)) + (Y-\mathbb{E}(Y)))^2\big] \\
&= \mathbb{E}\big[(X-\mathbb{E}(X))^2 + (Y-\mathbb{E}(Y))^2 + 2\cdot(X-\mathbb{E}(X))\cdot(Y-\mathbb{E}(Y))\big] \\
&= \mathbb{E}\big[(X-\mathbb{E}(X))^2\big] + \mathbb{E}\big[(Y-\mathbb{E}(Y))^2\big] + \mathbb{E}[2\cdot(X-\mathbb{E}(X))\cdot(Y-\mathbb{E}(Y))] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + \underbrace{2\cdot\mathrm{Cov}(X,Y)}_{0} \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y)
\end{aligned} \tag{4}$$

■

6. Zadanie 6 - Przypomnienie estymatrów, jego bias'u i wariancji wrac z przykładami

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

6.1. Definicje

Estymator - To funkcja mająca za zadanie przybliżyć (wyestymować) wybrany parametr.

$$\theta \approx \hat{\theta}(X_1, X_2, \cdots, X_n) \tag{5}$$

Bias estymatora - Jest wartość opisująca tendencje estymatora do zawyżania lub zaniżania wartości estymowanej

$$\text{Bias}\big(\hat{\theta}\big) = \mathbb{E}\big(\hat{\theta}\big) - \theta \tag{6}$$

- Estymator nazywamy obciążonym gdy $\text{Bias}\big(\hat{\theta}\big) \neq 0$
- Estymator nazywamy nieobciążonym gdy $\text{Bias}\big(\hat{\theta}\big) = 0 \Rightarrow \mathbb{E}\big(\hat{\theta}\big) = \theta$

Wariancja estymatora - służy ocenianiu jak daleko, zazwyczaj, będzie estymator od estymowanej wartości

$$\text{Var}\big(\hat{\theta}\big) = \mathbb{E}\Big(\big(\mathbb{E}\big(\hat{\theta}\big) - \hat{\theta}\big)^2\Big) \tag{7}$$

MSE - Mean Square Error (ang. Błąd średniokwadratowy) tak samo jak wariancja

$$\text{MSE}\big(\hat{\theta}\big) = \mathbb{E}\Big(\big(\hat{\theta}(X) - \theta\big)^2\Big) \tag{8}$$

6.2. Przykłady

6.2.1. Nieobciążony

Typ estymatora	Przykład	Obciążenie
Średnia z próby	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mathbb{E}(\hat{\mu}) = \mu \Rightarrow \text{Bias}(\hat{\mu}) = 0$
Wariancja z $n-1$	$\hat{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$	$\mathbb{E}(\hat{\sigma}^2_{n-1}) = \sigma^2 \Rightarrow \text{Bias}(\hat{\sigma}^2_{n-1}) = 0$

6.2.2. Obciążony

Typ estymatora	Przykład	Obciążenie
Wariancja z n	$\hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$	$\mathbb{E}(\hat{\sigma}^2_n) = \frac{n-1}{n}\sigma^2 \Rightarrow \text{Bias}(\hat{\sigma}^2_n) = -\frac{\sigma^2}{n}$
$X \sim \text{Exp}(\lambda)$	MLE $\hat{\lambda} = \frac{1}{\bar{X}}$	$\mathbb{E}\big(\hat{\lambda}\big) = \frac{n}{n-1} \Rightarrow \text{Bias}\big(\hat{\lambda}\big) = \frac{1}{n-1}\lambda$

7. Zadanie 7 - dopasowywanie elastyczności modelu do scenariusza

Szukamy odpowiedniego modelu dla problemu związanego z uczeniem nadzorowanym. 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. $\sigma^2 = \mathbb{V}\text{ar}(\varepsilon)$, 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?

Przypadek	Elastyczność modelu	Uzasadnienie
a) n - duże, p - małe	Bardziej elastyczny	Elastyczny model może wykorzystać dużą liczbę obserwacji aby dopasować się i dzięki temu obniża bias
b) n - małe p - duże	Mniej elastyczny	Duża liczba predyktorów grozi overfittingiem, należy najpierw dokonać redukcji wymiaru tak aby p było mniejsze od n a następnie wykorzystać mniej elastyczny model aby uzyskać model zwiększyć wydajność na małej ilości danych
c) silna nieliniowość danych	Bardziej elastyczny	Elastyczny model lepiej dopasuje się do danych i lepiej przedstawi nieliniowość, nie uśredniając jej tak jak zrobiłby to model mniej elastyczny
d) Wysoka wariancja szumu	Mniej elastyczny	Mniej elastyczny model uśredni dane i zniweluje w ten sposób wysoką wariancję szumu

7.1. Wnioski

	Model mniej elastyczny	Model bardziej elastyczny
Zalety	• stabilny, z mała wariancją • łatwiejszy do interpretacji • bardziej odporny na szum • sprawdza się lepiej dla małych n względem p	• niski bias • świetny dla dużych n • uchwyci złożone zależności
Wady	• duży bias • słaby przy dużej nieliniowości danych	• duża wariancja • niezrozumiały sposób działania (black box) • duży koszt obliczeń • kiepski dla małych n • ryzyko overfittingu