

Klasyfikacja

Piotr Kotłowski

Koło Naukowe Data Science, Wydział MiNI PW

3 styczeń 2025

Porównanie: Klasyfikacja vs. Regresja

Cel: Przewidywanie **kategorycznej** (dyskretnej) etykiety klasy.

- **Dane wyjściowe:** Skończony zbiór klas (np. 0 lub 1; A, B lub C).
- Naszym celem jest stworzenie odpowiedniej funkcji $y = f(x)$, która przypisuje wejście do kategorii.
- **Przykłady:**
 - Rozpoznawanie cyfr (0-9)
 - Diagnoza medyczna (Chory/Zdrowy)
- **Metryki oceny:** Zależnie od problemu (np. Accuracy, F1).

- **Definicja:** Problem, w którym zmienna celu y przyjmuje jedną z dwóch wartości, zazwyczaj oznaczanych jako:
 - 0: Klasa negatywna (np. "brak spamu")
 - 1: Klasa pozytywna (np. "spam")

Funkcja predykcyjna jako prawdopodobieństwo

W wielu modelach (np. regresji logistycznej) nie przewidujemy etykiety wprost. Zamiast tego naszym celem jest stworzenie funkcji $f(x)$, która zwraca **prawdopodobieństwo** przynależności do klasy pozytywnej:

$$\hat{y} = f(x) = P(y = 1 \mid x)$$

Wartość tej funkcji mieści się w przedziale $[0, 1]$.

* Gdzie x oznacza wektor cech (zmienne wejściowe/atruty) opisujący daną obserwację.

Klasyfikacja binarna: Reguła decyzyjna

- **Reguła decyzyjna:** Aby przypisać konkretną klasę, stosujemy próg odcięcia (threshold), najczęściej 0.5:

$$\text{Decyzja} = \begin{cases} 1 & \text{gdy } f(x) > 0.5 \\ 0 & \text{gdy } f(x) \leq 0.5 \end{cases}$$

- **Elastyczność progu:** Wartość 0.5 nie jest sztywna. Możemy (i powinniśmy) ją przesuwać w zależności od celów biznesowych i kosztów błędu.
 - **Niski próg (np. 0.1):** Gdy chcemy wykryć jak najwięcej przypadków pozytywnych (np. wykrywanie raka), godząc się na fałszywe alarmy.
 - **Wysoki próg (np. 0.9):** Gdy zależy nam na pewności i chcemy uniknąć fałszywych alarmów (np. systemy rekomendacji, filtry spamowe).

Funkcja Sigmoidalna (Logistyczna)

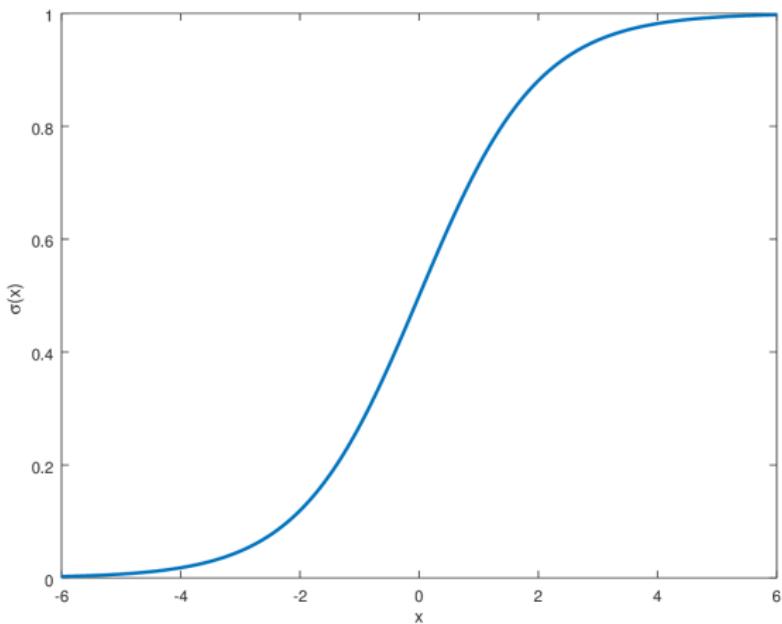
Wzór matematyczny:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Kluczowe właściwości:

- **Zbiór wartości:** Przedział $(0, 1)$. Dzięki temu idealnie nadaje się do mapowania wyników na **prawdopodobieństwo**.
- **Kształt:** Krzywa w kształcie litery "S".
- **Punkt środkowy:** Dla $z = 0$ wartość wynosi 0.5 (punkt maksymalnej niepewności).
- **Asympoty:** Dąży do 1 dla dużych wartości dodatnich i do 0 dla dużych wartości ujemnych.

Funkcja Sigmoidalna (Logistyczna)



Model Regresji Logistycznej

Funkcja predykcyjna

$$f(x) = P(y = 1 \mid x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Elementy równania:

- x : Wektor cech wejściowych
- w : Wektor wag – parametry, których uczy się model.
- b : Wyraz wolny – przesunięcie funkcji.
- $\sigma(\cdot)$: Funkcja sigmoidalna mapująca wynik do przedziału $(0, 1)$.

Od Prawdopodobieństwa do Log-Odds

Jeśli przekształcimy równanie sigmoidy, otrzymamy liniową zależność względem wag.

1. Szansa (Odds): Stosunek prawdopodobieństwa sukcesu (p) do porażki ($1 - p$):

$$\text{Odds} = \frac{p}{1 - p} = e^{w^T x + b}$$

2. Logarytm szans (Log-Odds / Logit): Logarytmując stronami, otrzymujemy:

Funkcja Logit

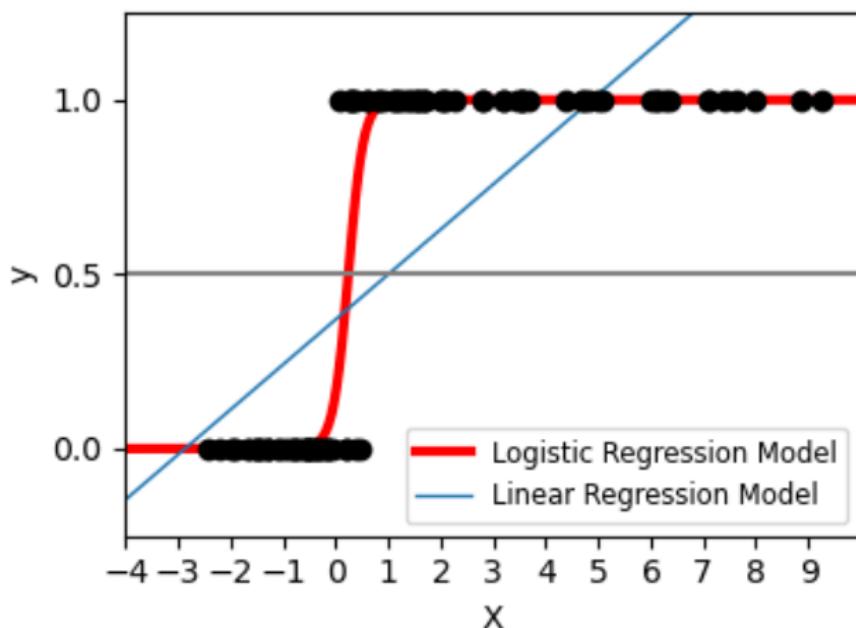
$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{Logit (Log-Odds)}} = \underbrace{w^T x + b}_{\text{Model Liniowy}}$$

Estymacja parametrów: Dlaczego nie MNK?

1. Czy istnieje rozwiążanie jawne (closed-form)?

- **NIE.** W przeciwieństwie do regresji liniowej (gdzie mamy wzór $(X^T X)^{-1} X^T y$), w regresji logistycznej nie da się wyznaczyć wzoru analitycznie.
- **Rozwiążanie:** Musimy stosować metody iteracyjne, takie jak **Spadek Gradientu** (Gradient Descent) lub metoda Newtona.

Regresja logistyczna vs Regresja liniowa



Drzewa Decyzyjne (Decision Trees)

Intuicja: Model ten naśladuje ludzki sposób podejmowania decyzji poprzez serię pytań logicznych ("If-Then-Else").

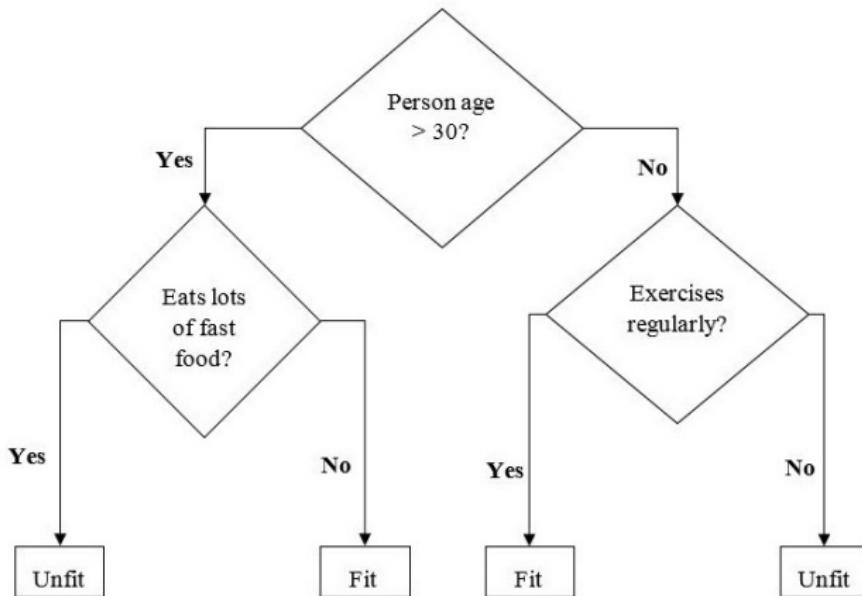
Budowa:

- **Korzeń (Root Node):** Punkt startowy zawierający całą populację.
- **Węzeł decyzyjny (Decision Node):** Punkt podziału (split) na podstawie warunku (np. $x_1 > 5$).
- **Liść (Leaf Node):** Końcowy węzeł zawierający decyzję (klasę lub wartość liczbową).

Kluczowe zalety:

- **Interpretowalność:** Model typu "White Box" – łatwy do wyjaśnienia.
- **Nieliniowość:** Model nie zakłada liniowej zależności danych. Potrafi tworzyć skomplikowane granice decyzyjne, dzieląc przestrzeń na prostokątne obszary.

Drzew Decyzyjne



Algorytm podziału węzła (Best Split Strategy)

- ① **Iteracja po cechach:** Algorytm rozważa każdą zmienną (kolumnę) X_j dostępną w zbiorze danych.
- ② **Sortowanie:** Dla danej cechy wartości unikalne są sortowane rosnąco: $v_1 < v_2 < \dots < v_n$.
- ③ **Skanowanie progów (Thresholds):** Algorytm sprawdza potencjalne miejsca podziału t (zazwyczaj średnie sąsiednich wartości). Dla każdego t :
 - **Podział:** Przypisz rekordy z $x \leq t$ do lewego węzła, a $x > t$ do prawego.
 - **Ocena:** Oblicz wartość funkcji kryterialnej dla tego konkretnego podziału.
- ④ **Wybór optimum (Greedy):** Spośród wszystkich sprawdzonych cech i progów, wybierany jest ten, który daje **maksymalny zysk informacyjny**.
- ⑤ **Rekurencja:** Proces jest powtarzany niezależnie dla nowo powstałego lewego i prawego dziecka, aż do spełnienia kryterium stopu (np. max głębokość).

Metryki oceny klasyfikacji

Wszystkie metryki opierają się na czterech podstawowych wynikach zliczanych w Macierzy Pomyłek (Confusion Matrix).

		Positive Negative	
		True ✓ Positive (TP)	False ✗ Positive (FP)
Predicted Label	Positive	True ✓ Positive (TP)	False ✗ Positive (FP)
	Negative	False ✗ Negative (FN)	True ✓ Negative (TN)

Precision = $\frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$

Recall = $\frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$

Accuracy = $\frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$

Interpretacja metryk: Precyza, Czułość i F1

Precyza (Precision)

$$\frac{TP}{TP + FP}$$

Pytanie: Jak bardzo mogę ufać, gdy model mówi "To jest klasa 1"?

Czułość (Recall)

$$\frac{TP}{TP + FN}$$

Pytanie: Czy model znalazł wszystkie interesujące nas przypadki?

F1-Score: Złoty środek

Średnia harmoniczna Precyzji i Czułości. Używamy jej, gdy zależy nam na równowadze lub gdy dane są niebalansowane.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Metryki oceny klasyfikacji

Dlaczego potrzebujemy różnych metryk? Accuracy bywa mylące! Jeśli w zbiorze mamy 99 pacjentów zdrowych i 1 chorego, model przewidujący zawsze "Zdrowy" ma Accuracy 99%, ale Recall 0% (jest bezużyteczny medycznie).

Inne kluczowe miary

W uczeniu maszynowym (szczególnie w scoringu kredytowym i modelach probabilistycznych) stosuje się także:

- **ROC-AUC** oraz **Gini**: Oceniają zdolność modelu do rozróżniania klas **niezależnie od przyjętego progu**.
- **Log-Loss (Entropia Krzyżowa)**: Ocenia nie tylko trafność, ale i **pewność** modelu (karze za bycie pewnym przy błędnej decyzji).

Laboratoria...