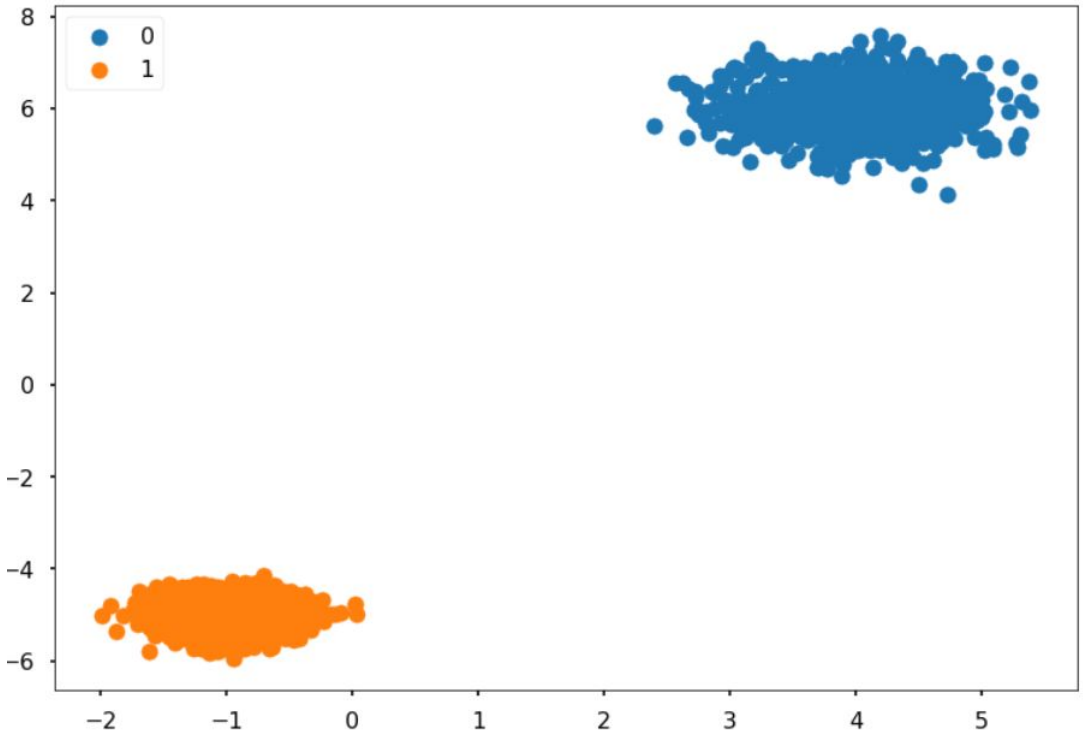
KNSI Golem Bootcamp 2021

Spotkanie 3 – klasyfikacja

Agenda

- Krótkie przypomnienie
- Regresja logistyczna (nie mylić z liniową!)
- Klasyfikacja binarna vs wieloklasowa
- Potężny notebook



Krótko o klasyfikacji

Kluczowe informacje

- Cel: zbudować model predykcyjny pozwalający przypisać obiekt do danej klasy
- Podstawowy model: regresja logistyczna, drzewo decyzyjne
- Przykład: mając dane o pasażerze Titanica przewidzieć czy uda mu się przeżyć katastrofę
- Główne typy klasyfikacji:
    - Binarna (rozumiana głównie jako prawda-fałsz)
    - Wieloklasowa (klasy są ponumerowane)
- Funkcja straty: Entropia krzyżowa

Regresja logistyczna

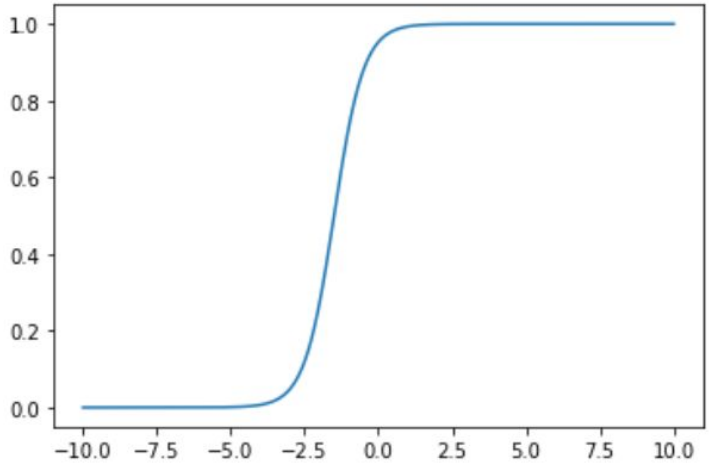
Jak działa?

Wzór

$$P = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

Przykład

$$\beta_0 = 3, \beta_1 = 2, X \in [-10, 10]$$



Sigmoida

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

Funkcja straty: entropia krzyżowa

$$H(p,q) = -\sum_{x \in X} p(x) * log(q(x))$$

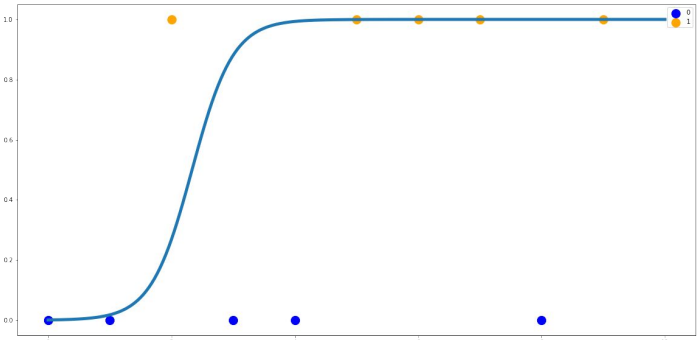
W naszym rozumieniu:
p(x) – prawdziwa klasa dla danego przypadku
q(x) – prawdopodobieństwo zwracane przez model

$$L(\beta, X, y) = -\sum_{x \in X} y * log(\sigma(\beta, x))$$

Przykład

Dane i model

$$X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$
$$y = \{0, 0, 1, 0, 0, 1, 1, 1, 0, 1\}$$
$$\beta_0 = -7, \beta_1 = 3$$



Wartość funkcji straty

$$L(\beta, X, y) = -\sum_{x \in X} y * log(\sigma(\beta, x))$$
$$-\sum_{x \in X} (1 - y) * log(1 - \sigma(\beta, x)) \approx 25.4$$

Wieloklasowe zadanie klasyfikacji

Główne różnice:

- Inna reprezentacja każdej z klasy (One Hot Encoding), przykład:

	Y
0	1
1	2
2	3
3	2
4	4

staje się

y_1	y_2	y_3	y_4
1	0	0	0
0	1	0	0
0	0	1	0
0	1	0	0
0	0	0	1

- Trochę inna, ale bardzo zbliżona forma funkcji straty opartej na entropii krzyżowej
- Modele zwracają wektor prawdopodobieństw przynależności do danej klasy
- Niektóre modele trzeba “przerobić”, by mogły działać przy takim zadaniu

Inny model: drzewo decyzyjne

Q&A