

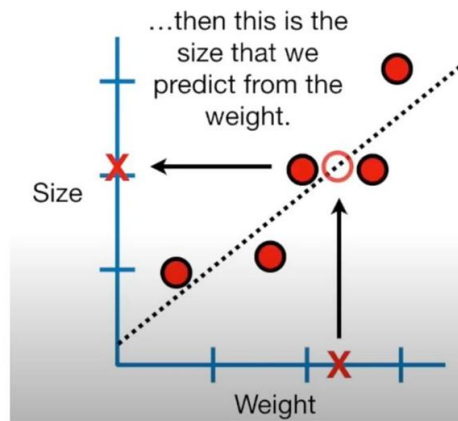
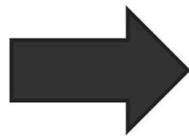
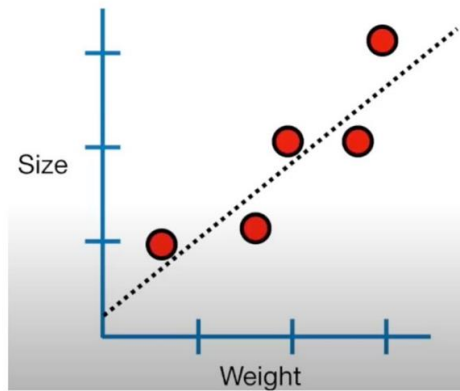
Regresja logistyczna

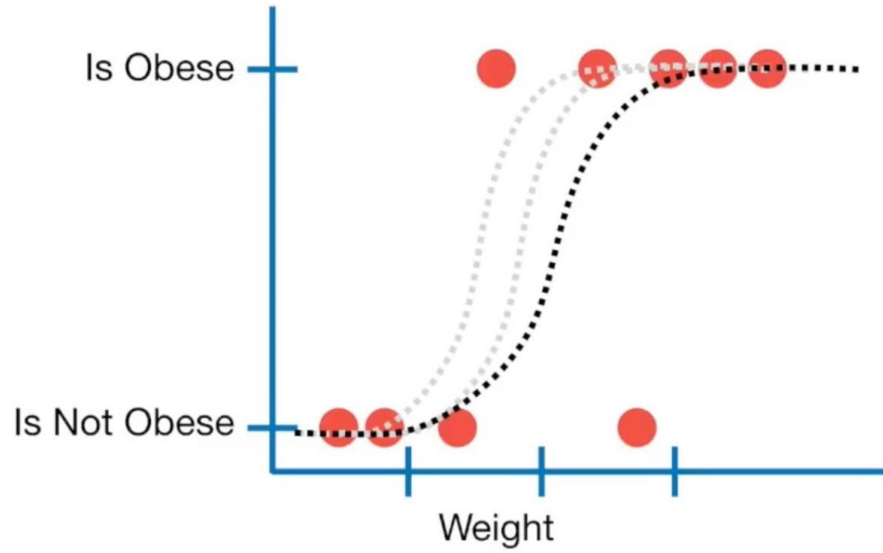


Roadmapa

- Czym jest regresja logistyczna?
- Czy i czym różni się względem regresji liniowej?
- Jakie są jej zastosowania?
- Implementacja i ocena wykonanych modeli

Regresja liniowa







Regresja logistyczna (regresja logitowa)

- Model używany w przypadku, gdy zmienna odpowiedzi Y jest kategoriowa
- Zakłada, że zmienna objaśniana ma rozkład dwupunktowy, czyli jest opisana na skali dychotomicznej (przyjmuje tylko dwie wartości)
- Używana w klasyfikacji, pozwala na oszacowanie prawdopodobieństwa przynależności do danej klasy
- Model regresji logistycznej jest uogólnionym modelem liniowym (GLM), w którym użyto logitu jako funkcji wiążącej



Uogólniony model liniowy

$$E(\mathbf{Y}|\mathbf{X}) = \mu = g^{-1}(\mathbf{X}\beta)$$

Rozkład normalny

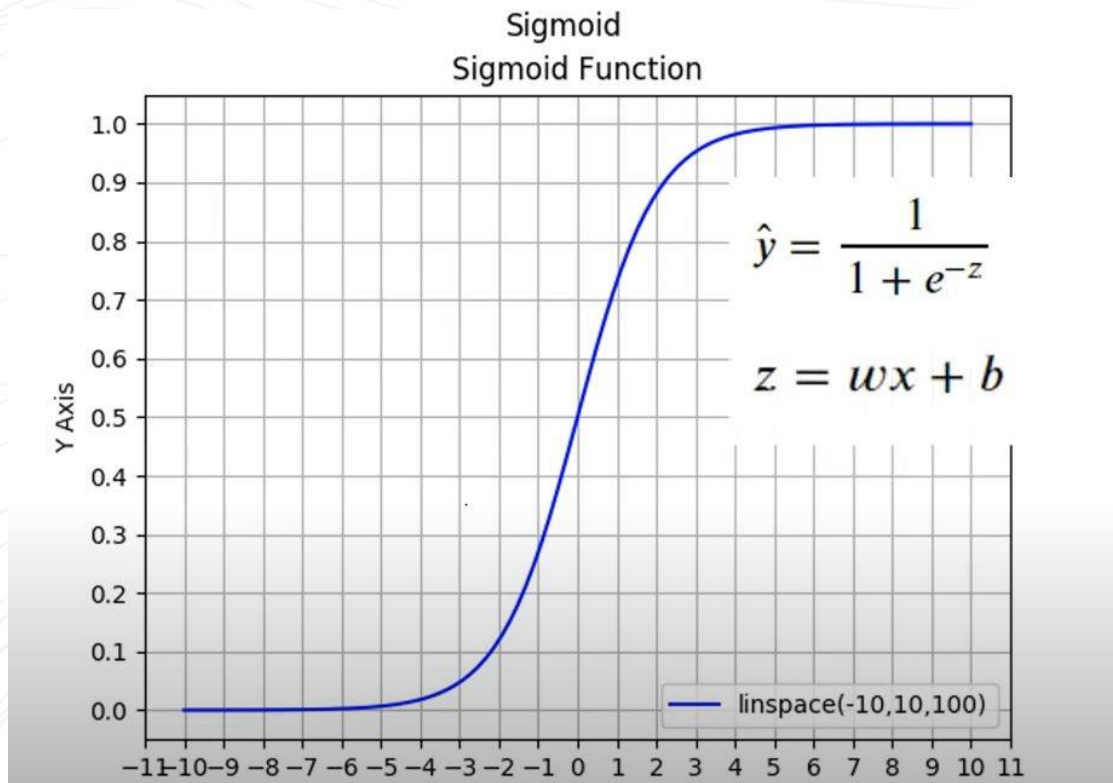
$$\mathbf{X}\beta = \mu$$

Rozkład Bernoulliego

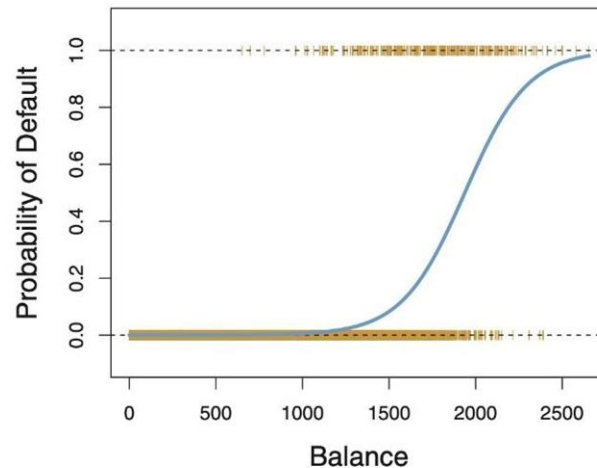
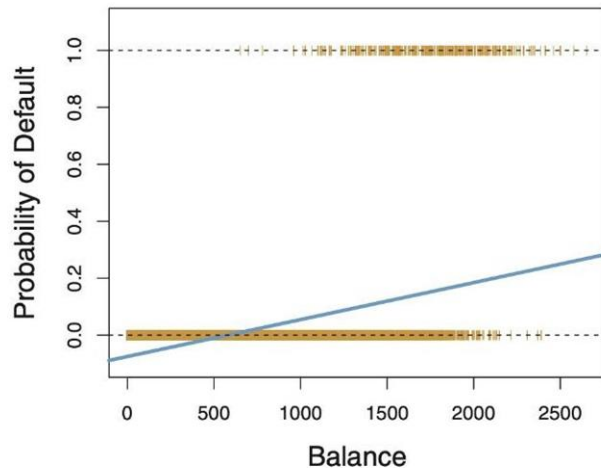
$$\mathbf{X}\beta = \ln\left(\frac{\mu}{1 - \mu}\right)$$

logit

Funkcja logistyczna



Regresja liniowa vs logistyczna



Prawdopodobieństwo klasy “default” (brak spłaty karty kredytowej) modelowane przy pomocy regresji liniowej (lewa strona) i regresji logistycznej (prawa strona)

Regresja liniowa vs logistyczna

	Linear Regression	Logistic Regression
1 Definition	To predict a continuous dependent variable based on values of independent variables	To predict a categorical dependent variable based on values of independent variables
2 Variable Type	Continuous dependent variable	Categorical dependent variable
3 Estimation method	Least square estimation	Maximum like-hood estimation
4 Equation	$Y = b_0 + b_1x + e$	$\log \left[\frac{Y}{1 - Y} \right] = C + B1X1 + B2X2 + \dots$
5 Best fit line	Straight line	Curve
6 Relationship between DV & IV	Linear relationship between the dependent and independent variable	Linear relationship is not mandatory
7 Output	Predicted integer value	Predicted binary value (0 or 1)
8 Applications	Business domain, forecasting sales	Classification problems, cybersecurity, image processing

<https://www.youtube.com/watch?v=OCwZyYH14uw>



infoShare
ACADEMY

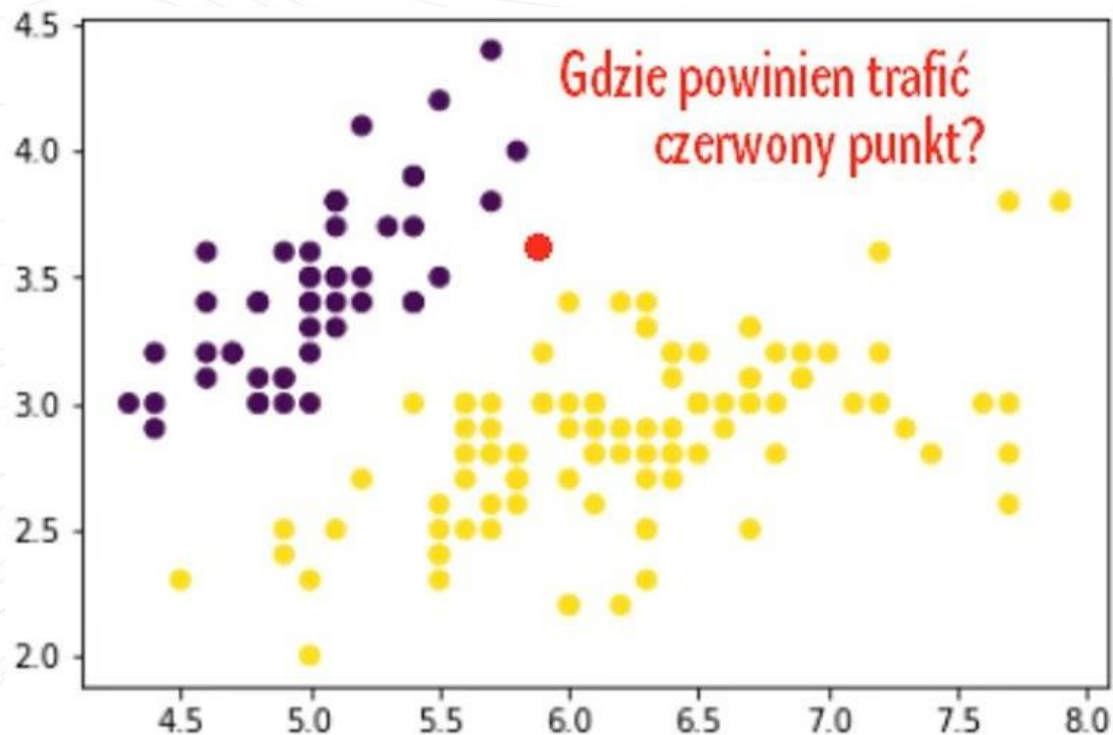


Zmienne w modelu

Skale pomiarowe

1. Skala nominalna (gatunki irysów)
2. Skala porządkowa ({zimno, ciepło, gorąco})
3. Skala interwałowa (daty)
4. Skala ilorazowa (wzrost)

$$Y = \begin{cases} 1 & \text{if epileptic seizure;} \\ 2 & \text{if stroke;} \\ 3 & \text{if drug overdose.} \end{cases}$$

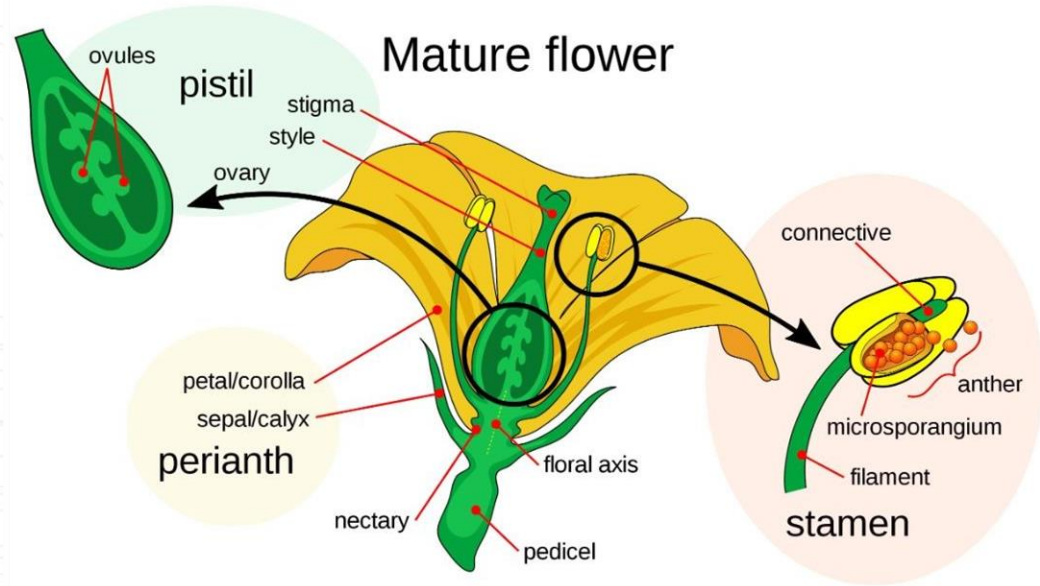


Klasyfikacja irysów

Klasy: trzy gatunki irysów

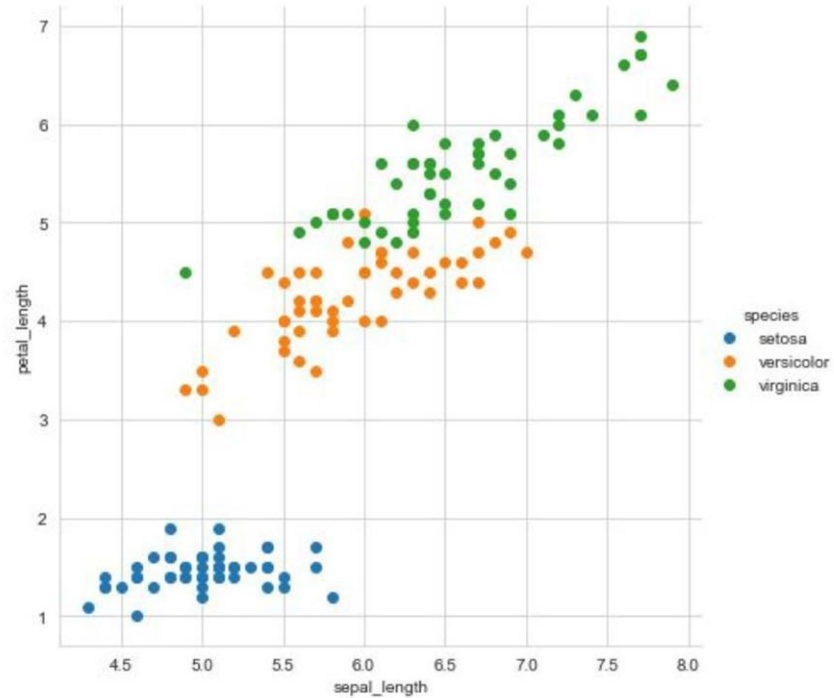
Cechy:

- długość działek kielicha (sepal)
- szerokość działek kielicha
- długość płatka (petal)
- szerokość płatka



<https://en.wikipedia.org/wiki/Sepal>

Iris dataset



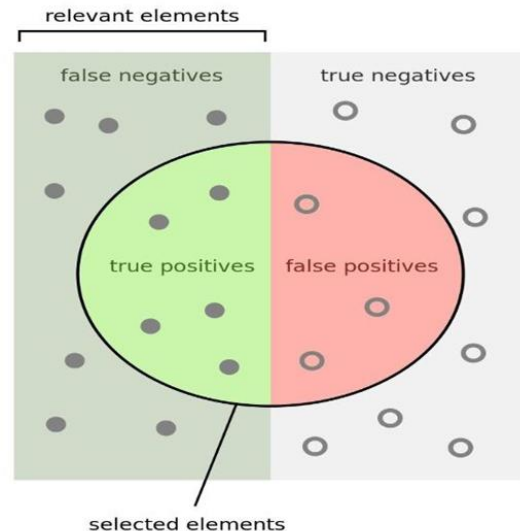


Metody walidacji klasyfikacji

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

		Predicted		
		0	1	
Actual	0	TN Type I error	FP Type I error	Specificity = $TN/(TN+FP)$
	1	FN Type II error	TP	Recall or Sensitivity = $TP/(TP+FN)$
		Negative Rate = $TN/(FN+TN)$	Precision = $TP/(TP+FP)$	

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

$$\text{F1 - Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$



Dodatkowe materiały:

Dokumentacja sklearn – LogisticRegression

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

One-Vs-Rest for Multi-Class Classification

<https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification>

Tutorial – regresja logistyczna

<https://realpython.com/logistic-regression-python>

Dzięki!