Systemy pomiarowo-diagnostyczne

Metody uczenia maszynowego – wykład II

bogumil.konopka@pwr.edu.pl

2015/2016

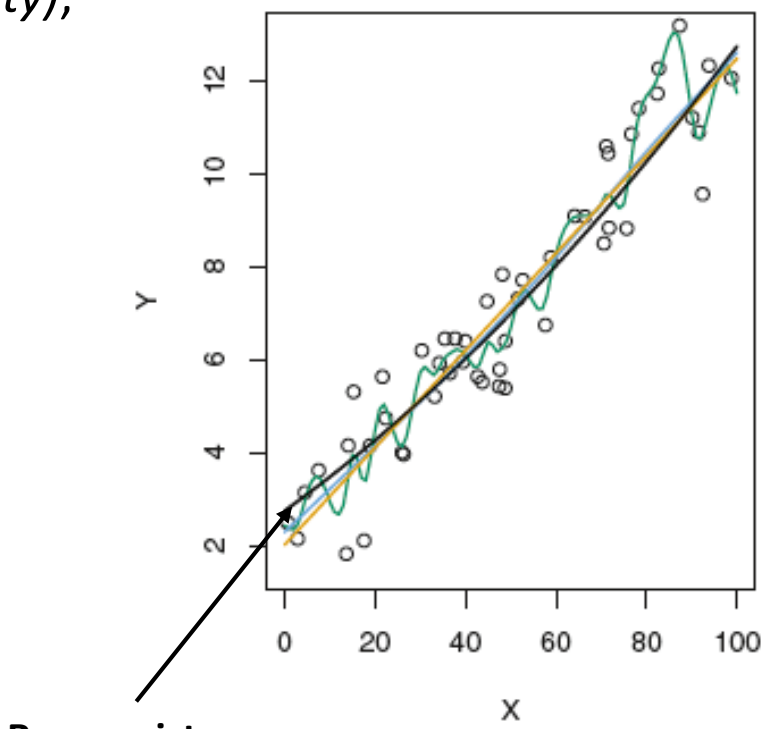
Określenie rzeczywistej dokładności modelu

Zbiór treningowym vs zbiór testowy

- Zbiór treningowy – zbiór wykorzystywany przy budowie modelu
- Zbiór testowy – zbiór niedostępny przy budowie modelu

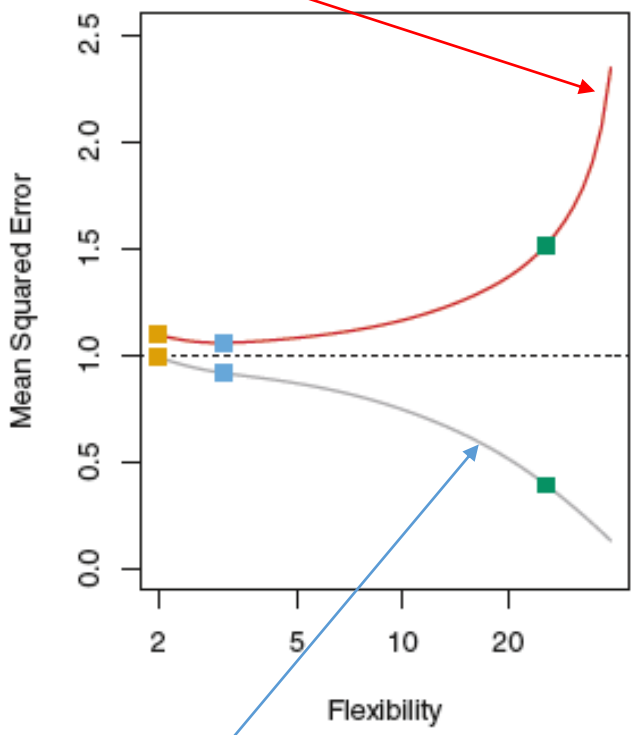
Elastyczność modelu przy danych liniowych

- Model liniowy (dwa parametry)
  - „mała” elastyczność (*flexibility)*,
  - mały błąd treningowy
  - mały błąd testowy
- Wielomian z kilkoma parametrami
  - „umiarkowana” elastyczność
  - mały błąd treningowy
  - mały błąd testowy
- Wielomian z dużą liczbą parametrów
  - duża elastyczność
  - mały błąd treningowy
  - duży błąd testowy



Rzeczywista zależność

MSE testowy

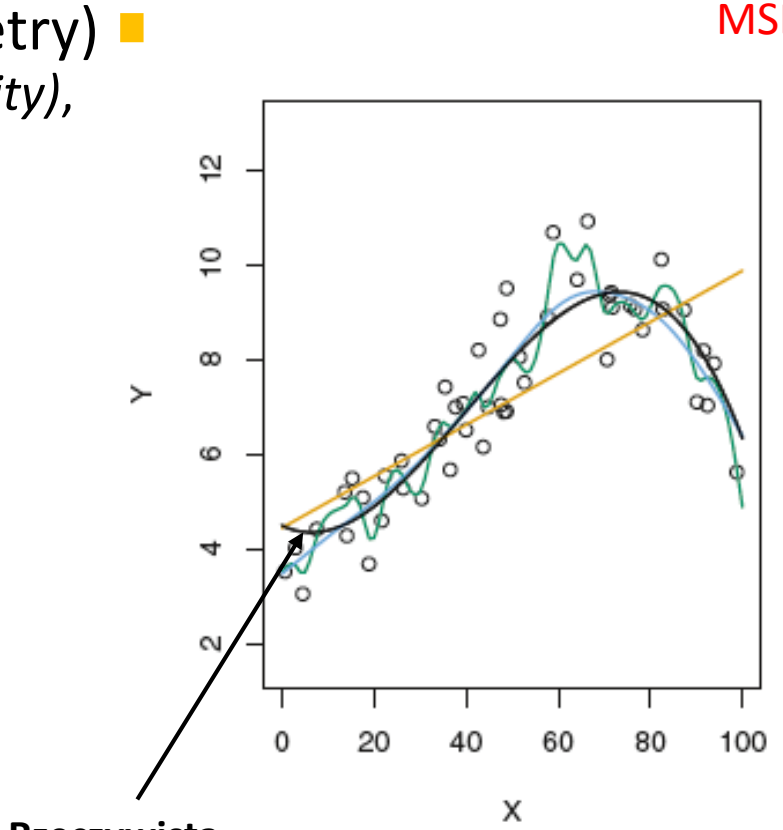


MSE treningowy

ITSL, James et al. 2013

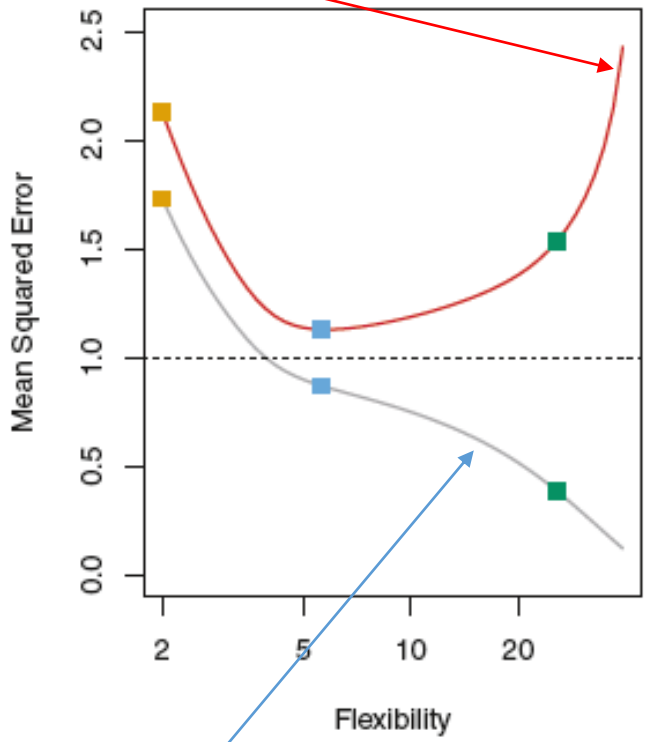
Elastyczność modelu, a MSE (przykład 2)

- Model liniowy (dwa parametry)
  - „mała” elastyczność (*flexibility)*,
  - duży błąd treningowy
  - duży błąd testowy
- Wielomian z kilkoma parametrami
  - „umiarkowana” elastyczność
  - mały błąd treningowy
  - mały błąd testowy
- Wielomian z dużą liczbą parametrów
  - duża elastyczność
  - mały błąd treningowy
  - duży błąd testowy



Rzeczywista zależność

MSE testowy

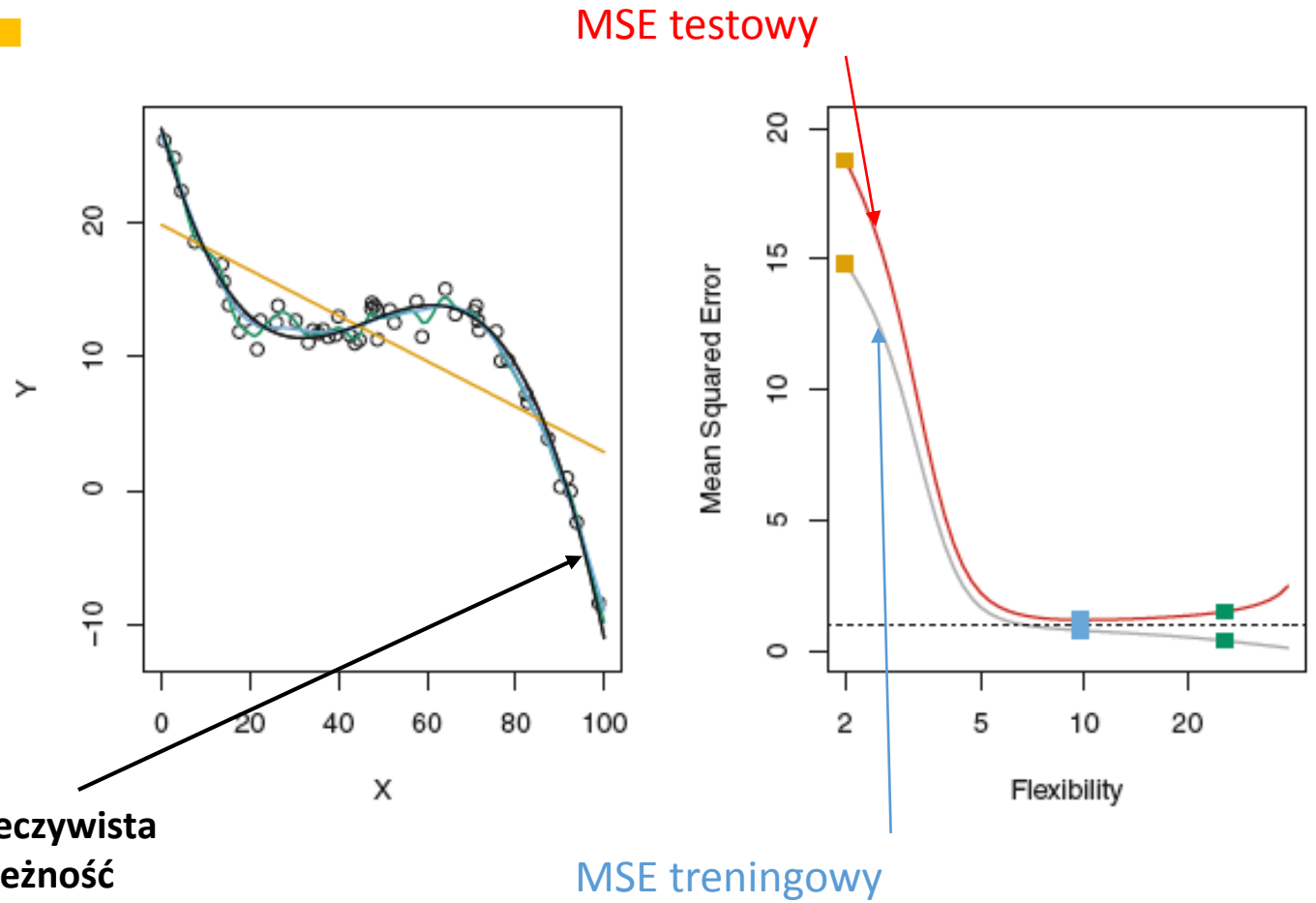


MSE treningowy

ITSL, James et al. 2013

Elastyczność modelu, a MSE (przykład 3)

- Model liniowy (dwa parametry)
  - „mała” elastyczność (*flexibility)*,
  - duży błąd treningowy
  - duży błąd testowy
- Wielomian z kilkoma parametrami
  - „umiarkowana” elastyczność
  - mały błąd treningowy
  - mały błąd testowy
- Wielomian z dużą liczbą parametrów
  - duża elastyczność
  - mały błąd treningowy
  - mały błąd testowy



ITSL, James et al. 2013

Obciążenie vs zmienność modelu

(ang. bias vs variance)

- Błąd może zawsze zostać rozłożony zgodnie z formułą:

$$E\left(y_0 - \hat{f}(x_0)\right) = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\varepsilon)$$

Zmienność modelu

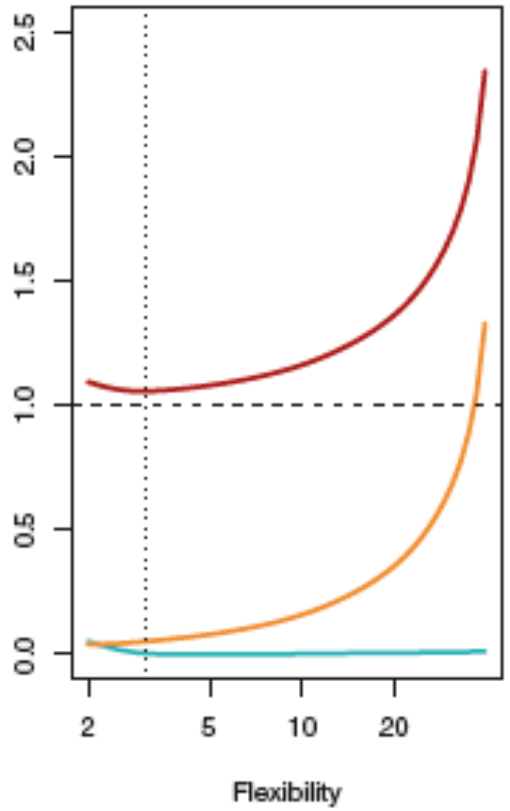
Obciążenie modelu

Błąd nieredukowalny

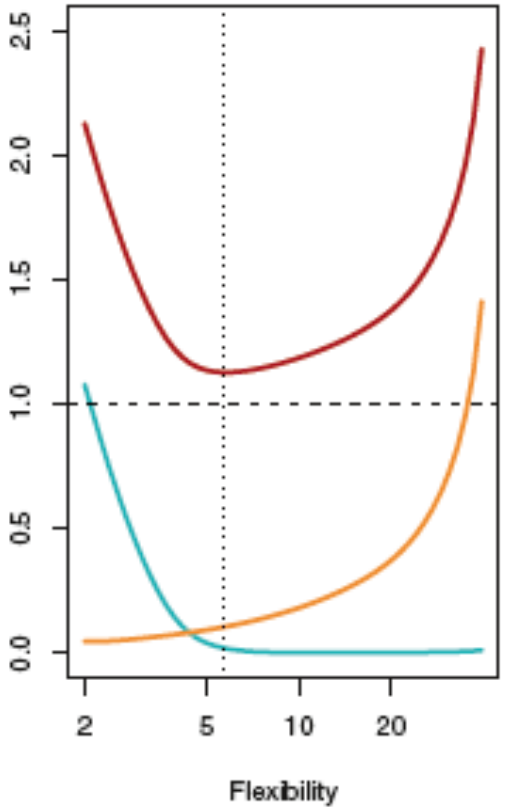
Obciążenie vs zmienność modelu

(ang. bias vs variance)

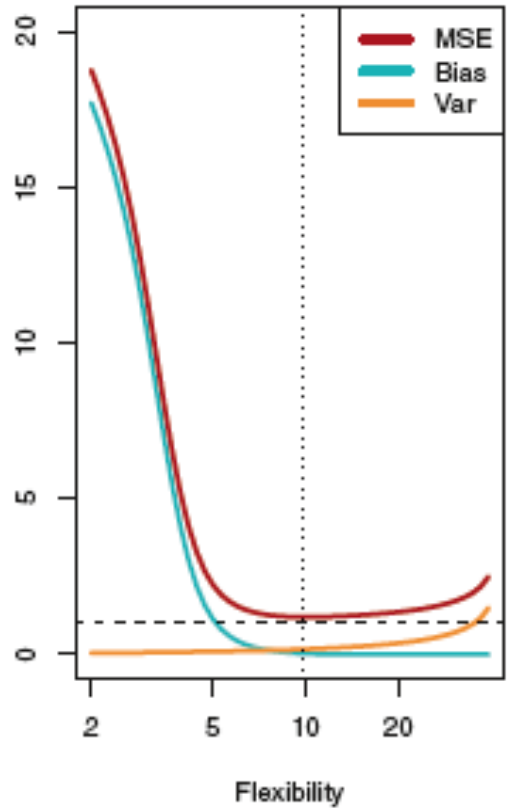
Dane liniowe



Dane nieliniowe



Dane mocno nieliniowe



$$E\left(y_0 - \hat{f}(x_0)\right) = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\varepsilon)$$

ITSL, James et al. 2013

Wykład II - plan

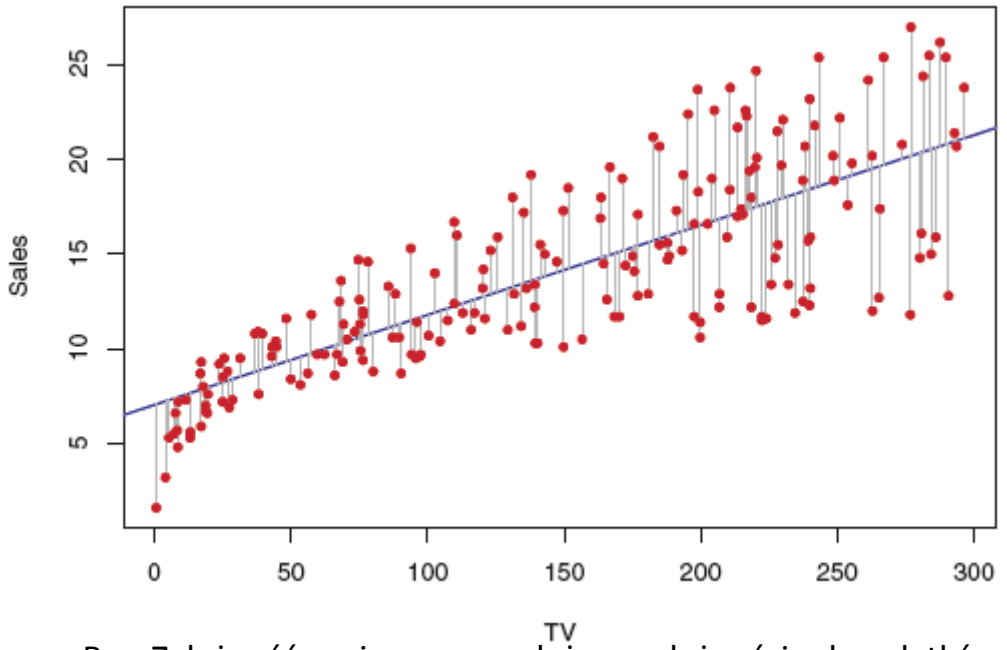
- Regresja liniowa
- Regresja logistyczna
- Ocena skuteczności klasyfikatorów:
  - Macierze pomyłek
  - Krzywe ROC

Regresja liniowa

- Metoda uczenia maszynowego z nadzorem
- Postać modelu:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Cele analizy:
  - Czy jest związek pomiędzy zmiennymi wyjaśniającymi, a zmienną wyjaśnianą?
  - Jak silny jest ten związek?
  - Jak silny jest wkład od poszczególnych zmiennych wyjaśniających?
  - Jak dokładnie możemy określić związek każdej zmiennej wyjaśniającej z odpowiedzią?
  - Jak dokładnie możemy przewidzieć zmienną wyjaśnianą
  - Czy związek jest liniowy?
  - Czy jest efekt synergii pomiędzy zmiennymi?



Rys. Zależność poziomu sprzedaży w zależności od wydatków na reklamy w telewizji

ITSL, James et al. 2013

Model prostej regresji liniowej - przykład

- Zbiór uczący i sformułowanie problemu

(tablica)

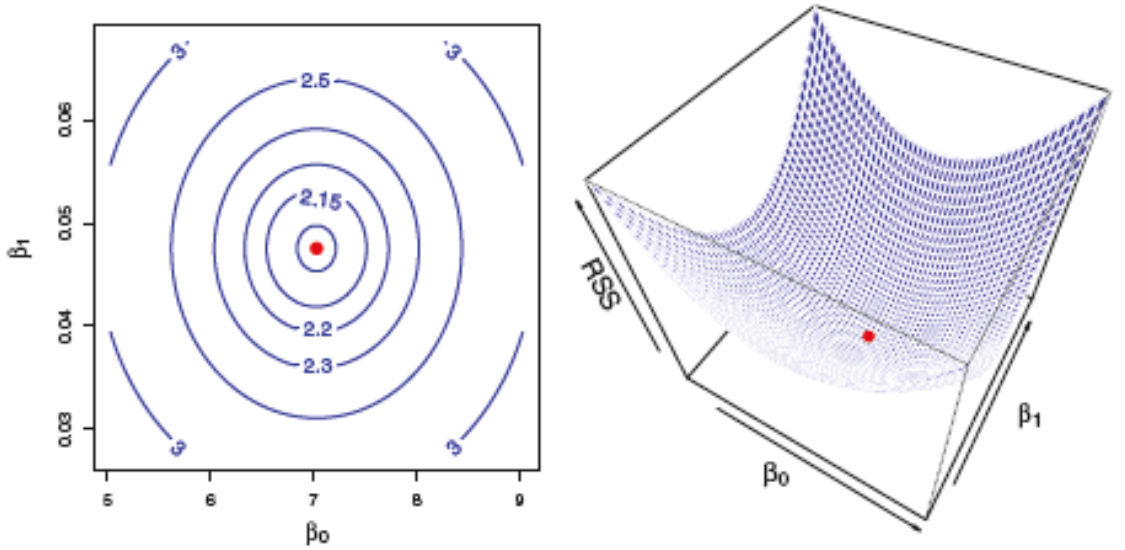
- Dopasowanie współczynników:
  - Minimalizacja RSS - Sumy kwadratów reszt („Residual Sum of Squares”)
  - Pochodne cząstkowe RSS po współczynnikach = 0

(tablica)

Minimalizacja sumy kwadratów reszt (RSS)

- Rezultat rozwiązania problemu minimalizacji RSS to:
  - Zestaw parametrów modelu, dający optymalne dopasowanie modelu do danych

- Istnieje wiele sposobów rozwiązania problemu minimalizacji RSS (np. algorytm najszybszego spadku – „gradient descent”)



ITSL, James et al. 2013

Ocena dokładności estymacji parametrów modelu

Przedziały ufności

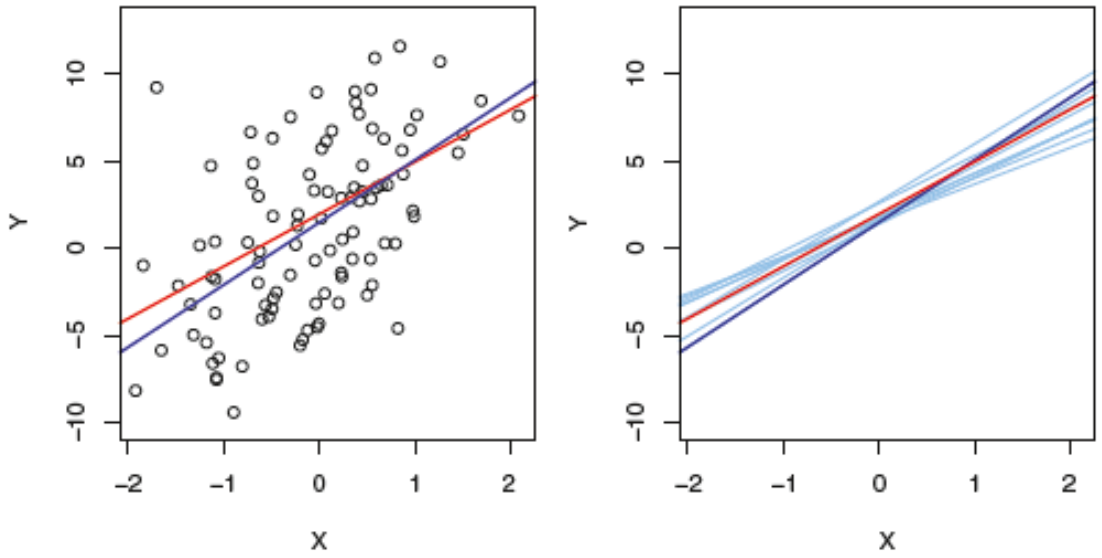
- Uzyskane parametry $\widehat{\beta_0}$ i $\widehat{\beta_1}$ – to tylko estymatory prawdziwych wartości
  - W zależności od punktów wykorzystanych w zbiorze uczącym, wyniki będą się różnić
- Przedziały ufności:

$$\left[\hat{\beta} - t^* SE(\hat{\beta}), \hat{\beta} + t^* SE(\hat{\beta})\right],$$

$$gdzie$$

$$t^* - kwantyl\ z\ rozkładu\ t - Studenta\ z\ (n-2)\ stopniami\ swobody$$

- $SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$
- $SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$



Ocena dokładności estymacji parametrów modelu

Test t-Studenta

- Test t-Studenta z df = n-2
- Celem testu jest sprawdzenie czy wartość współczynnika jest równa 0

$$h_0: \beta = 0$$
$$h_a: \beta \neq 0$$

- Statystyka testowa:

$$t = \frac{\beta - 0}{SE(\beta)}$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Ocena dokładności modelu

- Błąd standardowy reszt – RSE – Residual Standard Error

(tablica)

- Statystyka R^2

(tablica)

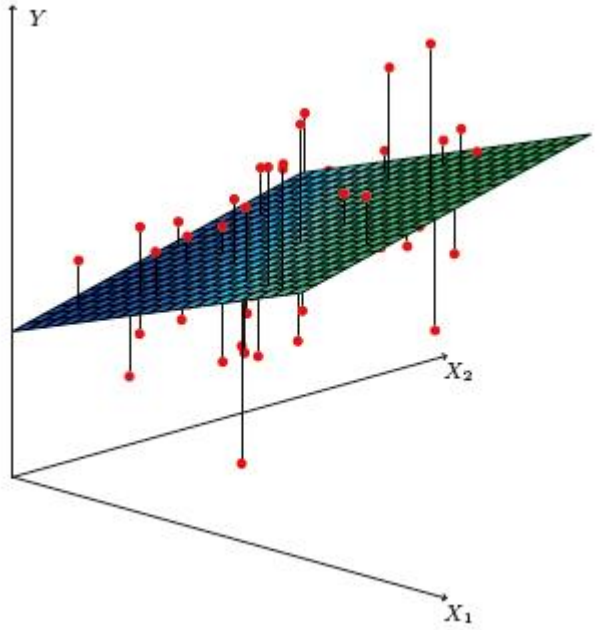
- F-statystka

(tablica)

Wielokrotna regresja liniowa

- Zbiór uczący i sformułowanie problemu:

(tablica)



- Dopasowanie modelu:

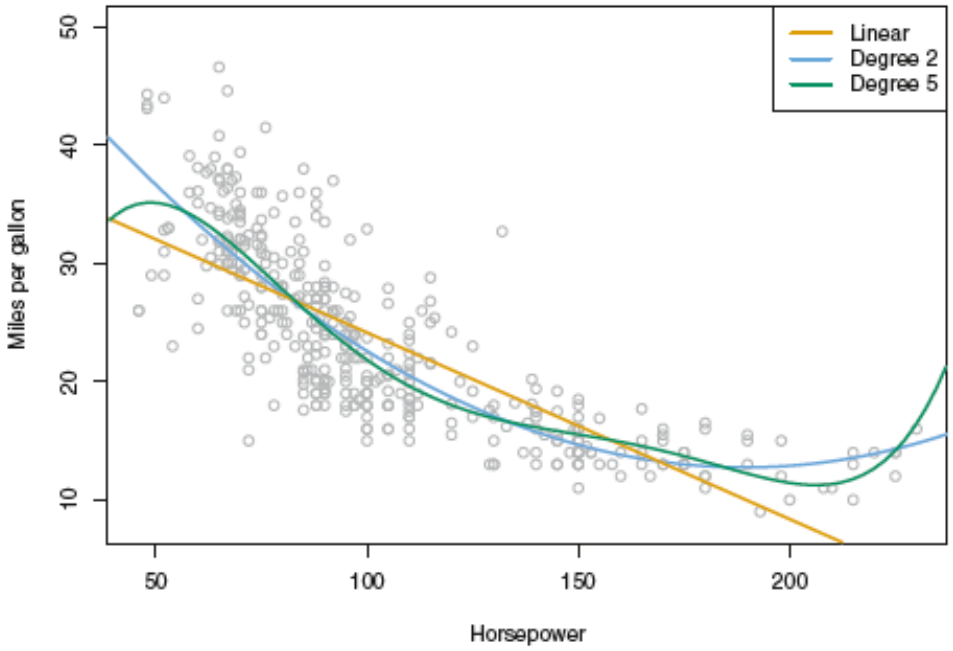
pełne wyprowadzenie „Guide to Inteligent Data Analysis” p. 234

Regresja wielomianowa

- Model liniowy może być łatwo rozszerzony do zależności wielomianowych

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2$$

Wystarczy utworzyć dodatkową zmienną wyjaśniającą



Więcej o regresji liniowej

- Algorytmy doboru zmiennych:
  - *Forward selection*
  - *Backward selection*
  - *Mixed selection*
- Regresja liniowa odporna na wartości odstające (*Robust linear regression*)
- Pojęcie *regularyzacji (regularization)*:
  - Metoda *LASSO*
  - Metoda *Ridge regression*

Problem klasyfikacji

- Celem jest przyporządkowanie obiektu do klasy w oparciu o zmienne wejściowe
  - Zmienna wyjaśniana jest zmnienną kategoryczną,
    np. „nowotwór złośliwy” / „nowotwór niezłośliwy”
- Sformułowanie formalne dla przypadku z dwiema klasami:

(tablica)

Klasyfikacja, a model liniowy

- Problemy z modelem liniowym:
  - W modelu liniowym $\hat{f}(x)$ może być >1 lub < od 0
  - Interpretacja jest trudna
  - Problemem jest również określenie parametru odcięcia
- Przy klasyfikacji chcemy żeby:

$$0 \leq \hat{f}(x) \leq 1$$

Regresja logistyczna

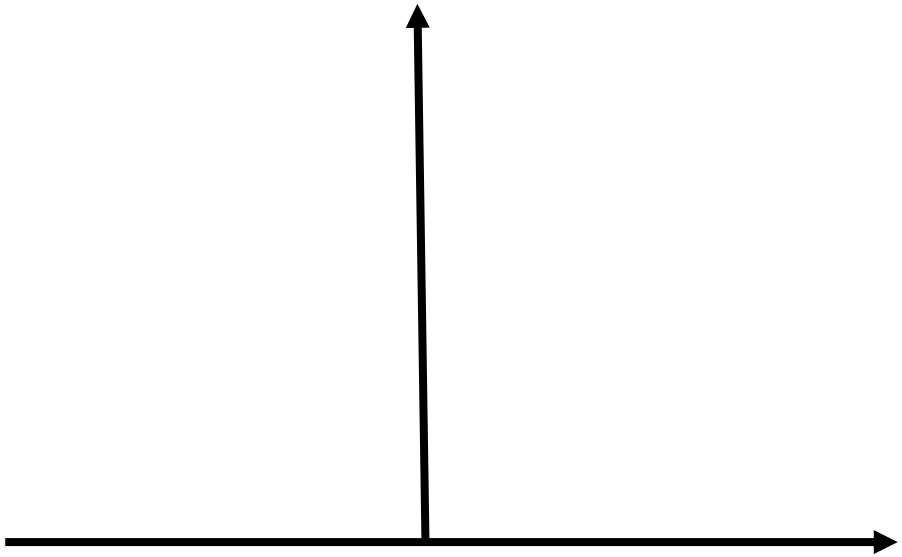
- Przy klasyfikacji chcemy żeby:

$$0 \leq \hat{f}(x) \leq 1$$

- Funkcja logistyczna:

$$\hat{f}(x) = \frac{1}{1 + e^{-z}}$$

$$\hat{f}(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$



Interpretacja modelu logistycznego

- $\hat{f}(x) = P(y = 1|X)$

- Przykład:

Jeżeli, dla modelu klasyfikującego nowotwory

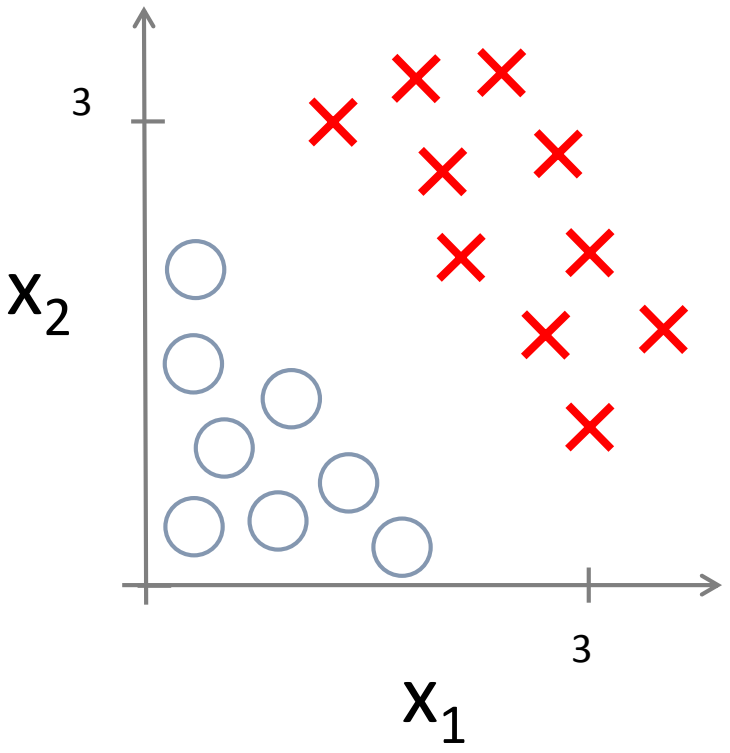
$$\hat{f}(x) = 0.7$$

To możemy powiedzieć:

„Jest 70% szans na to, że nowotwór jest złośliwy”

Graficzna interpretacja ‚z'

Dane w przestrzeni atrybutów

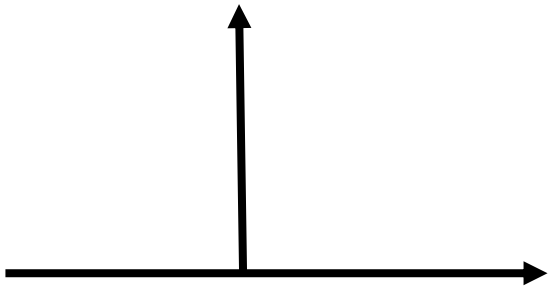


$$\hat{f}(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

Model dopasowany do danych:

$$\hat{f}(x) = \frac{1}{1 + e^{-(-3 + x_1 + x_2)}}$$

Kiedy y = 1?



Granica decyzji:

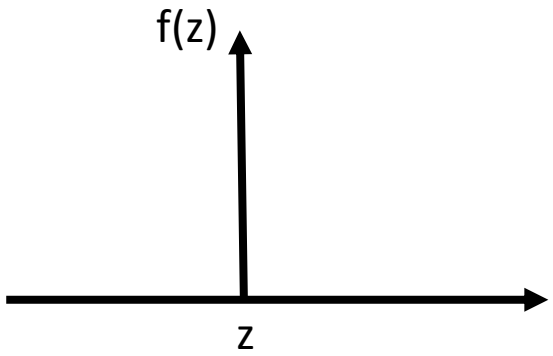
Dopasowanie współczynników

- Najczęściej współczynniki modelu dopasowuje się *metodą maksymalnej wiarygodności* (maximum likelihood)
- Intuicja algorytmu:
  - Wyszukiwane są takie wartości współczynników, dla których prawdopodobieństwo uzyskania prawidłowego przyporządkowania do klas jest największe
  - Maksymalizowana funkcja ma postać:

$$l(\beta) = \prod_{i:y_i=1} p(x_i;\beta) \prod_{j:y_j=0} (1 - p(x_j;\beta))$$

Ocena dokładności współczynników

- Współczynniki ocenia tak jak dla regresji liniowej:
  - Przedziały ufności
  - Test t-Studenta



	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance, income, *and student status. Student status is encoded as a dummy variable* student[Yes], *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* income *was measured in thousands of dollars.*

$$\hat{f}(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

Zadanie domowe na punkt z aktywności

- Jaka jest interpretacja współczynników w modelu regresji logistycznej?
- Wskazówki:
  - β_0 :Wychodząc od $\hat{f}(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1)}}$ wyprowadzić β_0 lub e^{β_0}, pamiętając, że $\hat{f}(x) = p(y = 1|x)$
  - β_1: Wyprowadzić e^{β_1} wyliczając stosunek $\frac{e^{-(\beta_0+\beta_1 x_1)}}{e^{-(\beta_0+\beta_1 x'_1)}}$, gdy $x'_1 = x_1 + 1$

Ocena dokładności klasyfikatora (klasyfikator binarny)

- Możliwe wyniki klasyfikacji:
  - **TP** – True Positive – Prawidłowo zaklasyfikowany przypadek pozytywny
  - **FP** – False Positive – Fałszywy pozytywny – Przypadek negatywny zaklasyfikowany jako pozytywny
  - **TN** – True Negative – Prawidłowo zaklasyfikowany przypadek negatywny
  - **FN** – False Negative – Fałszywy negatywny – Przypadek pozytywny zaklasyfikowany jako negatywny

Macierz pomyłek

		Rzeczywista klasa		
		Pozytywna	Negatywna	
Przewidywana klasa	Pozytywna	TP	FP	$\frac{TP}{P_{predicted}}$
	Negatywna	FN	TN	$\frac{TN}{N_{predicted}}$
		$\frac{TP}{P_{total}}$	$\frac{TN}{N_{total}}$	$\frac{TP+TN}{P_{total}+N_{total}}$

Positive Predictive Value/Precision – celność przewidywania przypadków pozytywnych

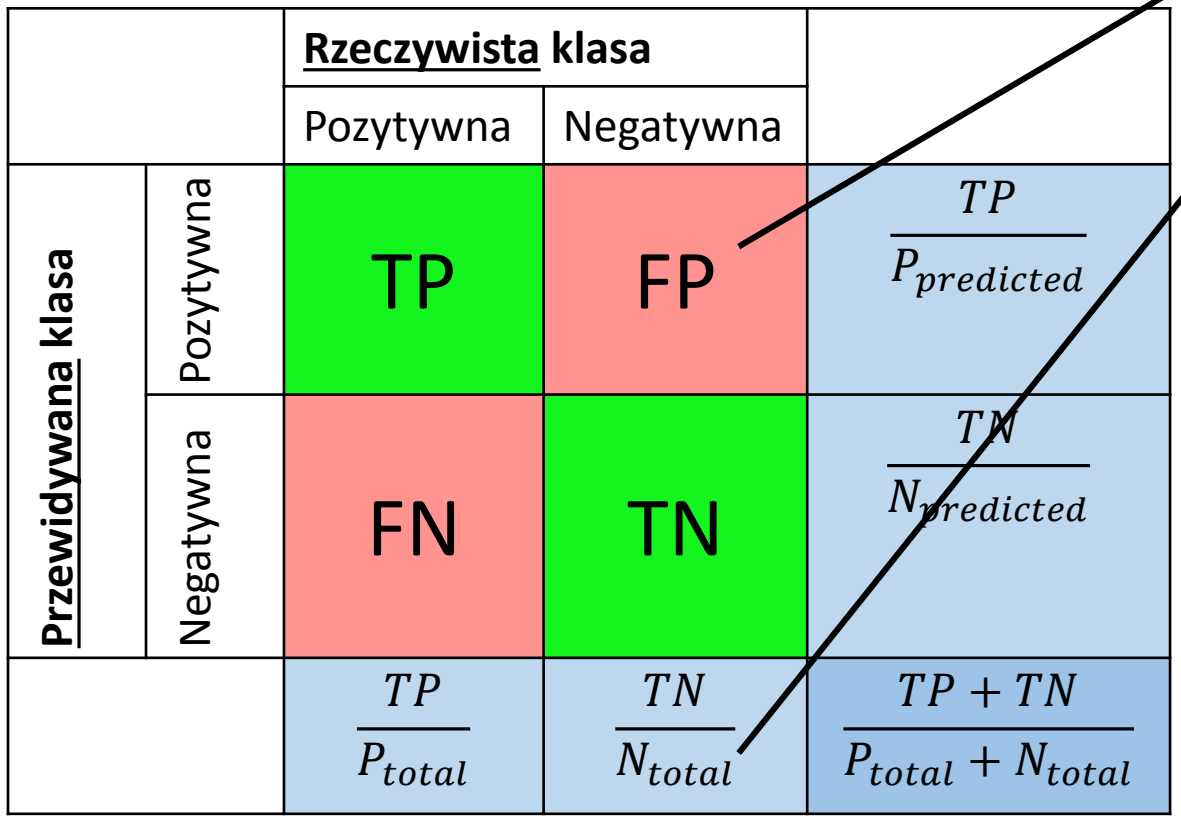
Negative Predictive Value – celność przewidywania przypadków negatywnych

Accuracy (ACC) – dokładność

Sensitivity - czułość

Specificity - swoistość

Macierz pomyłek



$\frac{FP}{N_{total}}$

False Positive Rate – odsetek predykcji fałszywie pozytywnych

Positive Predictive Value/Precision – celność przewidywania przypadków pozytywnych

Negative Predictive Value – celność przewidywania przypadków negatywnych

Accuracy (ACC) – skuteczność

Sensitivity - czułość

Specificity - swoistość

Pełna macierz pomyłek z wszystkimi parametrami

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	

http://en.wikipedia.org/wiki/Sensitivity_and_specificity

Dlaczego ogólna skuteczność klasyfikacji (ACC) nie wystarcza?

Zbiór liczy **50** pacjentów chorych (P) oraz **50** zdrowych (N)

		Rzeczywista klasa		
		P = 50	N = 50	
Przewidywana klasa	P	45		
	N		47	

Zbiór liczy **1000** pacjentów chorych (P) oraz **100** zdrowych (N)

		Rzeczywista klasa		
		P = 1000	N = 100	
Przewidywana klasa	P	997		
	N		1	

Dlaczego ogólna skuteczność klasyfikacji (ACC) nie wystarcza?

Zbiór liczy **50** pacjentów chorych (P) oraz **50** zdrowych (N)

		Rzeczywista klasa		
		P = 50	N = 50	
Przewidywana klasa	P	45		
	N		47	

Przy równolicznych zbiorach próbek pozytywnych i negatywnych Acc jest ok.

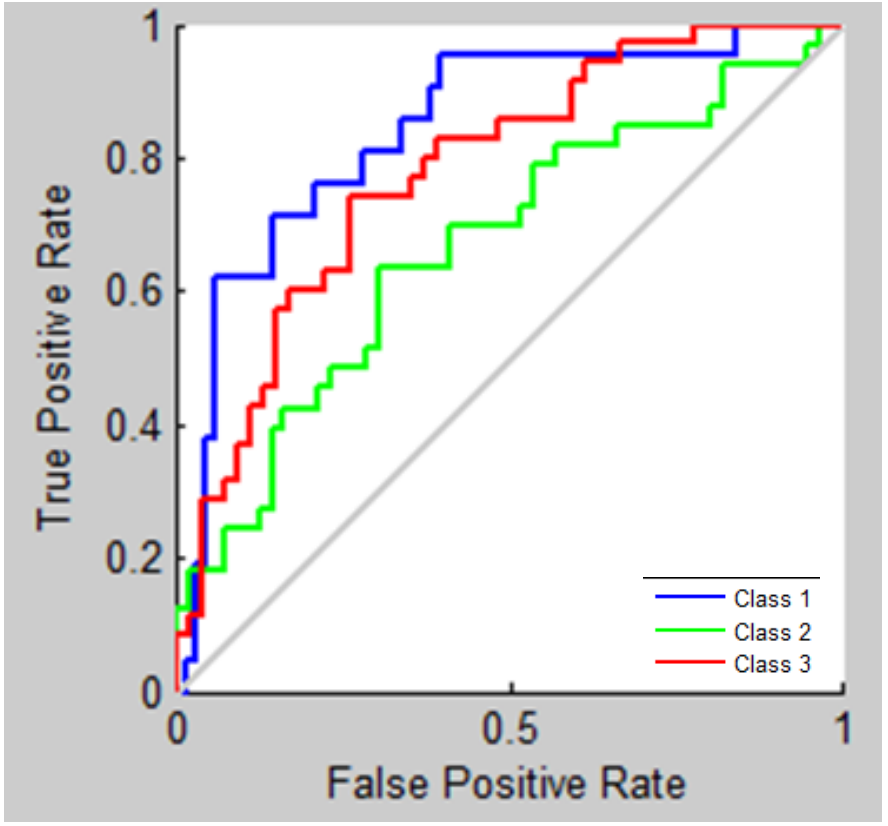
Zbiór liczy **1000** pacjentów chorych (P) oraz **100** zdrowych (N)

		Rzeczywista klasa		
		P = 1000	N = 100	
Przewidywana klasa	P	997		
	N		1	

Jeżeli jedna klasa jest **zdecydowanie** nadreprezentowana, wówczas Acc może zbyt optymistycznie oceniać klasyfikator

Krzywa ROC (Receiver Operating Characteristic)

- Cel stosowania:
  - Ocena klasyfikatora
  - Porównywanie klasyfikatorów
  - Wybór optymalnego progu odcięcia
- Pole pod krzywą (Area Under Curve - AUC) mówi o ogólnej skuteczności klasyfikatora
  - 1.0 – 0.9 bardzo dobry
  - 0.9 – 0.8 dobra
  - 0.8 – 0.7 dość dobra
  - 0.7 – 0.6 słaby
  - 0.6 – 0.5 bardzo słaby



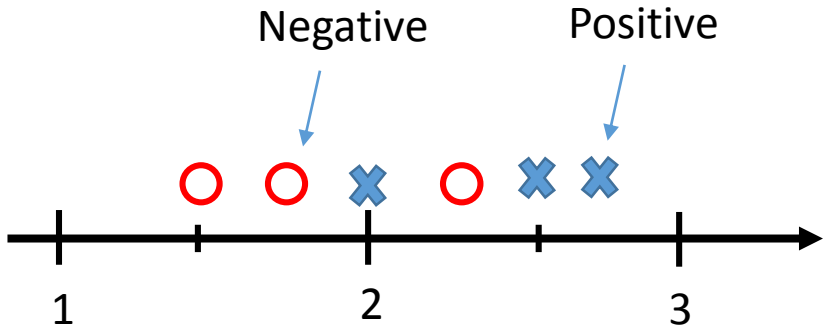
Krzywa ROC - algorytm

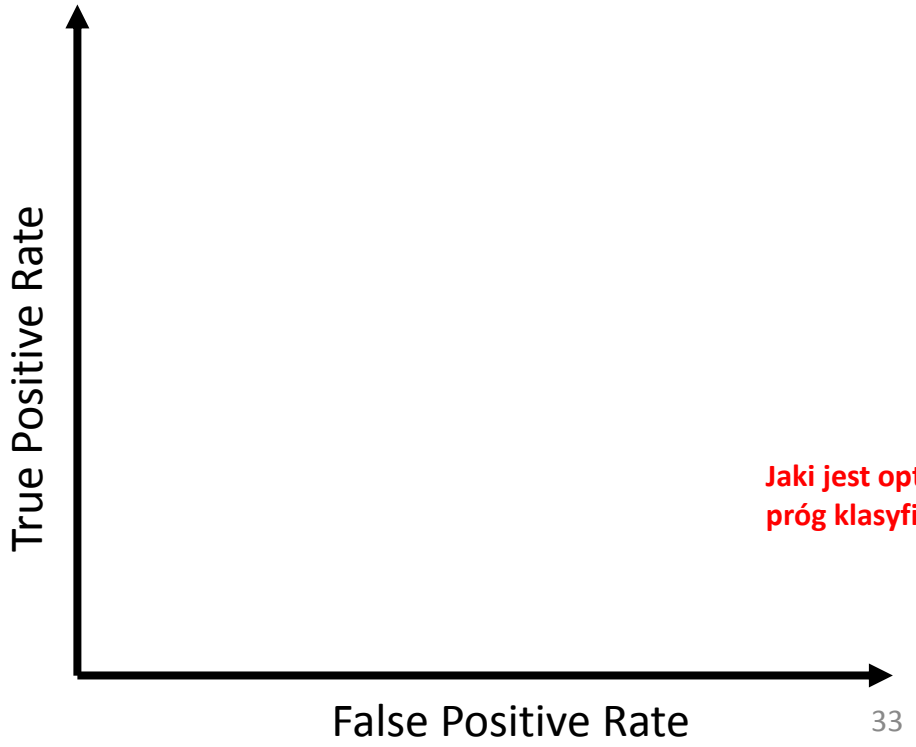
- Uszeregować od największego do najmniejszego
- Obliczać TPR i FPR przy kolejnych wartościach stężenia hormonu jako progach podziału na klasy

$$TPR = \frac{TP}{P_{total}} \qquad FPR = \frac{FP}{N_{total}}$$

$P_{total} = 3$ ✖

$N_{total} = 3$ ○

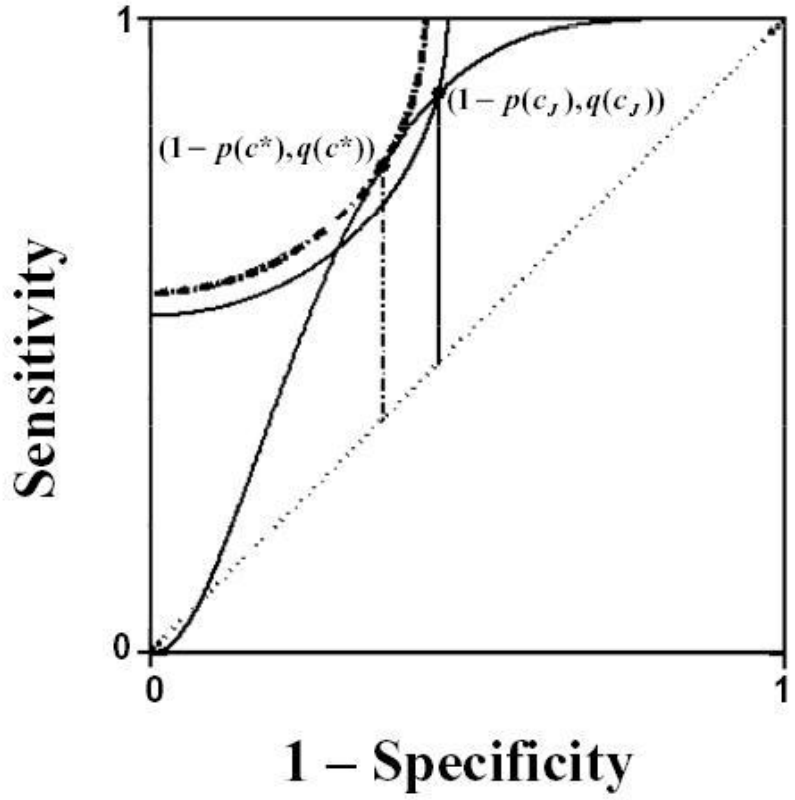




Jaki jest optymalny próg klasyfikacji?

Dobór optymalnego progu klasyfikacji

- Minimalizacja odległości od punktu (0,1) na krzywej ROC
- Maksymalizacja różnicy pomiędzy True Positive Rate i False Positive Rate – tzw. Indeks Youden'a



Perkins NJ, Schisterman EF. The Inconsistency of "Optimal" Cut-points Using Two ROC Based Criteria. *American journal of epidemiology*. 2006;163(7):670-675. doi:10.1093/aje/kwj063.

Najważniejsze pytania?

- Jaką budowę musi mieć model rozwiązujący problem regresji?
- Jak interpretować współczynniki modelu w regresji liniowej?
- Jak ocenić skuteczność dopasowania modelu liniowego?
- Jaką budowę musi mieć model rozwiązujący problem klasyfikacji?
- Jak ocenić skuteczność regresji logistycznej?