

# Regresja logistyczna

Joanna Jaworek-Korjakowska

WEAIIB, Katedra Automatyki i Robotyki, ISS

semestr letni, 2019

# Pojęcie klasyfikacji i dyskryminacji

## Definicja

**Dyskryminacja i klasyfikacja** są wielowymiarowymi metodami zajmującymi się rozdzieleniem odrębnych zbiorów obiektów (lub obserwacji) oraz przydzieleniem nowych obiektów (obserwacji) do wcześniej zdefiniowanych zbiorów (grup). Analiza dyskryminacyjna (jako procedura rozdzielająca) jest często wykorzystywana w celu zbadania obserwowanych różnic (kiedy zwyczajne relacje nie są dobrze znane). Procedury klasyfikacji prowadzą natomiast do dobrze zdefiniowanych reguł, które mogą być wykorzystywane do przydzielenia nowego obiektu do danego zbioru.

# Cel użycia klasyfikacji i dyskryminacji

Główne cele wykorzystania dyskryminacji i klasyfikacji:

- (dyskryminacja)* Do opisu zarówno graficznego (w trzech lub mniej wymiarach) jak i algebraicznego, różniących cech obiektów (obserwacji) z kilku znanych zbiorów (populacji). Staramy się znaleźć "wyróżniki", których wartości liczbowe pozwalają rozdzielić zbiory tak bardzo jak to jest możliwe.
- (klasyfikacja)* Aby posortować obiekty (obserwacje) na dwie lub więcej klas. Nacisk kładzie się na uzyskiwanie reguły, która może być używana do optymalnego przypisania nowego obiektu do poszczególnej klasy.

## Uwaga:

Funkcja która rozdziela obiekty może czasami posłużyć jako 'rozdzielnik (allocator)' i odwrotnie, zasada, która przydziela obiekty może sugerować procedurę dyskryminacji. W praktyce cele 1 i 2 często się pokrywają, a różnica między rozdzieleniem i przydzieleniem staje się niewyraźna.

# Przykłady populacji i badanych dla nich zmiennych.

Populacje $\pi_1$ i $\pi_2$	Zmienne pomiarowe X
1. Wypłacalne i źle prosperujące firmy ubezpieczeniowe (o złym stanie finansowym)	aktywa ogółem, cena akcji i obligacji, wartość rynkowa akcji i obligacji, wysokość strat, nadwyżki
2. Dwa gatunki gwiazdnicy	długość kielicha kwiatowego i płatka, długość rysy na płatku, długość przykwiatku, średnica pyłku
3. Nabywcy nowego produktu i opieszali ludzie (którzy 'powoli' kupują)	wykształcenie, dochód, wielkość rodziny, ilość poprzednich zmian marek danego produktu
4. Osoby które dostaną się i nie dostaną na studia	wyniki egzaminu wstępniego, średnia ocen z liceum, liczba zajęć w liceum
5. Mężczyźni i kobiety	pomiary antropologiczne taki jak np. obwód i objętość wykonane na starożytnych czaszkach
6. Pozytywne lub negatywne ryzyko kredytowe	dochody, wiek, ilość kart kredytowych, wielkość rodziny
7. Alkoholicy i osoby nie będące alkoholikami	aktywność enzymu oksydazy monoaminowej, aktywność enzymu cyklazy adenylanowej

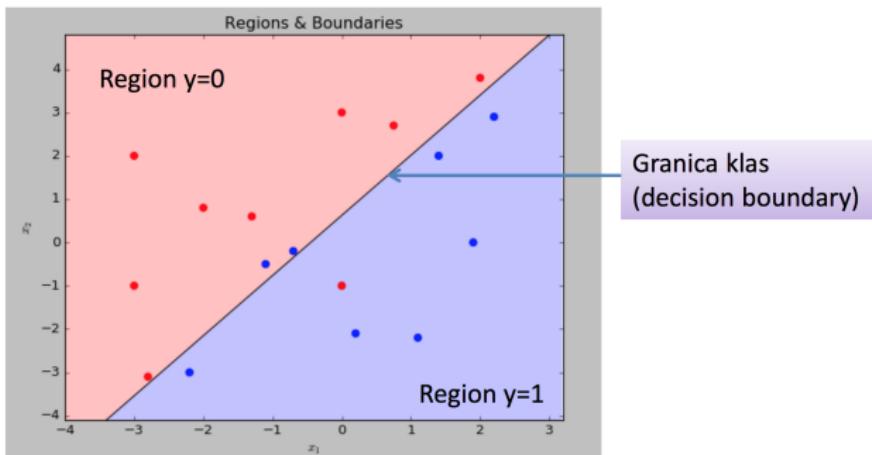
# Przykłady zagadnień klasyfikacji

- Klasyfikacja transakcji kartami płatniczymi (autoryzowane czy oszustwa)
- Klasyfikacja wniosków kredytowych (udzielić/odrzucić)
- Klasyfikacja roszczeń ubezpieczeniowych (uzasadnione czy próba wyłudzenia)
- Przewidywanie odejścia klientów firm telekomunikacyjnych (ang. churn)
- Kategoryzacja tekstów (np. wiadomości, artykułów) jako: finanse, pogoda, rozrywka, sport
- Filtrowanie spamu
- Detekcja twarzy, postaci, obiektów na obrazach
- Określanie, czy zmiany rakowe w komórkach są łagodne lub złośliwe
- Klasyfikacja struktury białek

# Porównanie klasyfikacji i regresji

- Podobieństwa
  - Model ma postać funkcji  $X \rightarrow Y$
  - W obu zagadnieniach problemem jest **wymiar  $X$**  (tzw. klątwa wymiarowości). Jeżeli wymiar  $X$  wynosi  $n$ , aby równomiernie pokryć  $X$   $k$  obserwacjami w kierunku każdego wymiaru potrzeba  $k^n$  obserwacji.
  - W obu przypadkach istotne są **zdolności generalizacji** modelu: wyznaczanie błędu testowego w zależności od złożoności, zjawisko nadmiernego dopasowania (overfitting)
  - Część modeli może być użyta zarówno do regresji, jak i klasyfikacji: drzewa regresji/decyzyjne, sieci neuronowe
- Różnice
  - W zagadnieniach klasyfikacji wartości wyjściowe są kategoryczne (dyskretnie, skończony zbiór wartości): 0/1, tak/nie
  - Możliwa jest klasyfikacja wielowartościowa (multilabel), wówczas  $Y = 2^C$ , gdzie  $C$  jest zbiorem etykiet, np. kategoryzacja tekstów
  - Stosowane są inne funkcje oceny

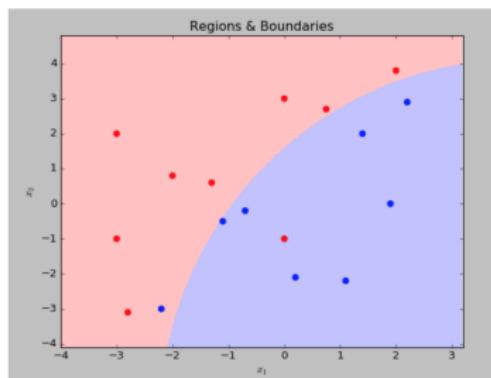
# Regiony decyzyjne i granice klas



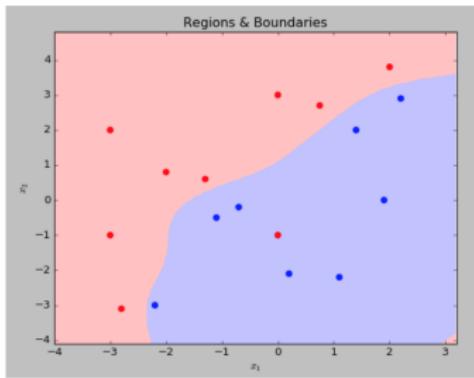
- Region decyzyjny dla danej klasy  $c_i$  to podzbiór obserwacji  $X(c_i)$ , którym klasyfikator przypisze klasę (decyzję)  $c_i$ .
- Granica klas to brzeg regionu

# Regiony decyzyjne i granice klas

- Regiony decyzyjne na ogół nie są wyznaczane analitycznie, ale są pochodną parametrów wyznaczonego modelu. Dla modeli nieparametrycznych mogą być określone wyłącznie przez testowanie wartości wejściowych.
- Kształty regionów mocno zależą od przyjętego modelu i jego złożoności



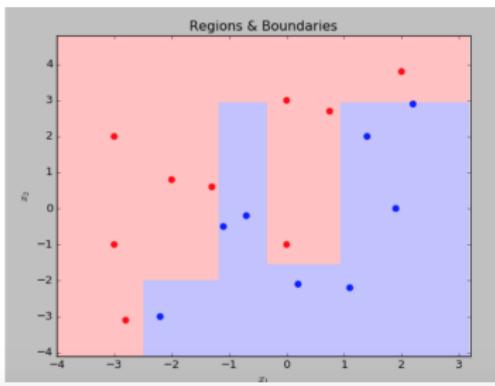
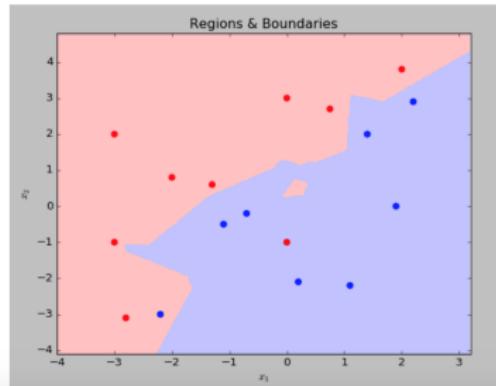
Naive Bayes



SVM + RBF kernel

# Regiony decyzyjne i granice klas

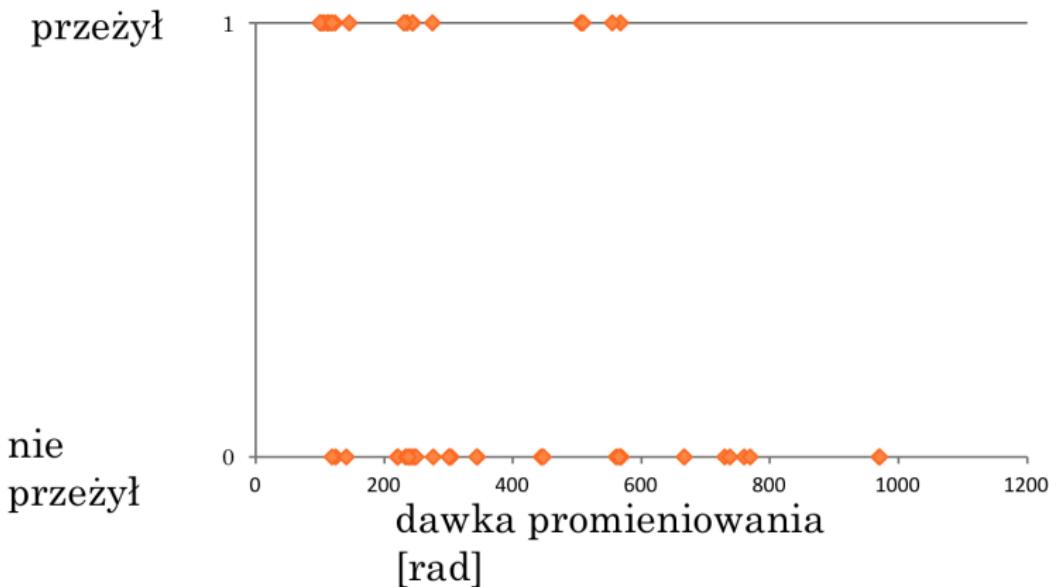
- W zależności od metody granice regionów mogą być krzywymi, separującymi dane, mieć postać łamanych, zawierać wyspy.
- Nie wszystkie obserwacje zbioru uczącego muszą być przypisane do regionu zgodnego z etykietą klasy.
- Złożone kształty regionów decyzyjnych najczęściej są oznaką nadmiernego dopasowania do danych uczących (dużej wariancji)
- Wizualizacje 2D mają raczej charakter poglądowy, niż znaczenie praktyczne



- **Generatywne (obserwacje uwarunkowane etykietami klas)**
  - Wyznaczają pełny model  $p(x|c_i)$
  - Używają reguły Bayesa do określenia granic klas
  - Przykłady: naiwny model Bayesa, Gaussian mixture model
  - Granice klas są zazwyczaj funkcjami kwadratowymi
- **Dyskryminatywne**
  - **Oparte na regresji:**
    - Modelują  $p(c_i|x)$  bezpośrednio
    - Przykłady: regresja logistyczna, sieci neuronowe
  - **Nie wykorzystujące bezpośrednio prawdopodobieństw, skupione na wyznaczaniu optymalnych granic klas:**
    - SVM (support vector machines): liniowe i nieliniowe regiony decyzyjne
    - Najbliższych sąsiadów (nearest neighbor) - granice klas w postaci łamanych
    - Drzewa decyzyjne – granice klas wzdłuż osi atrybutów

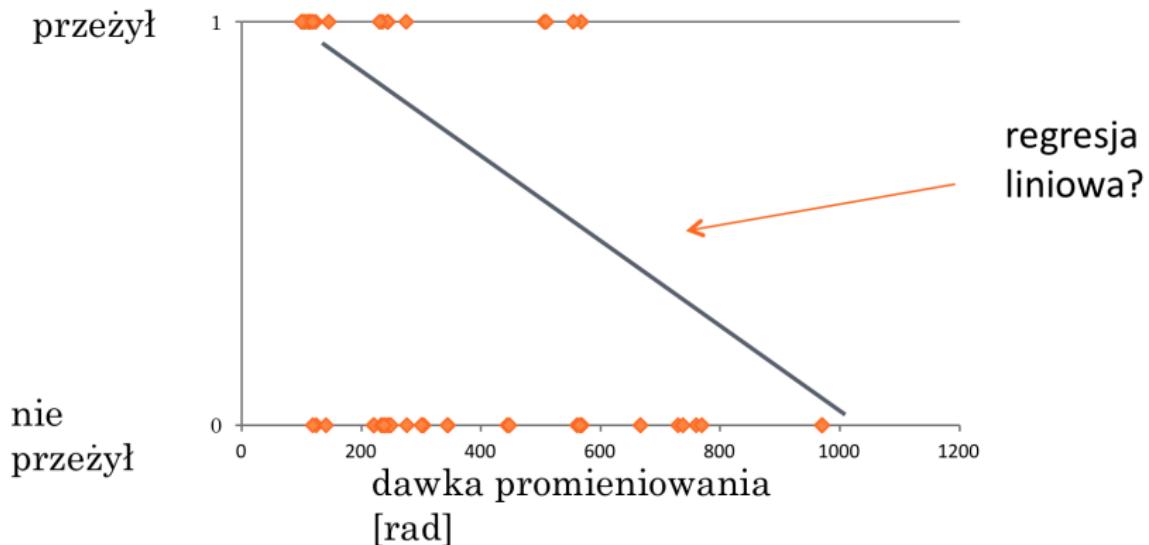
# Regresja logistyczna - przykład

120 myszy poddano różnym dawkom promieniowania w radach (dose) w określonym czasie, następnie sprawdzono czy osobnik przeżył kolejne 24 h .



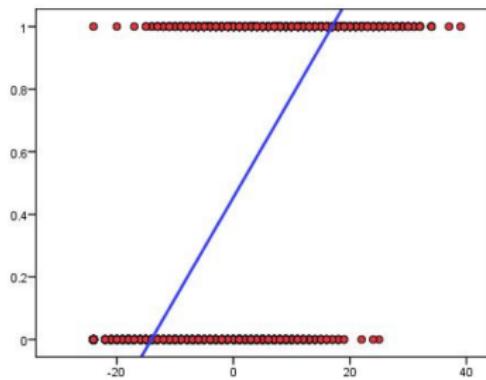
# Regresja logistyczna - przykład

120 myszy poddano różnym dawkom promieniowania w radach (dose) w określonym czasie, następnie sprawdzono czy osobnik przeżył kolejne 24 h .

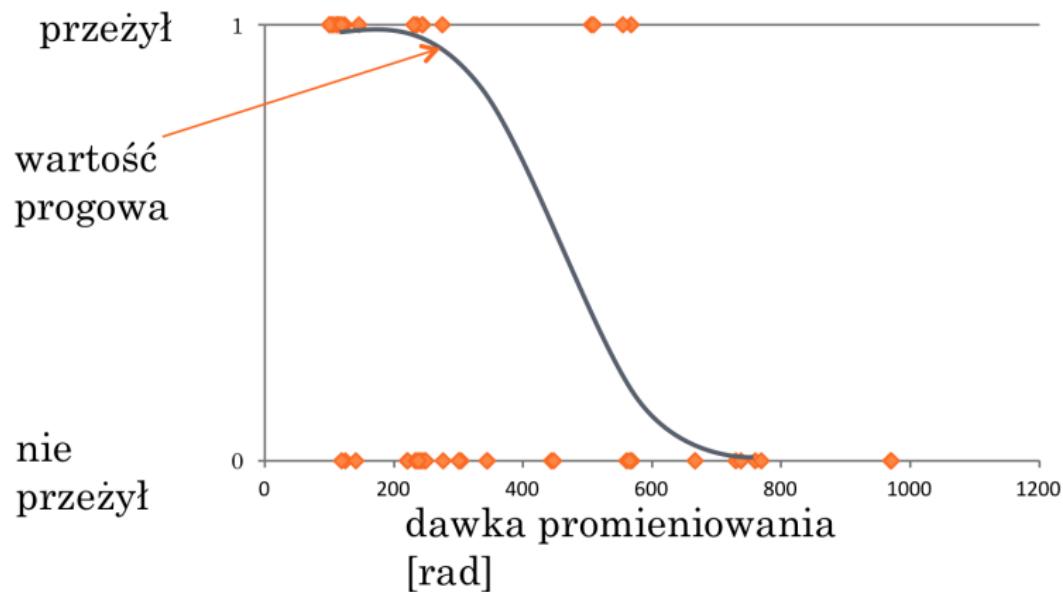


# Regresja logistyczna

- W regresji liniowej zmienne powinny być mierzone na skali ilościowej
- Problem z predykcją: dla dychotomicznej zmiennej objaśnianej regresja liniowa będzie szacowała wartości spoza akceptowalnego zakresu (poniżej 0 lub powyżej 1)
- Założenia do modelu nie będą są spełnione:
  - brak rozkładu normalnego dla reszt
  - brak jednorodności wariancji



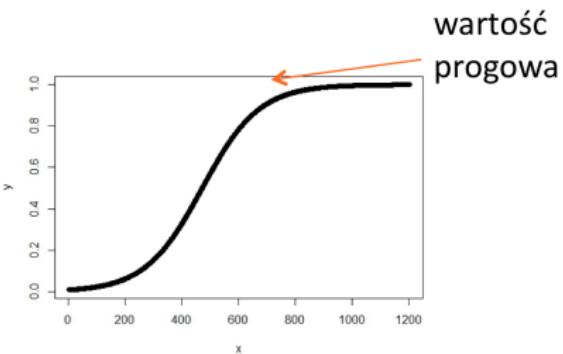
# Funkcja logistyczna



# Funkcja logistyczna

- Funkcja logistyczna

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



- Etapy zmian wartości funkcji logistycznej:
  - Dla początkowych argumentów funkcja przyjmuje **wartości bliskie zera / jedynki**
  - Od momentu osiągnięcia **wartości progowej następuje nagły wzrost / spadek** wartości funkcji
  - Po osiągnięciu pewnej wartości dla kolejnych wartości argumentów przyjmuje **wartości bliskie jedynki / zera**



# Reprezentacja hipotezy

Wcześniej powiedzieliśmy, że chcemy aby  $h_\theta(x)$  spełniał właściwość:

$$0 \leq h_\theta(x) \leq 1$$

W regresji liniowej używaliśmy hipotezy:

$$h_\theta(x) = \theta^T x$$

Teraz dokonujemy lekkiej modyfikacji, a dokładnie składamy funkcję hipotezy z regresji liniowej z nową funkcją  $g$ :

$$h_\theta(x) = g(\theta^T x)$$

Funkcja  $g$  to tzw. funkcja logistyczna (sigmoida):

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

Co sprowadza do tego że:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

# Funkcja logistyczna

Zauważamy, że sigmoida posiada asymptotę w  $y = 1$  dążącą do  $+\infty$ , oraz w  $y = 0$  dającą do  $-\infty$ .

## Jak interpretować logistyczną hipotezę?

$h_\theta(x)$  utożsamiamy z prawdopodobieństwem, że  $y = 1$  dla konkretnego wejścia  $x$ .

Przykład.

$$x = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 \\ tumorSize \end{pmatrix} \quad (1)$$

$$h_\theta(x) = 0.7$$

Stąd wynika, że prawdopodobieństwo  $y = 1$  dla danego wejścia  $x$  wynosi 0.7. W tym przypadku oznacza to, że prawdopodobieństwo złośliwości guza wynosi 70%.

# Funkcja logistyczna

Zapisując bardziej formalnie:

$$h_{\theta}(x) = P(y = 1|x; \theta) \quad (2)$$

Hipotezę interpretujemy jako prawdopodobieństwo, że  $y = 1$  dla danego wektora wejścia  $x$ , ze względu na parametr  $\theta$ .  
Z tego że prawdopodobieństwo sumuje się do 1:

$$P(y = 0|x; \theta) = 1 - P(y = 1|x; \theta) \quad (3)$$

# Jak wybrać parametr $\theta$

Dla regresji liniowej, używaliśmy błędu średniokwadratowego

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Zdefiniujmy to w następujący sposób:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(h_\theta(x^{(i)}), y^{(i)})$$

$$Cost(h_\theta(x), y) = \frac{1}{2} (h_\theta(x) - y)^2$$

Funkcja *Cost* reprezentuje koszt, który musi zapłacić masz algorytm, jeśli zwraca odpowiedź  $h_\theta(x)$ , gdy rzeczywistą wartością jest  $y$ .

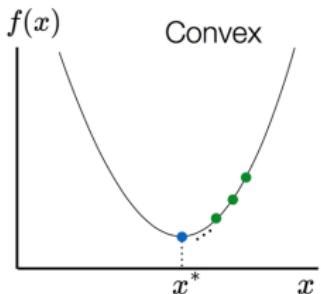
# Jak wybrać parametr $\theta$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(h_\theta(x^{(i)}), y^{(i)})$$

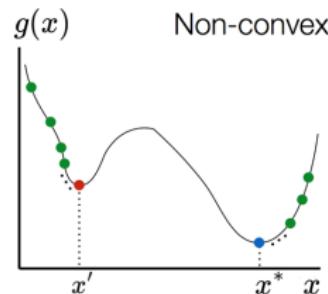
$$Cost(h_\theta(x), y) = \frac{1}{2}(h_\theta(x) - y)^2$$

Możemy użyć tej funkcji podczas minimalizacji regresji logistycznej, ale istnieje problem, mianowicie jest to funkcja niewypukła.

Sigmoida w hipotezie wprowadza nieliniowość, która powoduje w powyższej funkcji kosztu powstanie wielu minimów lokalnych.



Any local minimum is a global minimum



Multiple local minima may exist

# Funkcja kosztu dla regresji logistycznej

Użyjemy następującej funkcji kosztu:

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{gdy } y = 1 \\ -\log(1 - h_\theta(x)) & \text{gdy } y = 0 \end{cases}$$

Mamy

$$Cost(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x))$$

Możemy teraz podstawić powyższy wzór na  $J(\theta)$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(h_\theta(x^{(i)}), y^{(i)})$$

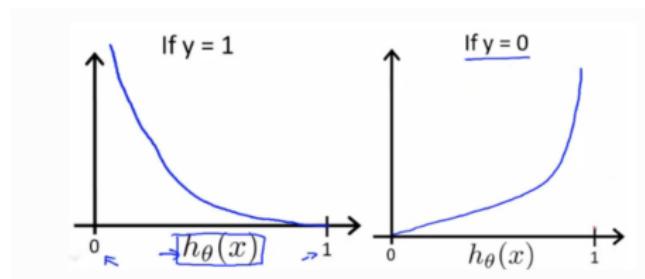
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m [-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Można się zastanawiać dlaczego zdecydowaliśmy się właśnie na tą funkcję kosztu. Poza tym że ma tą pozytywną cechę że jest wypukła, może ona zostać wprowadzona z praw statystyki przy użyciu zasady maksymalnego prawdopodobieństwa, czyli zasady mówiącej o tym jak wydajnie szukać parametrów różnych modeli.

# Korzyści z funkcji logarytmicznej

Korzyści wynikające z zastosowania logarytmu ujawniają się, gdy spojrzymy na wykresy funkcji kosztu dla  $y = 1$  i  $y = 0$ .

Te gładkie funkcje monotoniczne (zawsze rosnące lub zawsze malejące) ułatwiają obliczanie gradientu i minimalizowanie kosztów.



Funkcja ta ma kilka ciekawych, użytecznych wartości:

- Koszt = 0, gdy  $y = 1$ ,  $h_{\theta}(x) = 1$
- Ale gdy  $h_{\theta}(x) \rightarrow 0$ , to Koszt  $\rightarrow \infty$
- Ujmuję to intuicję, że gdy  $h_{\theta}(x) = 0$  (przewidujemy  $P(y = 1|x; \theta) = 0$ ), ale  $y = 1$ , to obciążamy nasz algorytm uczący bardzo wysokim kosztem.

# Gradient prosty

Naszym dalszym celem będzie znalezienie takich parametrów  $\theta$  dla których  $J(\theta)$  będzie minimalne. Można to osiągnąć za pomocą metody gradientu prostego, czyli powtarzać

$$\theta_j = \theta_j - \alpha \frac{d}{d\theta_j} J(\theta)$$

Gdzie  $\alpha$  jest pewną stałą wybraną przez nas, oznaczającą jak długie kroki robi algorytm w każdej iteracji.

Można obliczyć że

$$\frac{d}{d\theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Algorytm regresji liniowej od regresji logistycznej różni się tylko definicją funkcji hipotezy.

# Algorytmy optymalizacji

Istnieją inne algorytmy, których możemy użyć do minimalizacji, dostarczając im funkcję kosztu i gradient:

- Conjugate gradient
- BFGS
- L-BFGS

Zalety zaawansowanych algorytmów optymalizacji:

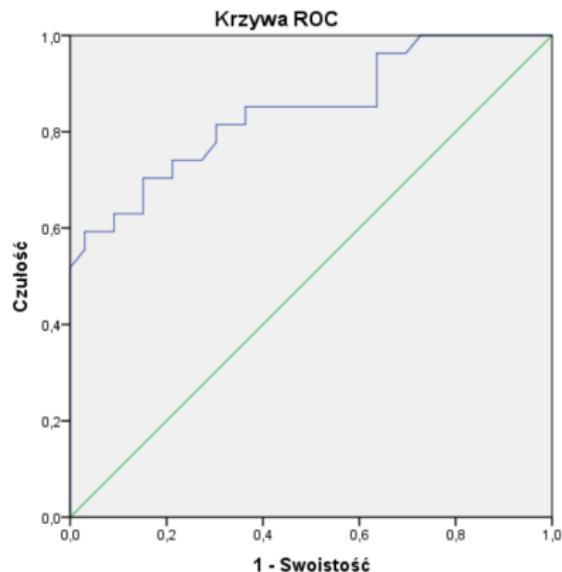
- Nie trzeba ręcznie wybierać wartości  $\alpha$  (algorytm automatycznie próbuje różnych wartości, dobierając odpowiednią)
- Często szybsze niż gradient descent

Wady:

- Bardziej złożone (z tego względu raczej nie powinniśmy implementować ich samodzielnie, tylko skorzystać z gotowych bibliotek)

# Jakość klasyfikatora

## JAKOŚĆ KLASYFIKACJI DLA ZMIENNEJ DIAGNOSTYCZNEJ



Krzywa wyznaczana dla wszystkich możliwych punktów odcięcia

Czułość (oś Y) opisuje częstość względną wystąpień prawdziwie dodatnich

1 – Swoistość (oś X) opisuje częstość względną wystąpień fałszywie dodatnich

Dobra klasyfikacja:  
Dla danego punktu odcięcia czułość > 1 - swoistość

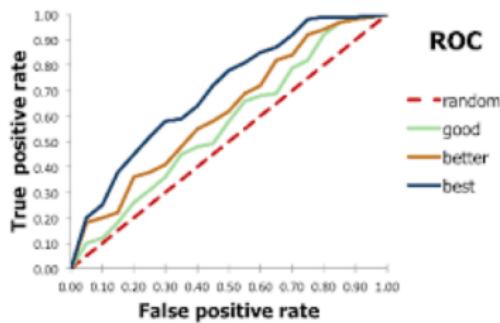
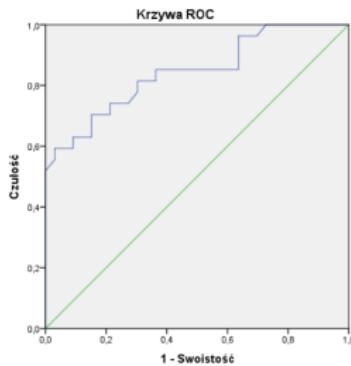
Gdy krzywa ROC pokrywa się z przekątną  $y = x$  (zielona linia), to decyzja podejmowana na podstawie zmiennej diagnostycznej jest tak samo dobra jak losowy podział badanych obiektów na grupy



# Jakość klasyfikatora

## JAKOŚĆ KLASYFIKACJI DLA ZMIENNEJ DIAGNOSTYCZNEJ

- Wielkość pola pod krzywą ROC mieści się w przedziale  $<0 ; 1>$
- Krzywa powstaje na podstawie wyznaczonych wartości czułości i swoistości
- Im większe pole tym dokładniej sklasyfikujemy dane do grupy na podstawie analizowanej zmiennej diagnostycznej



# Macierz pomyłek

True\predicted	$c_1$	$c_2$	$c_3$
$c_1$	5	0	2
$c_2$	1	7	2
$c_3$	1	4	3

$precision(c_i)$

$recall(c_i)$

nieprawidłowo sklasyfikowane

prawidłowo sklasyfikowane

The diagram shows a 3x4 confusion matrix with rows labeled by true classes  $c_1, c_2, c_3$  and columns labeled by predicted classes  $c_1, c_2, c_3$ . The matrix entries are:  $c_1 \rightarrow c_1: 5$ ,  $c_1 \rightarrow c_2: 0$ ,  $c_1 \rightarrow c_3: 2$ ;  $c_2 \rightarrow c_1: 1$ ,  $c_2 \rightarrow c_2: 7$ ,  $c_2 \rightarrow c_3: 2$ ;  $c_3 \rightarrow c_1: 1$ ,  $c_3 \rightarrow c_2: 4$ ,  $c_3 \rightarrow c_3: 3$ . A green double-headed vertical arrow on the left is labeled  $precision(c_i)$ . A blue double-headed horizontal arrow below the matrix is labeled  $recall(c_i)$ . A red double-headed arrow pointing to the right from the bottom of the matrix is labeled "nieprawidłowo sklasyfikowane" (incorrectly classified). A blue double-headed arrow pointing to the right from the right side of the matrix is labeled "prawidłowo sklasyfikowane" (correctly classified).

$$accuracy = \frac{\sum_{i=1}^k e[i, i]}{\sum_{i=1}^k \sum_{j=1}^k e[i, j]}$$

Wartości  $precision$  i  $recall$  mogą być wyznaczone dla poszczególnych klas:

$$precision(c_i) = \frac{e[i, i]}{\sum_{j=1}^k e[j, i]}$$

$$recall(c_i) = \frac{e[i, i]}{\sum_{j=1}^k e[i, j]}$$

Obliczenie miary F1 wymaga uśrednienia wyników. Może to być zrealizowane na dwa sposoby: mikro i makro

## Mikro

1. Dla każdej klasy  $c_i$  oblicz  $TP(c_i)$ ,  $FP(c_i)$  oraz  $FN(c_i)$
2. Oblicz  $precision = \frac{\sum_{c_i} TP(c_i)}{\sum_{c_i} (TP(c_i) + FP(c_i))}$
3. Oblicz  $recall = \frac{\sum_{c_i} TP(c_i)}{\sum_{c_i} (TP(c_i) + FN(c_i))}$
4. Oblicz  $F_1 = 2 * \frac{precision * recall}{precision + recall}$

## Makro

1. Dla każdej klasy  $c_i$ 
  1. oblicz  $TP(c_i)$ ,  $FP(c_i)$  oraz  $FN(c_i)$
  2. Wyznacz  $prec(c_i) = \frac{TP(c_i)}{TP(c_i)+FP(c_i)}$
  3. Wyznacz  $recall(c_i) = \frac{TP(c_i)}{TP(c_i)+FN(c_i)}$
2. Oblicz średnie wartości
  1.  $precision = \frac{\sum_{c_i} prec(c_i)}{|\{c_i\}|}$
  2.  $recall = \frac{\sum_{c_i} recall(c_i)}{|\{c_i\}|}$
3. Oblicz  $F_1 = 2 * \frac{precision*recall}{precision+recall}$