

Metody eksploracji danych

4. Klasyfikacja

Piotr Szwed

Katedra Informatyki Stosowanej AGH

2017

Wprowadzenie
Ocena klasyfiaktorów
Regresja Logistyczna

Zagadnienie klasyfikacji

- Dane:
 - Zbiór uczący: $D = \{(x_i, y_i)\}_{i=1,m}$
 - (x_i, y_i) – obserwacje, instancje
 - x_i - obserwacje (wartości atrybutów różnego typu)
 - $y_i \in \{c_1, \dots, c_k\}$ - etykiety klas
- Celem jest znalezienie modelu $c(x): x \rightarrow \{c_1, \dots, c_k\}$ pozwalającego na przypisanie etykiety klasy nieznanym wektorom x
- Klasyfikacja binarna $k = 2$ wówczas $c(x): x \rightarrow \{c_1, c_2\}$
- Efektywność otrzymanego modelu powinna być oceniana z użyciem odrębnego zbioru danych
 - Podział (test split) oryginalnego zbioru danych na D_{train} i D_{test}
 - Zastosowanie walidacji krzyżowej (CV, cross validation)

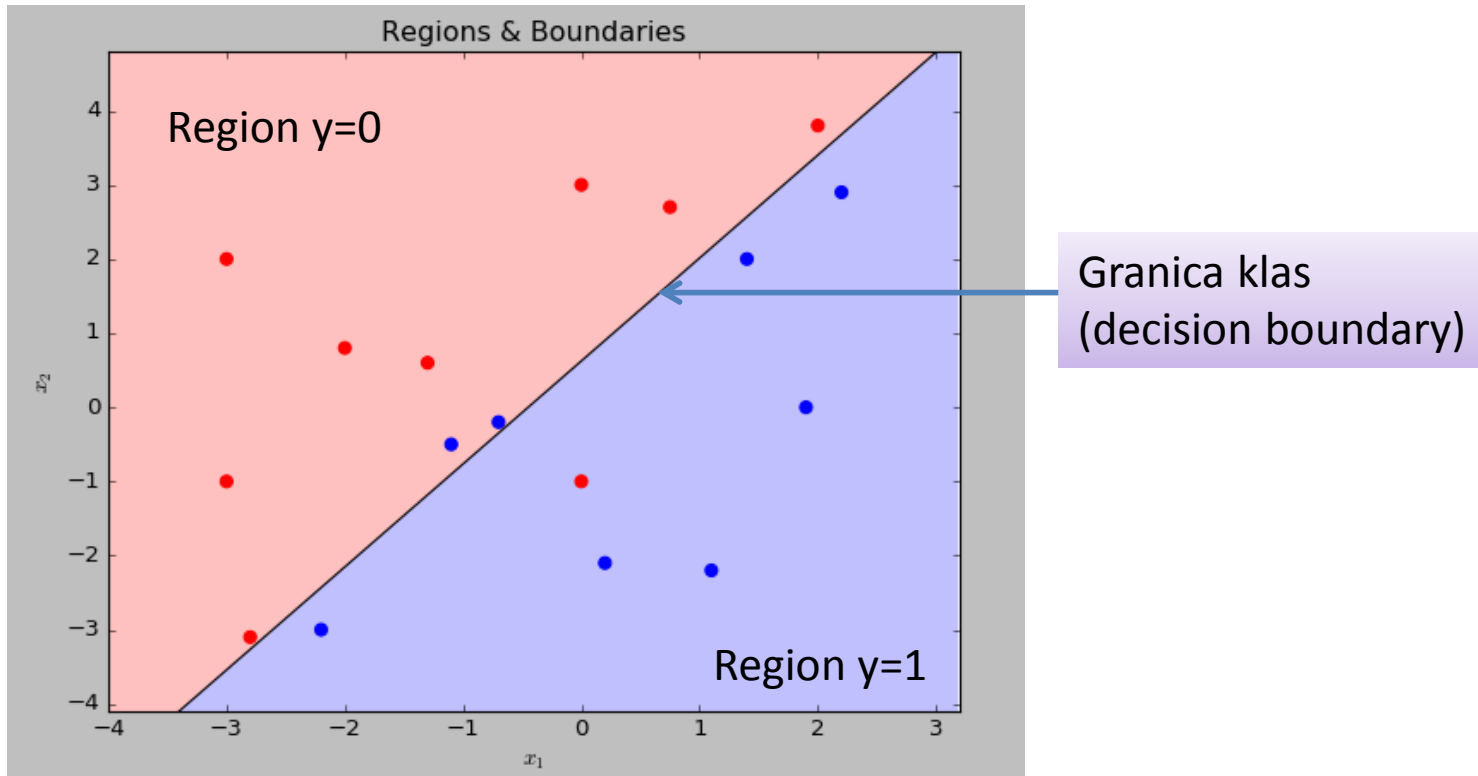
Przykłady zagadnień klasyfikacji

- Klasyfikacja transakcji kartami płatniczymi (autoryzowane czy oszustwa)
- Klasyfikacja wniosków kredytowych (udzielić/odrzuć)
- 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 (dyskretne, 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

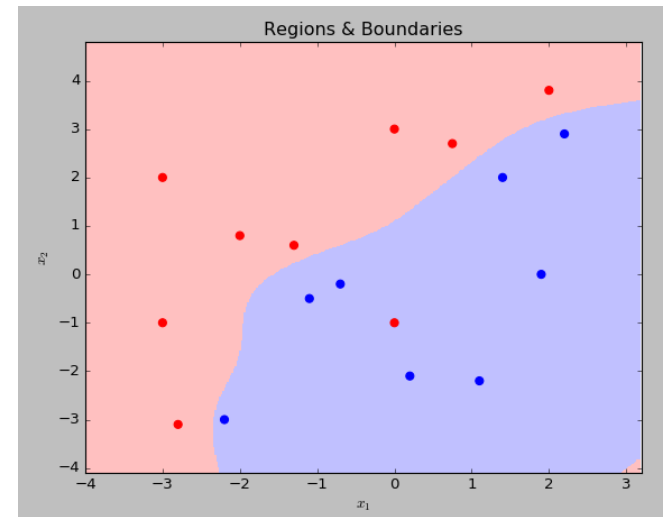
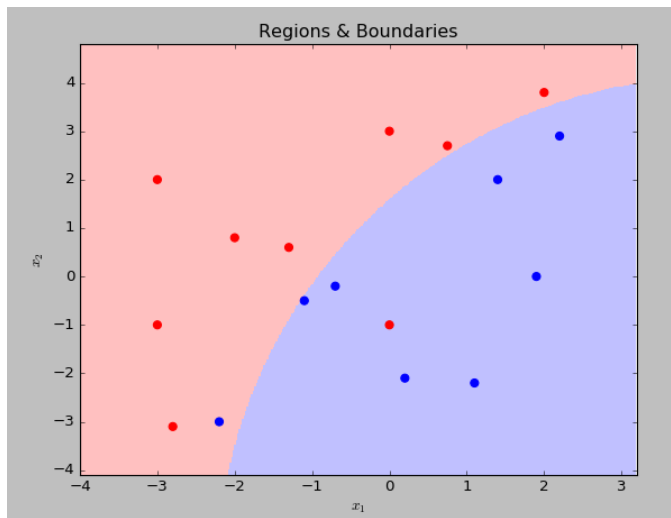
Regiony decyzyjne i granice klas



- Region decyzyjny dla danej klasy c_i to podzbiór obserwacji $X(c_i) = \{x \in X : c(x) = c_i\}$, którym klasyfikator przypisze klasę (decyzję) c_i .
- Granica klas to brzeg regionu

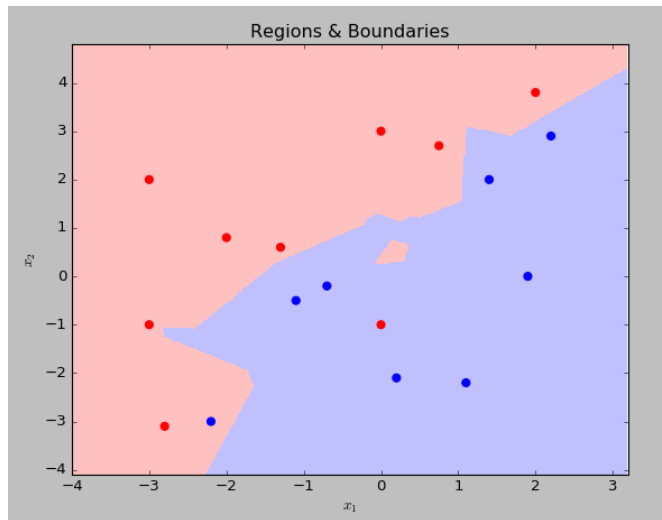
Regiony decyzyjne i granice klas

- 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

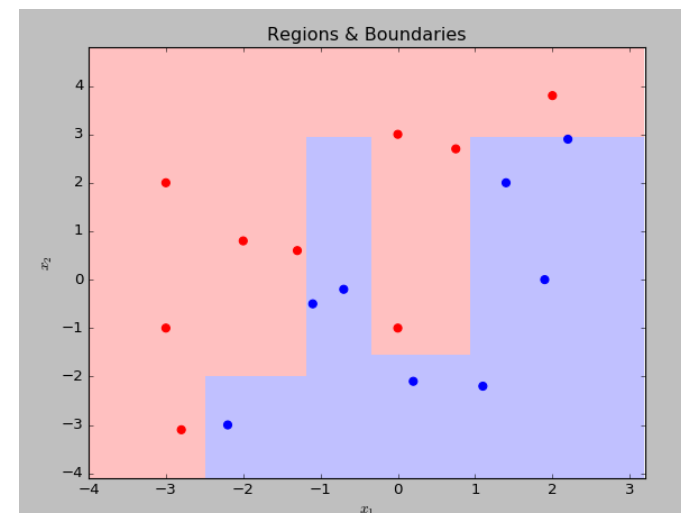


Regiony decyzyjne i granice klas: inne przykłady

- 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



k-NN (k=3)



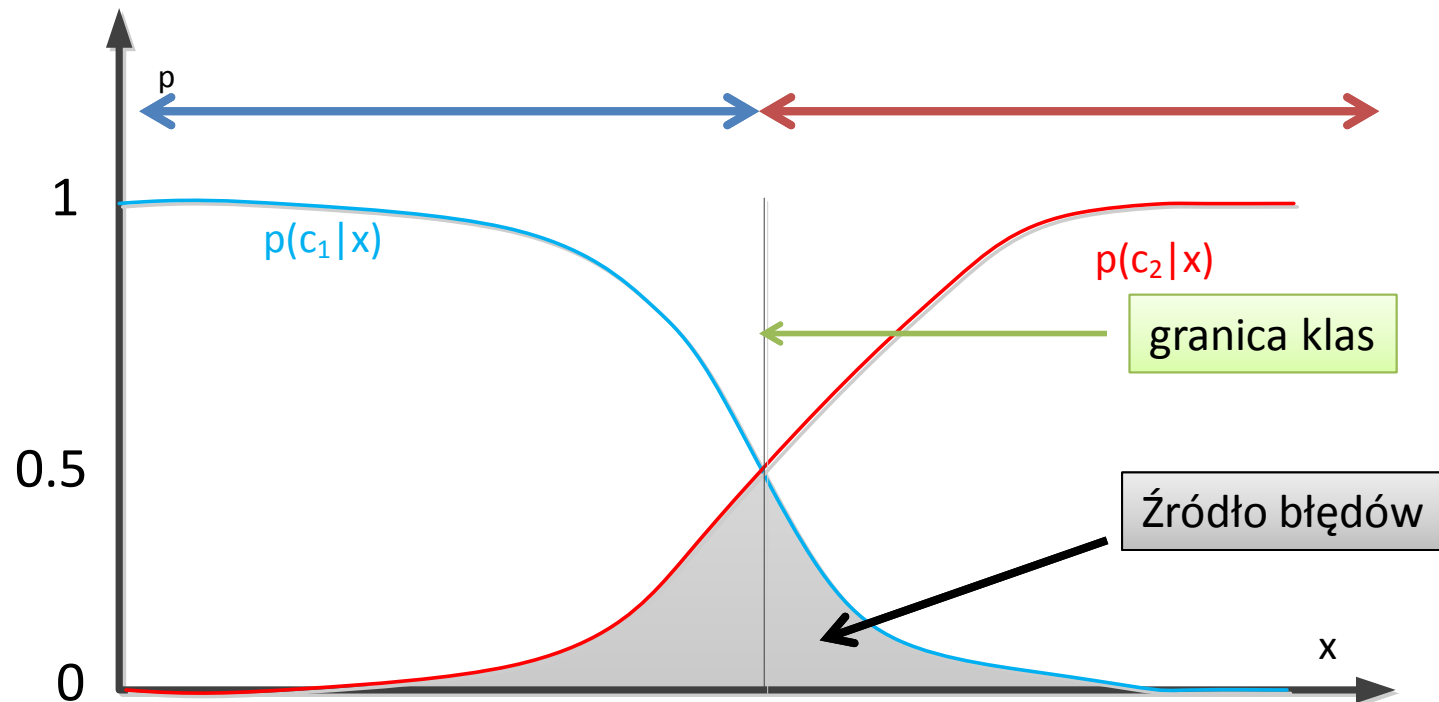
Drzewo decyzyjne

Perspektywa probabilistyczna

- Niech $C = \{c_1, \dots, c_k\}$
- $p(c_i)$ to prawdopodobieństwo wystąpienia klasy c_i w pewnym zbiorze obserwacji.
 - Jeżeli zbiorem tym jest zbiór danych uczących D i dla pewnej klasy c_j , $p(c_j)$ jest małe, wówczas zbiór ten nazywany jest niezrównoważonym (niezbalansowanym, ang. unbalanced).
- $p(c_i|x)$ to prawdopodobieństwo, przypisania do obserwacji x klasy c_i .
- Zakładając, że klasyfikator jest w stanie wyznaczyć $p(c_i|x)$, dla $i = 1, \dots, k$, wówczas optymalną decyzją jest wybór klasy:
$$c_m = \arg \max \{p(c_i|x) : i = 1, k\}$$

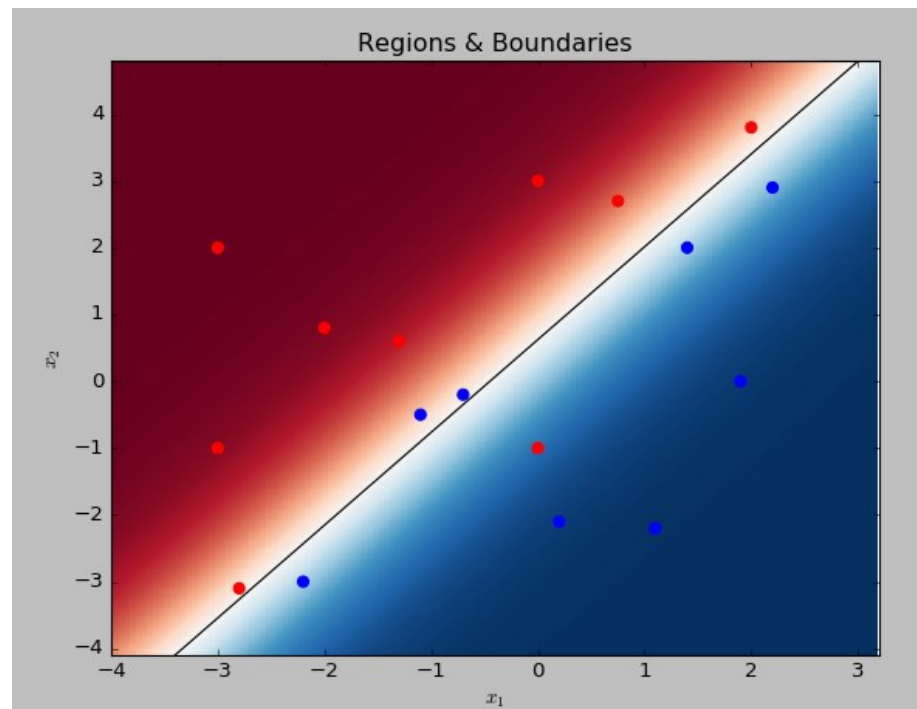
Jednowymiarowy przypadek klasyfikacji binarnej

- Regiony decyzyjne dla danej klasy : wartości, gdzie klasa jest bardziej prawdopodobna, czyli $p(c_1|x) > p(c_2|x)$
- Granica klas: $p(c_1|x) = p(c_2|x) = 0.5$
- **Sigmoidalna** funkcja prawdopodobieństwa: dowolny zakres x może zostać odwzorowany w przedział $[0,1]$



Przykład: regresja logistyczna

- Binarne zagadnienie klasyfikacji
- Granica klas jest hiperpłaszczyzną (linią w przypadku dwuwymiarowym)
- Prawdopodobieństwo jednej z klasy rośnie, a drugiej maleje wraz z oddalaniem się od granicy klas
- Na granicy: $p(c_1|x) = p(c_2|x) = \frac{1}{2}$



Modele Bayesowskie

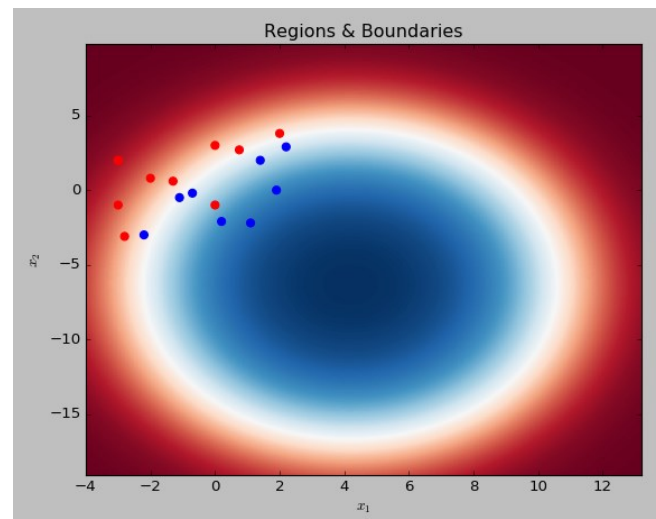
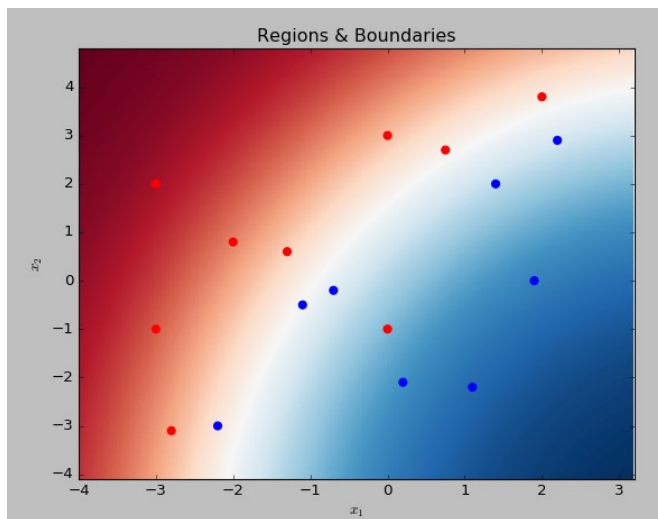
- Modele Bayesowskie wykorzystują prawdopodobieństwo wystąpienia obserwacji x dla klasy c_i , czyli prawdopodobieństwa warunkowe:

$$p(x|c_i), i = 1, \dots, k$$

- Następnie $p(c_i|x)$ jest wyznaczane za pomocą reguły Bayesa:

$$p(c_i | x) = \frac{p(x|c_i)p(c_i)}{p(x)},$$

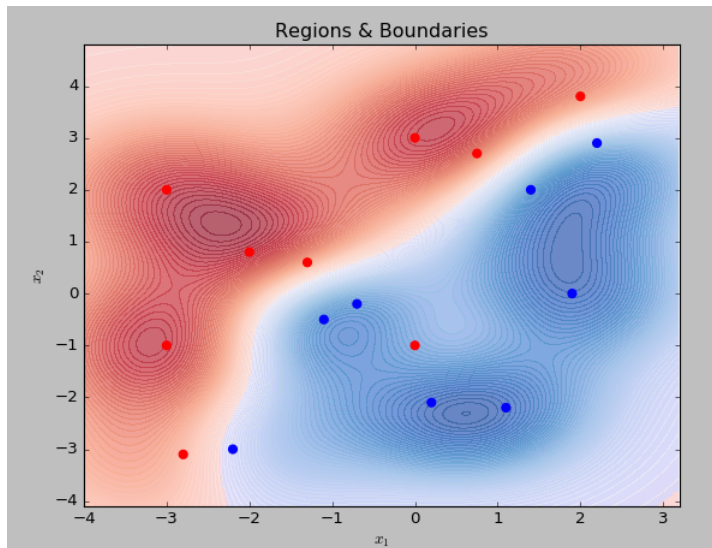
- Prawdopodobieństwo $p(x)$ jest obliczane jako: $p(x) = \sum p(x|c_i)p(c_i)$
- Ponieważ celem jest wyznaczenie $\arg \max \{p(c_i|x): i = 1, k\}$, prawdopodobieństwo $p(x)$ jako wspólny czynnik skalujący może zostać pominięte.



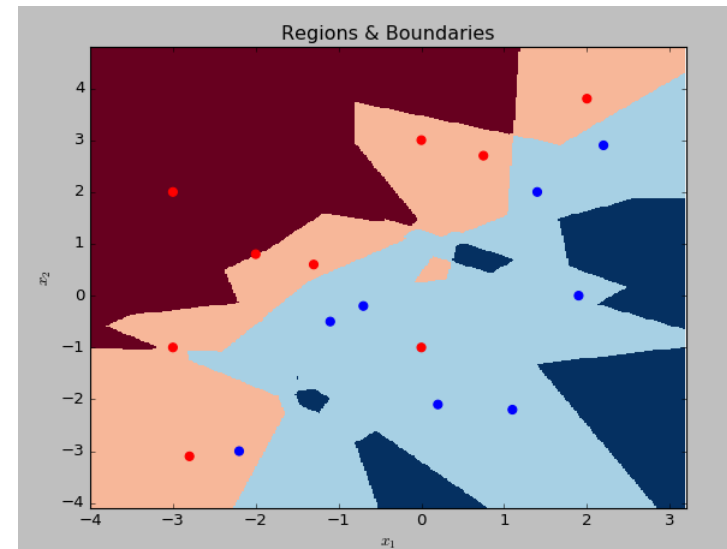
Rozkład
 $p(c_i | x)$
wyznaczony
przez naiwny
klasyfikator
Bayesa

Inne metody

- Wiele metod nie korzysta z modeli probabilistycznych. Regiony decyzyjne i granice klas są wyznaczane na podstawie różnych parametrów:
 - SVM: margines pomiędzy obserwacjami
 - drzewa: zysk informacyjny
 - kNN: klasa przeważająca w sąsiedztwie
- Dla części z nich prawdopodobieństwa $p(c_i | x)$ mogą zostać przypisane do obserwacji x po wyznaczeniu modelu.



SVM z kernelem RBF



kNN (k=3)

Podział klasyfikatorów

- **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

Ocena klasyfikatorów

Funkcje oceny klasyfikatora

- Podstawową funkcją oceny klasyfikatora jest trafność klasyfikacji.
- Oznaczmy przez $L(i, j)$ koszt błędnej klasyfikacji:

$$L(i, j) = \begin{cases} 1 & \text{gdy } i \neq j \\ 0 & \text{gdy } i = j \end{cases}$$

$$accuracy = \frac{1}{m} \sum_{i=1}^m L(y_i, c(x_i))$$

- Jednakże algorytmy uczące nie minimalizują bezpośrednio tej funkcji, raczej posługują się różnymi funkcjami zastępczymi, które są optymalizowane

Macierz pomyłek (błędów)

Macierz pomyłek (błędów)

Macierz pomyłek (ang. *confussion matrix*, także *contingency table*) zestawia dane dotyczące prawidłowo/nieprawidłowo sklasyfikowanych przypadków

Przykład

Klasyfikacja binarna, np. detekcja choroby za pomocą badania diagnostycznego. Testy przeprowadzone na grupie 50 pacjentów chorych i 50 pacjentów zdrowych

	Sklasyfikowany jako chory	Sklasyfikowany jako zdrowy	
Pacjent chory	34	16	← nieprawidłowo sklasyfikowane
Pacjent zdrowy	7	43	

Macierz pomyłek dla binarnej klasyfikacji

	Sklasyfikowany jako chory	Sklasyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

- P (przypadki pozytywne) $P = TP + FN$
- N (przypadki negatywne) $N = FP + TN$

Trafność (accuracy)

	Sklasyfikowany jako chory	Sklasyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

$$accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + FP + TN}$$

TP = 34	FN = 16
FP = 7	TN = 43

$$acc = \frac{34 + 43}{100} = 0.770$$

TP = 8	FN = 2
FP = 3	TN = 987

$$acc = \frac{8 + 987}{1000} = 0.995$$

TP = 0	FN = 10
FP = 0	TN = 990

$$acc = \frac{0 + 990}{1000} = 0.990$$

Wyniki niemal identyczne, ale w ostatnim przypadku badanie (klasyfikator) nigdy nie wykryje choroby

Czułość

W przypadku niezrównoważonych danych trafność jest złą miarą. Oczekujemy jednak, że badanie wskaże chorobę w przypadku, kiedy pacjent jest rzeczywiście chory.

Czułość (ang. *sensitivity*), także TPR (*true positive rate*), także przywołanie (ang. *recall*)

$$sensitivity = TPR = \frac{TP}{TP + FN}$$

	Sklassyfikowany jako chory	Sklassyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

Czułość wyznaczana dla dużej liczby instancji może być traktowana jako prawdopodobieństwo wykrycia choroby przez badanie (klasyfikator).

$$sensitivity = P(\text{Sklassyfikowany jako chory} | \text{Pacjent chory})$$

Recall

Przy ocenie klasyfikatorów częściej używanym terminem jest przywołanie (ang. *recall*).

Wyobraźmy sobie, że zadaniem klasyfikatora jest wybranie dokumentów na temat sportu w zbiorze milionów dokumentów. Interesuje nas, aby znalazł ich jak najwięcej.

$$recall = \frac{|\{\text{znalezione dokumenty}\} \cap \{\text{dokumenty na temat sportu}\}|}{|\{\text{dokumenty na temat sportu}\}|}$$

- Licznik – liczba zwróconych dokumentów na temat sportu (czyli TP)
- Mianownik – liczba wszystkich dokumentów na temat sportu w całym zbiorze (czyli $P = TP + FN$)

	Sklasyfikowany jako sport	Sklasyfikowany jako inne
sport	TP (true positive)	FN (false negative)
inne	FP (false positive)	TN (true negative)

Obliczmy czułość

	Sklassyfikowany jako chory	Sklassyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

$$\text{sensitivity} = \text{recall} = TPR = \frac{TP}{TP + FN}$$

TP = 34	FN = 16
FP = 7	TN = 43

$$TPR = \frac{34}{50} = 0.680$$

TP = 8	FN = 2
FP = 3	TN = 987

$$TPR = \frac{8}{10} = 0.800$$

TP = 0	FN = 10
FP = 0	TN = 990

$$TPR = \frac{0}{10} = 0$$

Specyficzność

Specyficzność (ang. *specificity*) jest analogiczną miarą dla przypadków negatywnych.

$$specificity = TNR = \frac{TN}{TN + FP}$$

	Sklassyfikowany jako chory	Sklassyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

Specyficzność (swoistość) wyznaczana dla dużej liczby instancji może być w tym przypadku interpretowana jako prawdopodobieństwo odrzucenia diagnozy choroby w przypadku zdrowych pacjentów:

$$specificity = P(\text{Sklassyfikowany jako zdrowy} | \text{Pacjent zdrowy})$$

Specyficzność

$$\text{specificity} = TNR = \frac{TN}{TN + FP}$$

	Sklassyfikowany jako chory	Sklassyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

TP = 34	FN = 16
FP = 7	TN = 43

$$TNR = \frac{43}{50} = 0.880$$

TP = 8	FN = 2
FP = 3	TN = 987

$$TNR = \frac{987}{990} = 0.996$$

TP = 0	FN = 10
FP = 0	TN = 990

$$TNR = \frac{990}{990} = 1$$

ROC (Receiver Operating Characteristic)

Miara rozwinięta podczas II Wojny Światowej na potrzeby oceny efektywności operatorów radaru. Operator ma ustalić czy punkt na ekranie to:

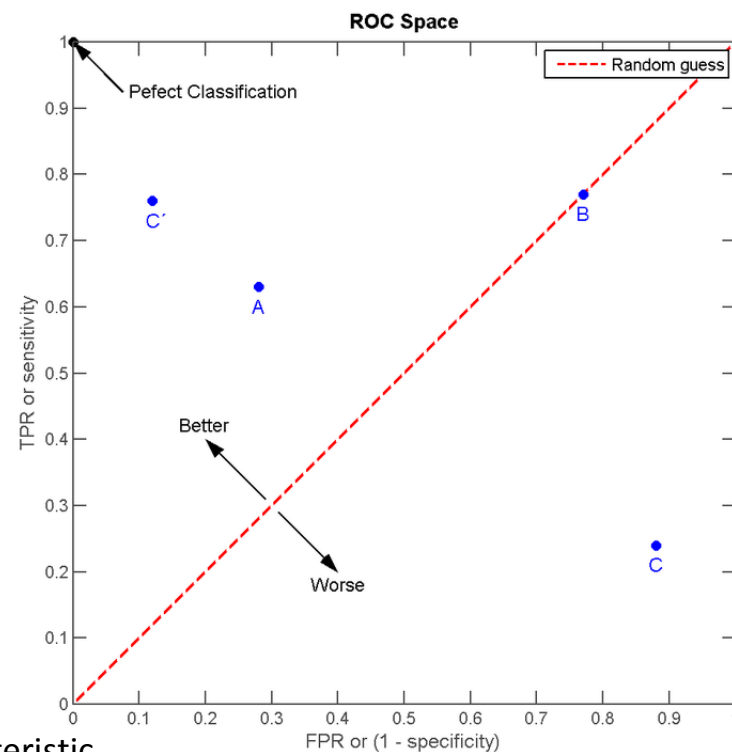
- samolot wroga (przypadek pozytywny)
- samolot własny lub stado ptaków (przypadek negatywny)

Przestrzeń ROC pokazuje zależność pomiędzy

$$1. FPR = \frac{FP}{TN+FP} = 1 - \textit{specificity}$$

$$2. TPR = \frac{TP}{TP+FP} = \textit{sensitivity}$$

- Najlepszy klasyfikator jest w punkcie (0,1)
- Klasyfikator losowy leży na ukośnej linii
- Zły klasyfikator jest poniżej linii (ale wtedy wystarczy zanegować wyniki i będzie dobrym klasyfikatorem)



Źródło: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Krzywa ROC

Niektóre klasyfikatory poza etykietą klasy wyjściowej zwracają liczbową miarę oceny klasyfikacji, np. prawdopodobieństwo.

Ocena C	Data	Ocena CPP	Klasa	Prawdop. nie_zdal
3.0	2016-09-01	2.0	nie_zdal	0.961375
3.0	2016-09-01	3.0	nie_zdal	0.763757
3.0	2016-09-01	3.5	nie_zdal	0.538138
3.0	2016-09-01	4.0	zdal	0.295734
3.0	2016-09-01	4.5	zdal	0.131445
3.0	2016-09-01	5.0	zdal	0.051721
3.5	2016-09-01	2.0	nie_zdal	0.93792
3.5	2016-09-01	3.0	nie_zdal	0.662431
3.5	2016-09-01	3.5	zdal	0.414256
3.5	2016-09-01	4.0	zdal	0.203115
3.5	2016-09-01	4.5	zdal	0.084132
3.5	2016-09-01	5.0	zdal	0.032046
4.0	2016-09-01	2.0	nie_zdal	0.901676
4.0	2016-09-01	3.0	nie_zdal	0.543614
4.0	2016-09-01	3.5	zdal	0.300348
4.0	2016-09-01	4.0	zdal	0.133984
4.0	2016-09-01	4.5	zdal	0.052814
4.0	2016-09-01	5.0	zdal	0.019699
4.5	2016-09-01	2.0	nie_zdal	0.847709

Dla prawdopodobieństw progamiem klasyfikacji jest 0.5.

Jakie byłyby wyniki gdybyśmy zmienili próg? Na przykład przyjęli :

- 0.05
- 0.13
- 0.20
- 0.53
- 0.76
- itd

Krzywa ROC

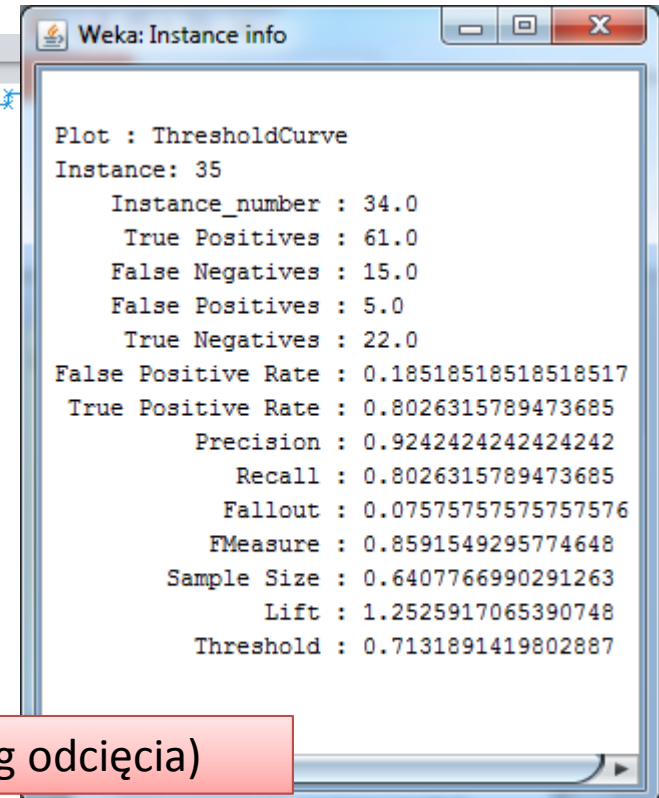
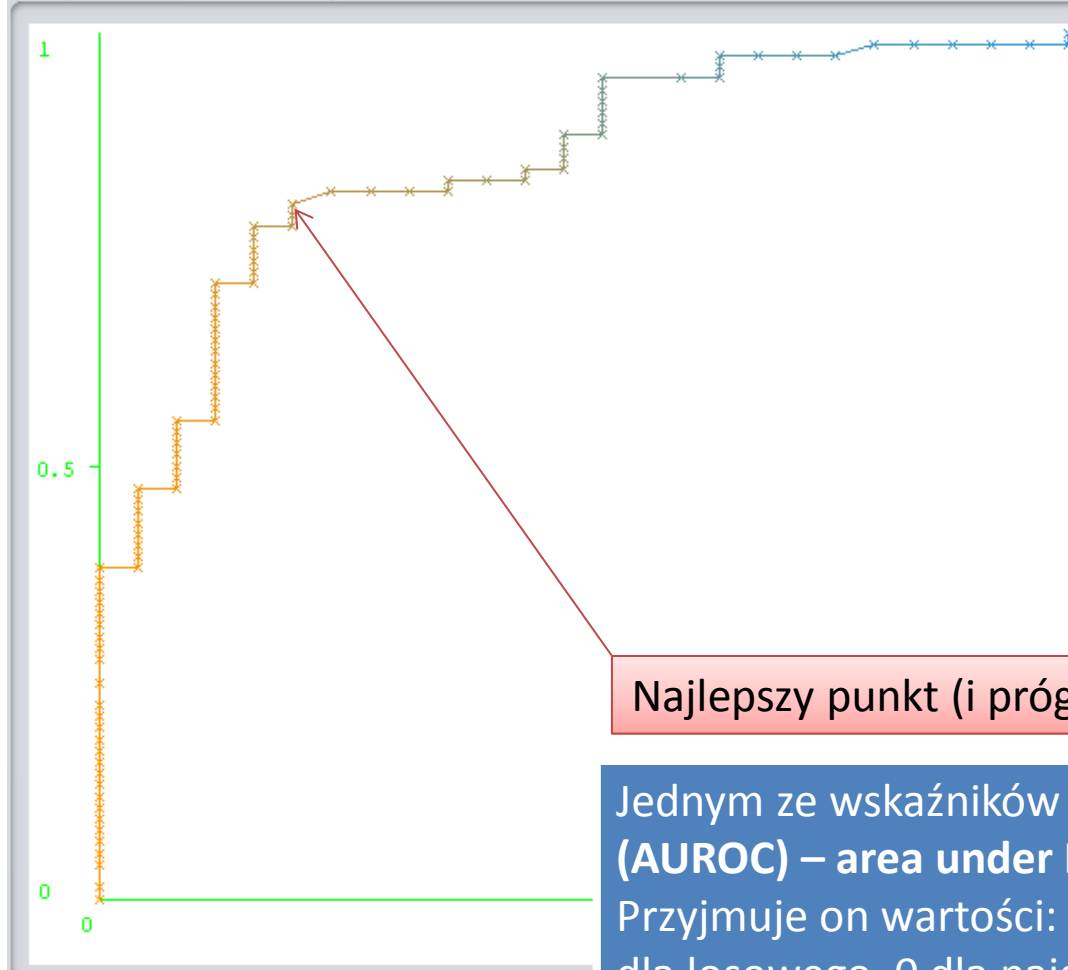
- Posortujmy wyniki klasyfikacji według prawdopodobieństw
- Przyjmijmy kolejno rosnące prawdopodobieństwa jako progi i obliczmy punkty (FPR , TPR)

Ocena C	Data	Ocena CPP	Klasa	Prawdop. nie_zdal
5.0	2016-09-01	5.0	zdal	0.007349
4.5	2016-09-01	5.0	zdal	0.012051
4.0	2016-09-01	5.0	zdal	0.019699
5.0	2016-09-01	4.5	zdal	0.02013
3.5	2016-09-01	5.0	zdal	0.032046
4.5	2016-09-01	4.5	zdal	0.032737
3.0	2016-09-01	5.0	zdal	0.051721
4.0	2016-09-01	4.5	zdal	0.052814
5.0	2016-09-01	4.0	zdal	0.053928
3.5	2016-09-01	4.5	zdal	0.084132
4.5	2016-09-01	4.0	zdal	0.085847
3.0	2016-09-01	4.5	zdal	0.131445
4.0	2016-09-01	4.0	zdal	0.133984
5.0	2016-09-01	3.5	zdal	0.136564
3.5	2016-09-01	4.0	zdal	0.203115
4.5	2016-09-01	3.5	zdal	0.206708
3.0	2016-09-01	4.0	zdal	0.295734
4.0	2016-09-01	3.5	zdal	0.300348
5.0	2016-09-01	3.0	zdal	0.305003

Krzywa ROC

Wynikiem będzie krzywa ROC

Plot (Area under ROC = 0.8596)



Najlepszy punkt (i próg odcięcia)

Jednym ze wskaźników oceny klasyfikatorów jest **AUC (AUROC)** – area under ROC, czyli pole pod krzywą ROC. Przyjmuje on wartości: 1 dla idealnego klasyfikatora, 0.5 dla losowego, 0 dla najgorszego klasyfikatora,

Dokładność (precision)

	Sklassyfikowany jako chory	Sklassyfikowany jako zdrowy
Pacjent chory	TP (true positive)	FN (false negative)
Pacjent zdrowy	FP (false positive)	TN (true negative)

Dokładność (ang. precision) zdefiniowana jest jako:

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

Miara ta pozwala określić w ilu przypadkach (procentowo) klasyfikator zwrócił poprawny wynik.

TP = 34	FN = 16
FP = 7	TN = 43

$$precision = \frac{34}{34 + 7} = 0.829$$

TP = 50	FN = 50
FP = 50	TN = 850

$$precision = \frac{50}{50 + 50} = 0.5$$

Dokładność i czułość - precision vs. recall

Dokładność (precision) i czułość (recall) często mogą zwracać rozbieżne wyniki, co świadczy o niskiej jakości klasyfikatora

TP = 50	FN = 0
FP = 950	TN = 0

$$\textit{precision} = \frac{50}{50 + 950} = 0.05 \quad \textit{recall} = \frac{50}{50 + 0} = 1.0$$

TP = 3	FN = 47
FP = 0	TN = 950

$$\textit{precision} = \frac{3}{3 + 0} = 1.0 \quad \textit{recall} = \frac{3}{3 + 47} = 0.06$$

Miara F1

Miara F_1 (lub po prostu F) jest średnią harmoniczną *precision* i *recall*

$$F_1 = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

TP = 50	FN = 0
FP = 950	TN = 0

$$\textit{precision} = 0.05 \quad \textit{recall} = 1.0 \quad F_1 = 0.05$$

TP = 3	FN = 47
FP = 0	TN = 950

$$\textit{precision} = 1.0 \quad \textit{recall} = 0.06 \quad F_1 = 0.06$$

TP = 34	FN = 16
FP = 7	TN = 43

$$\textit{precision} = 0.83 \quad \textit{recall} = 0.68 \quad F_1 = 0.74$$

Macierz pomyłek dla wielu klas

True \ predicted	c_1	c_2	c_3	
c_1	5	0	2	nieprawidłowo sklasyfikowane
c_2	1	7	2	
c_3	1	4	3	prawidłowo sklasyfikowane

$precision(c_i)$ (green arrow pointing to rows)
 $recall(c_i)$ (blue arrow pointing to columns)

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

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

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

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

Miara F1 dla wielu klas

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 $F_1 = 2 * \frac{precision * recall}{precision + recall}$

Miara F1 dla wielu klas

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 $F_1 = 2 * \frac{precision*recall}{precision+recall}$

F1 mikro vs makro

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

Micro Precision	0.6
Micro Recall	0.6
Micro F1	0.6
Macro Precision	0.596429
Macro Recall	0.593074
Macro F1	0.594746

True\predicted	c_1	c_2	c_3
c_1	5	0	2
c_2	1	497	2
c_3	1	4	3

Micro Precision	0.980583
Micro Recall	0.980583
Micro F1	0.980583
Macro Precision	0.694429
Macro Recall	0.711624
Macro F1	0.702921

Silne niezrównoważenie

Współczynnik Kappa

Współczynnik Kappa Cohena pozwala określić stopień zgodności etykiet w zbiorze uczącym oraz zwróconych przez klasyfikator.

Wyobraźmy sobie, że mamy dwóch „oceniających” opisanych dwiema funkcjami:

$$\begin{aligned} f_1(x): X &\rightarrow \{c_1, \dots, c_k\} \\ f_2(x): X &\rightarrow \{c_1, \dots, c_k\} \end{aligned}$$

- Pierwszym jest prawdziwe przypisanie etykiet (*ground truth*)
- Drugim jest zbudowany klasyfikator

Zgodność ocen może mieć dwa źródła:

1. Rzeczywiście obaj „oceniający” uważają, że obserwacji powinna być przypisana ta sama etykieta
2. Zbieżność ma charakter losowy, czyli pierwszy oceniający wybiera etykieta c_i dla obserwacji x z prawdopodobieństwem $p_1(f_1(x) = c_i|x)$, drugi z prawdopodobieństwem $p_2(f_2(x) = c_i|x)$. Oba zdarzenia wyboru są niezależne, więc

$$p(f_1(x) = f_2(x) = c_i|x) = p_1(f_1(x) = c_i|x) \cdot p_2(f_2(x) = c_i|x)$$

Współczynnik Kappa

Współczynnik Kappa jest obliczany jako:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

p_0 – to zaobserwowane prawdopodobieństwo zgodności

p_e – to korekta przypadkowej zgodności dla niezależnych losowych wyborów etykiet

Dla macierzy pomyłek E oznaczmy przez $m = \sum_{i=1}^k \sum_{j=1}^k e[i, k]$

$$p_0 = \frac{\sum_{i=1}^k e[i, i]}{m} = \textit{accuracy}$$

$$p_e = \sum_{i=1}^k \left(\frac{\sum_{j=1}^k e[i, j]}{m} \cdot \frac{\sum_{j=1}^k e[j, i]}{m} \right)$$

Współczynnik Kappa - przykład

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



Dzielimy przez sumę elementów
(zamieniamy na
prawdopodobieństwa)

	c_1	c_2	c_3
c_1	0.2	0	0.08
c_2	0.04	0.28	0.08
c_3	0.04	0.16	0.12

Współczynnik Kappa - przykład

Wyznaczamy $P_1(c_i)$ i $P_2(c_i)$ jako sumy elementów w wierszach i kolumnach. Czyli np. $P_2(c_1)$ opisuje prawdopodobieństwo, że klasyfikator przypisze etykietę c_1 .

	c_1	c_2	c_3	
c_1	0.2	0	0.08	$P_1(c_1) = 0.28$
c_2	0.04	0.28	0.08	$P_1(c_2) = 0.40$
c_3	0.04	0.16	0.12	$P_1(c_3) = 0.32$
	$P_2(c_1) = 0.28$	$P_2(c_2) = 0.44$	$P_2(c_3) = 0.28$	

$$p_0 = \sum_{i=1}^k p_{ii} = 0.2 + 0.28 + 0.12 = 0.6$$

$$p_e = \sum_{i=1}^k p_1(c_i) \cdot p_2(c_i) = 0.28 \cdot 0.28 + 0.4 \cdot 0.44 + 0.32 \cdot 0.28 = 0.344$$

Stąd $\kappa = 0.39$

Współczynnik Kappa - interpretacja

True \ predicted	c_1	c_2	c_3
c_1	5	0	2
c_2	1	497	2
c_3	1	4	3

$$\kappa = 0.65$$

Ocena zgodności dla różnych wartości κ

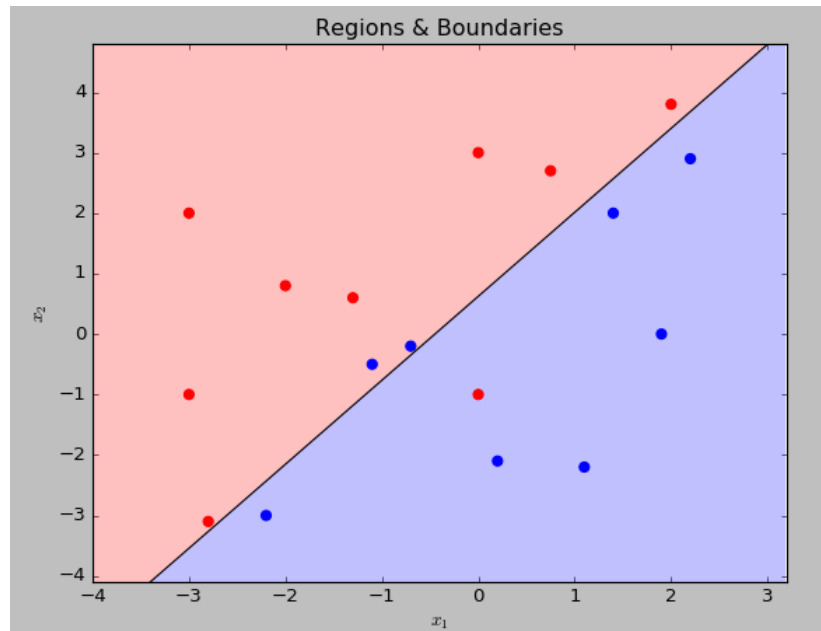
- $(0, 0.2)$ – nieistotna
- $(0.21, 0.40)$ – słaba
- $(0.41, 0.60)$ – średnia
- $(0.61, 0.80)$ – zasadnicza
- $(0.81, 1.0)$ -- doskonała zgodność

Regresja logistyczna

Regresja logistyczna

Regresja logistyczna jest zagadnieniem klasyfikacji binarnej:

- Wartości x są wektorami n -wymiarowymi: $x \in \mathbb{R}^n$
- Zbiór decyzji (klas) Y zawiera dwa elementy, np. $Y = \{0,1\}$
- Granica klas jest $n-1$ wymiarową hiperpłaszczyzną, czyli np. prostą w przypadku 2D



Regresja logistyczna: hiperpłaszczyzna

- Równanie hiperpłaszczyzny (x_i oznacza i -ty element wektora):

$$w_0 + w_1x_1 + \dots + w_nx_n = 0$$

- Wektor normalny: $z = [w_1, \dots, w_n]$

- Norma z (długość): $\|z\| = \sqrt{\sum w_i^2}$

- Odległość punktu x_k od hiperpłaszczyzny:

$$d(x_k) = \frac{|w_0 + \sum_{i=1}^n w_i x_{ki}|}{\|z\|}$$

- Przekształcając X do R^{n+1} : $x \rightarrow [1: x]$

możemy zapisać:

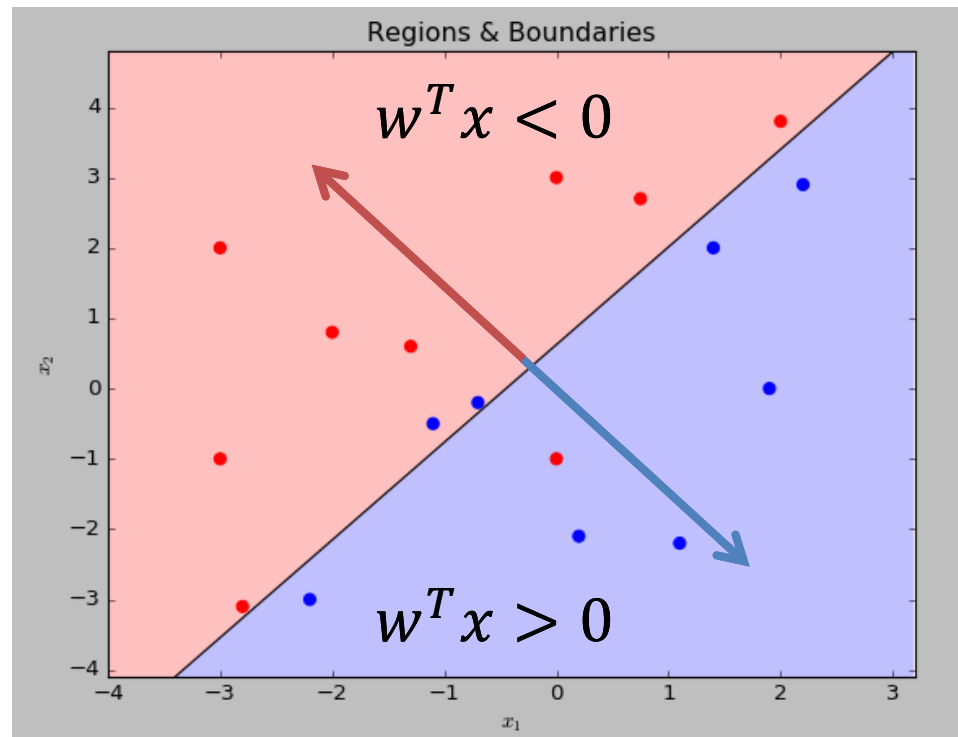
– równanie hiperpłaszczyzny: $w^T x = 0$

– odległość punktu od hiperpłaszczyzny: $d(x_k) = \frac{|w^T x_k|}{\|z\|}$

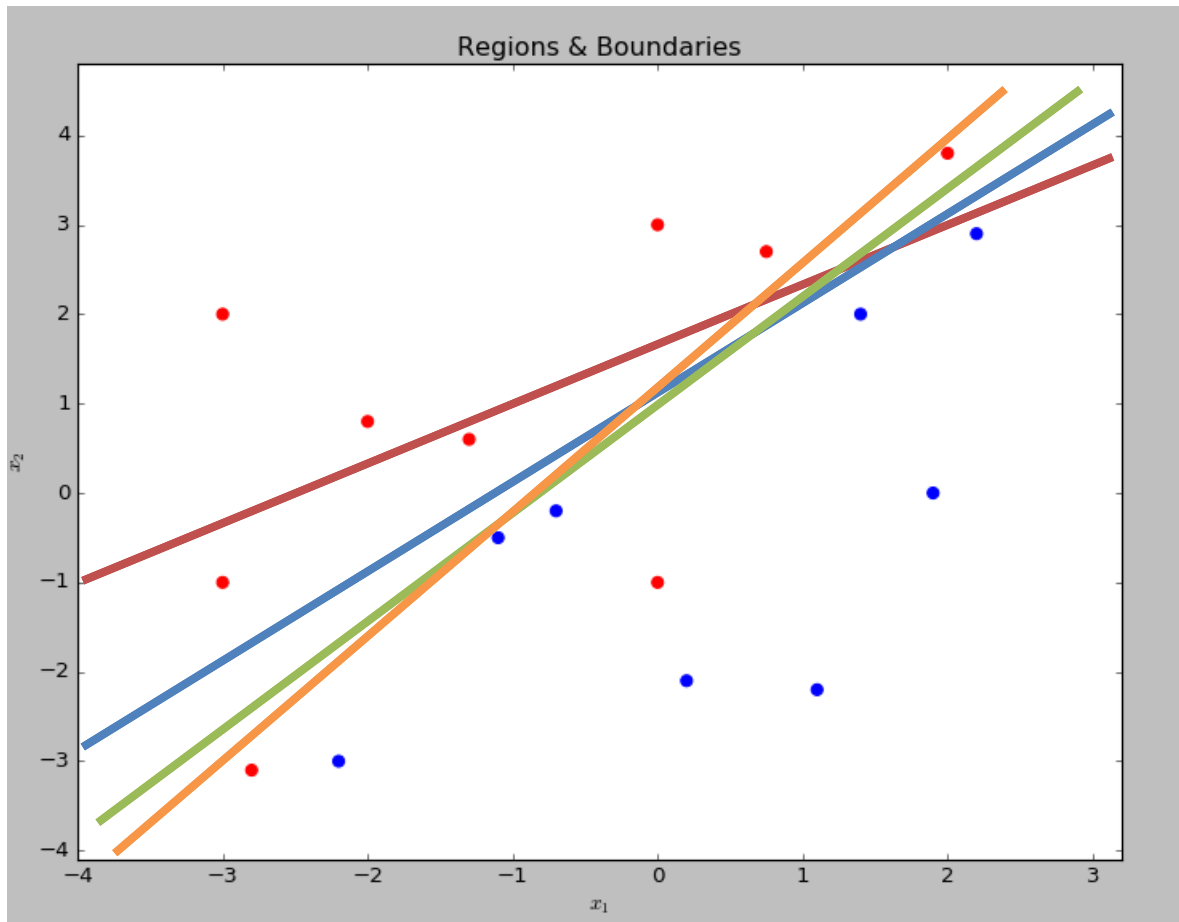
Regresja logistyczna: funkcja klasyfikacji

Najprostszą formą wyboru funkcji klasyfikacji $c(x)$ jest przypisanie klasy na podstawie wartości $w^T x$

$$c(x) = \begin{cases} 0 & \text{gdy } w^T x < 0 \\ 1 & \text{gdy } w^T x \geq 0 \end{cases}$$



Jak dobrać granicę klas?



4 sklasyfikowane
błędnie

2 sklasyfikowane
błędnie

2 sklasyfikowane
błędnie

1 sklasyfikowany
błędnie

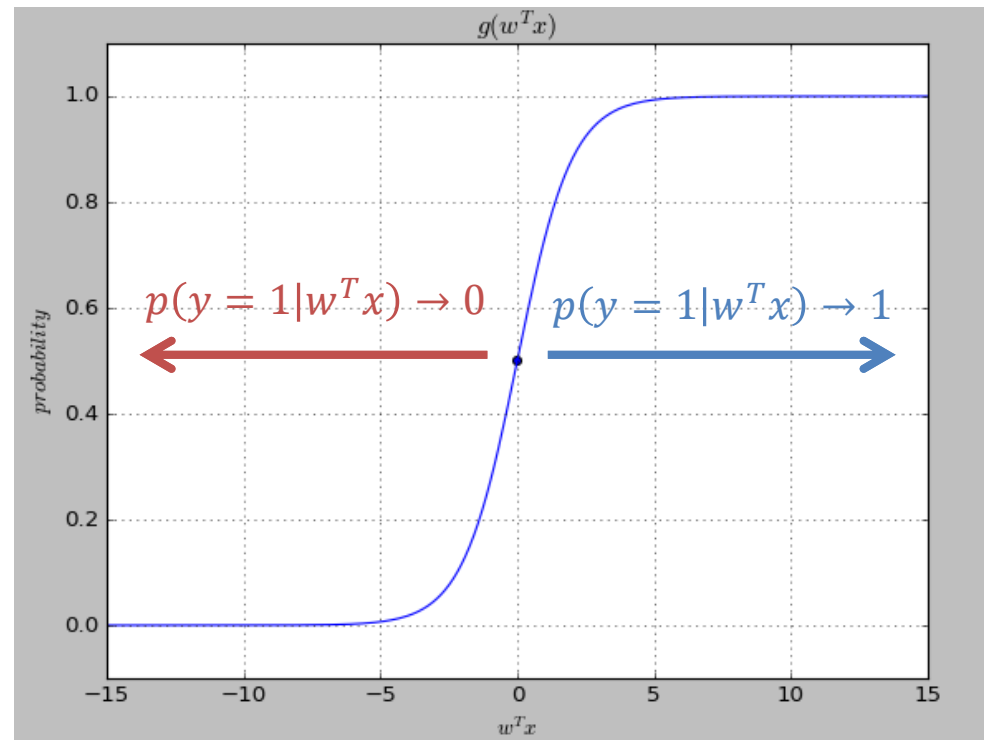
- Może warto wziąć pod uwagę odległość od prostej?
- Jeżeli prosta leży blisko punktów, wówczas model wykazuje wrażliwość na drobne zmiany w ich położeniu.

Regresja logistyczna: model

W modelu regresji logistycznej prawdopodobieństwo $p(c|x)$ jest wyznaczone bezpośrednio.

Ogólna idea:

- dla x w pobliżu granicy klas ($w^T x = 0$) prawdopodobieństwa dla obu klas wynoszą w przybliżeniu $\frac{1}{2}$
- wraz z oddalaniem się od granicy klas: $|w^T x| \rightarrow \infty$ prawdopodobieństwo $p(c_1|x)$ jednej z klas rośnie do 1, a drugiej maleje do 0.



Uogólniony model liniowy

Regresja logistyczna jest jednym z przypadków **uogólnionego modelu liniowego** (GLM, generalized linear model):

$$L(E[y|x]) = w^T x$$

- $E[y]$ – wartość oczekiwana zmiennej wyjściowej
- $L(x)$ - funkcja wiążąca (ang. link function)
- GLM stosuje się w przypadku, kiedy np. zmienna wyjściowa ma wartości dyskretne lub należy zamodelować inne rozkłady statystyczne niż normalny
- Zakładając, że mamy dwie klasy oznaczone przez 1 i 0, wartość oczekiwana wynosi:

$$\begin{aligned} E[y|x] &= \sum_y y p(y|x) = 1 \cdot p(y = 1|x) + 0 \cdot p(y = 0|x) \\ &= p(y = 1|x) \end{aligned}$$

- Dla binarnych problemów klasyfikacji wystarczające jest aproksymowane prawdopodobieństwo wystąpienia wybranej klasy:

$$p(y = 0 | x) = 1 - p(y = 1 | x)$$

Szansa

- Rozważmy wyrażenie:

$$odds = \frac{p}{1 - p}$$

- Jest to iloraz prawdopodobieństwa wystąpienia zdarzenia do zdarzenia przeciwnego,

np. jeśli prawdopodobieństwo wygranej kandydata w wyborach wynosi 0.8, to szansa $odds = \frac{0.8}{0.2} = 4:1$

- Szansa przypisania x etykiety klasy 1:

$$odds(y = 1|x) = \frac{p(y = 1|x)}{1 - p(y = 1|x)}$$

- Szansa $odds$ przybiera wartości z przedziału $[0, \infty]$

Logit

Rozważamy uogólniony model liniowy $L(E[y|x]) = L(p) = w^T x$,
gdzie L – funkcja wiążąca (ang. link function)

- Załóżmy, że funkcją wiążącą jest funkcja $logit(p)$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

- Dla $p \in [0,1]$, wartości $logit(p) \in [-\infty, \infty]$.
Funkcja $logit(p)$ pozwala więc na powiązanie zakresu wartości prawdopodobieństw $[0,1]$ z zakresem zmienności modelu liniowego $w^T x \in [-\infty, \infty]$
- Równanie uogólnionego modelu liniowego dla regresji logistycznej ma postać:

$$logit(p(y = 1|x)) = \ln\left(\frac{p(y = 1|x)}{1 - p(y = 1|x)}\right) = w^T x$$

Interpretacja logit

Zestawienie wyników egzaminu

- *wykl* – na ilu wykładach student był obecny
- *zal* – ocena z zaliczenia (min 3.0)
- *zdal* – czy po przystąpieniu zdał egzamin w pierwszym terminie (1/0)

Współczynniki regresji:

$$\text{logit}(p) = -11.63 + 0.24 * \text{wykl} + 2.70 * \text{zal}$$

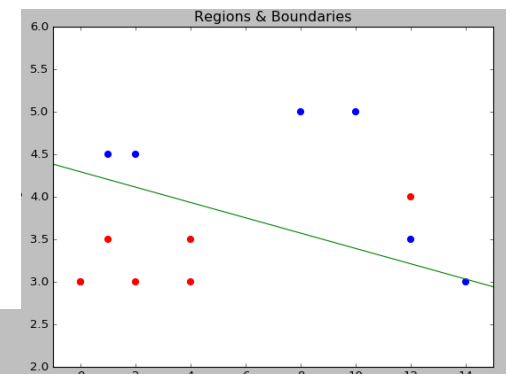
Dodatkowa obecność na jednym wykładzie:

- zwiększa $\text{logit}(p)$ o 0.24
- zwiększa szanse zdania
razy $e^{0.24} = 1.27$ (czyli o 27%)

Ocena z zaliczenie o stopień wyższa:

- zwiększa $\text{logit}(p)$ o 2.7
- zwiększa szanse zdania
razy $e^{2.7} = 14.87$ (czyli o 1387%)

wykl	zal	zdal
14	3.0	1
12	3.5	1
12	4.0	0
10	5.0	1
2	4.5	1
4	3.5	0
0	3.0	0
1	4.5	1
2	3.0	0
8	5.0	1
4	3.0	0
0	3.0	0
1	3.5	0



Prawdopodobieństwo $p(y|x)$

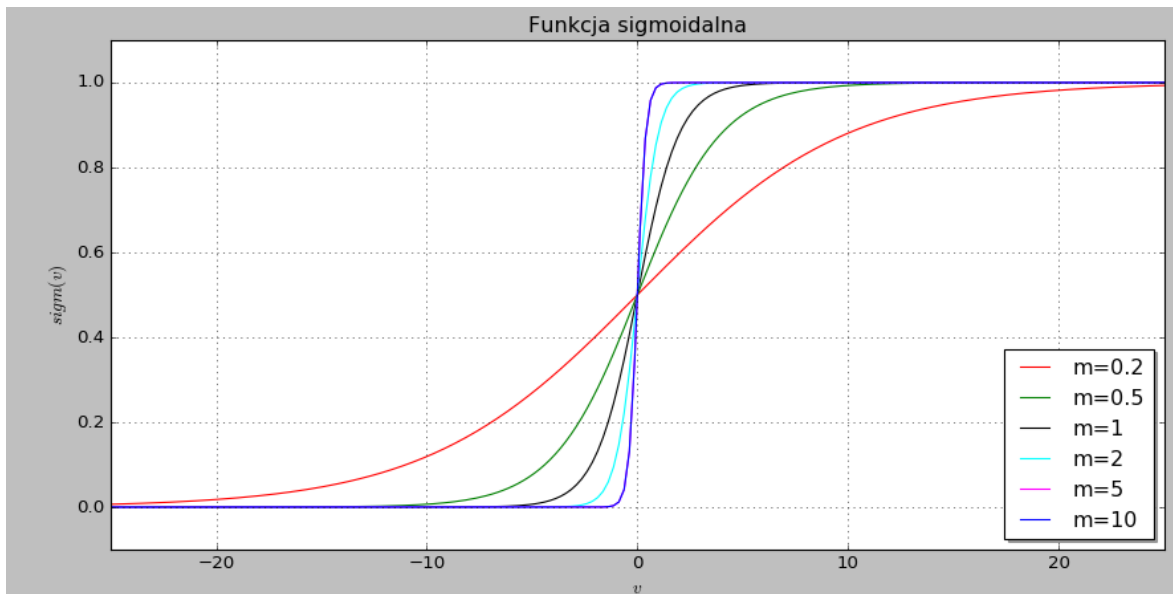
Przekształcając równanie:

$$\ln \left(\frac{p(y = 1|x)}{1 - p(y = 1|x)} \right) = w^T x$$

otrzymujemy równoważną postać:

$$p(y = 1|x) = \frac{1}{1 + \exp(-w^T x)} = \frac{\exp(w^T x)}{1 + \exp(w^T x)}$$

W równaniu występuje funkcja sigmoidalna: $\text{sigm}(v) = \frac{1}{1 + \exp(-mv)}$

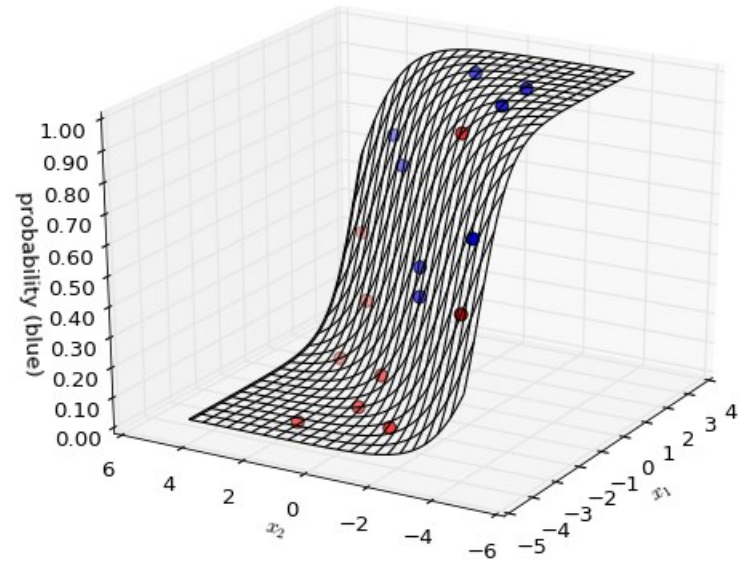
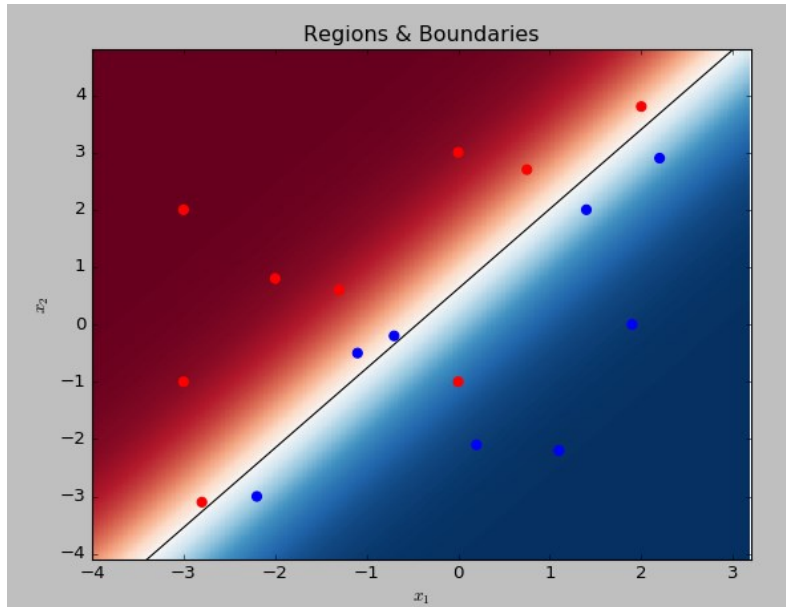


Wektor w jest celowo **nieznormalizowany**.

- Małe wartości wag (małe m) powodują, że krzywa $p(y = 1|x)$ jest płaska
- Duże wartości – krzywa coraz bardziej przypomina funkcję skokową

Przykład

$$w^T x = 0.748 + 1.665 x_1 - 1.198 x_2$$



- Po lewej: obserwacje i granica klas:
 - biały obszar $p(y = blue|x) \approx p(y = red|x) \approx 0.5$
 - brązowy: $p(y = red|x) \rightarrow 1$
 - granatowy: $p(y = blue|x) \rightarrow 1$
- Po prawej: wykres gęstości prawdopodobieństwa $p(y = blue|x)$

Dobór parametrów hiperpłaszczyzny

- Dobór parametrów hiperpłaszczyzny jest problemem optymalizacyjnym:
 - definiujemy funkcję kosztu
 - przeprowadzamy optymalizację
- Dany jest zbiór uczący: $D = \{(x_i, y_i)\}_{i=1,m}$
- **Dla jakich parametrów regresji w obserwacje ze zbioru D są najbardziej prawdopodobne?**
 - $p(y_i = 1|w, x_i)$ – określone na podstawie poszukiwanego modelu prawdopodobieństwo, że $y_i = 1$
 - $p(y_i = 0|w, x_i)$ – prawdopodobieństwo, że $y_i = 0$
- Niezależnie od wartości y_i prawdopodobieństwo wystąpienia tej wartości dla x_i wynosi:
$$p(y_i = 1|w, x_i)^{y_i} \cdot p(y_i = 0|w, x_i)^{(1-y_i)}$$
$$= p(y_i = 1|w, x_i)^{y_i} \cdot (1 - p(y_i = 1|w, x_i))^{(1-y_i)}$$
 - Jeżeli $y_i = 1$, to **drugi czynnik** ma wartość $p^0 = 1$
 - Jeżeli $y_i = 0$, to **pierwszy czynnik** ma wartość 1

Estymacja największej wiarygodności

- Prawdopodobieństwo wystąpienia zbioru obserwacji:

$$L(w) = \prod_{i=1}^m p(y_i = 1|w, x_i)^{y_i} \cdot (1 - p(y_i = 1|w, x_i))^{(1-y_i)}$$

Estymacja największej wiarygodności (ang. MLE, Maximum likelihood estimation)

Wybierane są parametry w^* , dla których prawdopodobieństwo zaobserwowania zbioru uczącego $L(w)$ przybiera wartość maksymalną:

$$w^* = \arg \max L(w)$$

Podstawiając:

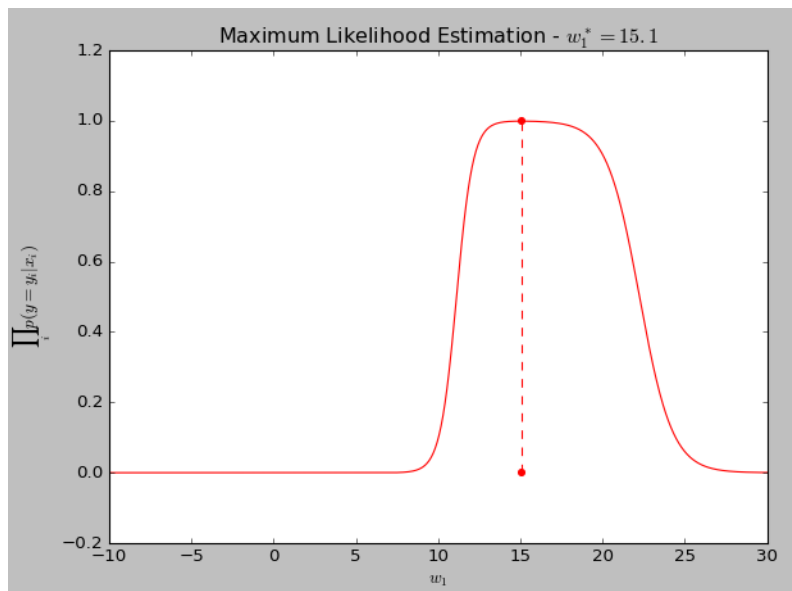
$$p(y_i = 1|x_i, w) = \frac{1}{1 + \exp(-w^T x_i)}$$

otrzymujemy:

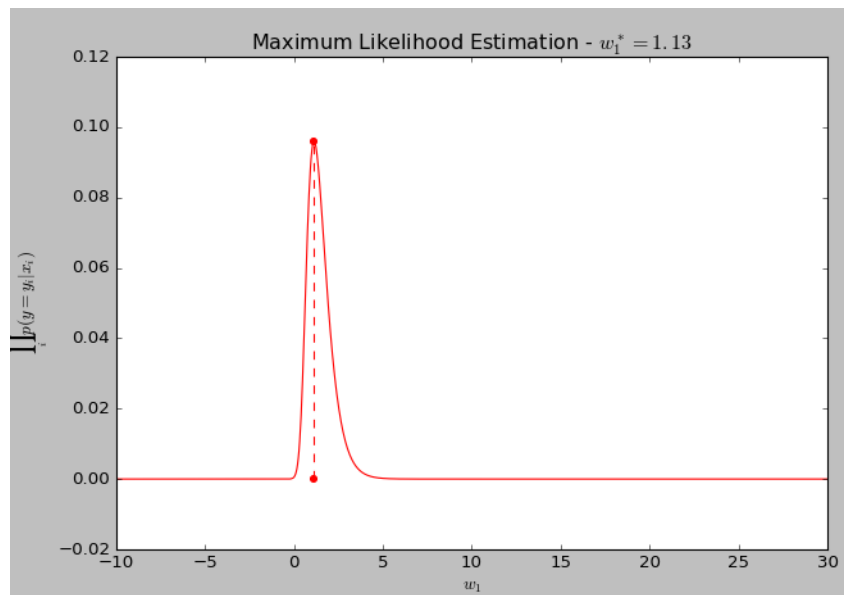
$$L(w) = \prod_{i=1}^m \left(\frac{1}{1 + \exp(-w^T x_i)} \right)^{y_i} \cdot \left(\frac{\exp(-w^T x_i)}{1 + \exp(-w^T x_i)} \right)^{(1-y_i)}$$

Przebieg $L(w)$

x	-4	-3	-2	-1	0	1
y	0	0	0	1	1	1



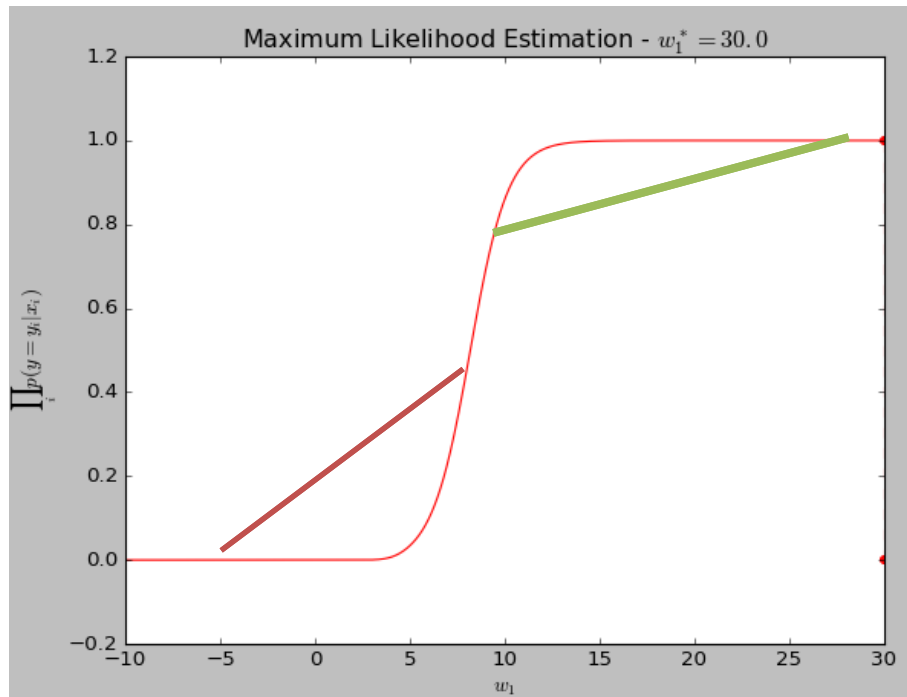
x	-3	-2	-1	1	2	3	4	5
y	0	0	0	1	0	1	1	1



- W jednowymiarowym zagadnieniu powinny być optymalizowane dwa parametry: w_0 i w_1 . Wykresy sporządzono dla optymalnej wartości w_0^*
- Po lewej – dane są liniowo separowalne, więc prawdopodobieństwo $L(w)$ osiąga wartość 1
- Po prawej – nie są i musi być przyjęty kompromis. Stąd $\max L(w) \approx 0.1$

Przebieg $L(w)$

x	-4	-3	-2	-1	0	1
y	0	0	0	0	1	1



- Szczególny przebieg – granica klas dla wartości 0
- Praktycznie wszystkie wartości w_1 powyżej 15 są dobre

- Funkcja $L(w)$ jest na ogół funkcją wklęsłą
- Jej maksimum istnieje lub można wyznaczyć wartości bliskie optymalnym.
- Brak rozwiązania analitycznego

Logarytm $L(w)$

- Dla dużych m iloczyn prawdopodobieństw wchodzących w skład $L(w)$ byłby bardzo mały – problem niestabilny numerycznie
- Z tego powodu wyznaczany jest logarytm $L(w)$ - **iloczyn czynników** zastąpiony jest **sumą ich logarytmów**
- Dla jednego punktu:

$$ll_i(w) = \ln \left(\left(\frac{1}{1 + \exp(-w^T x_i)} \right)^{y_i} \cdot \left(\frac{\exp(-w^T x_i)}{1 + \exp(-w^T x_i)} \right)^{(1-y_i)} \right)$$

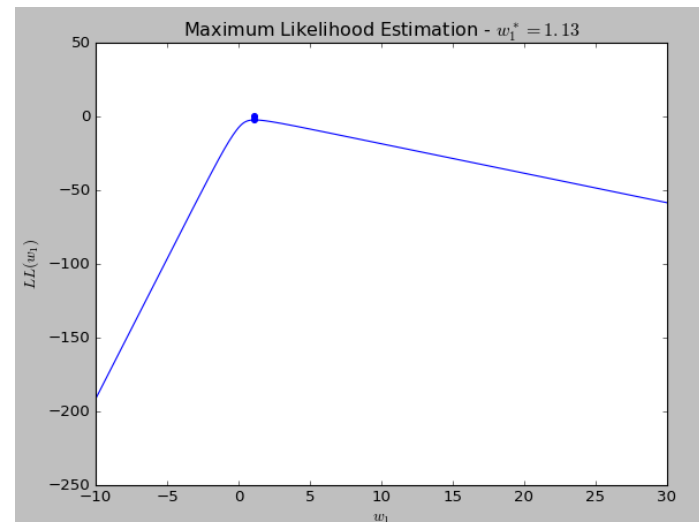
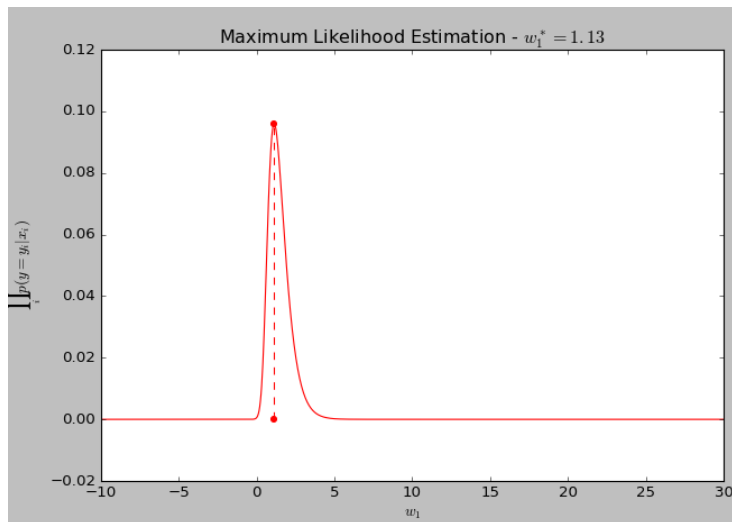
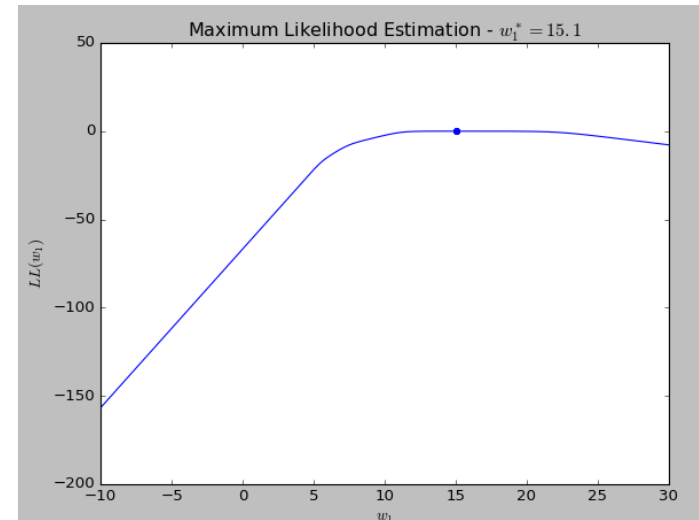
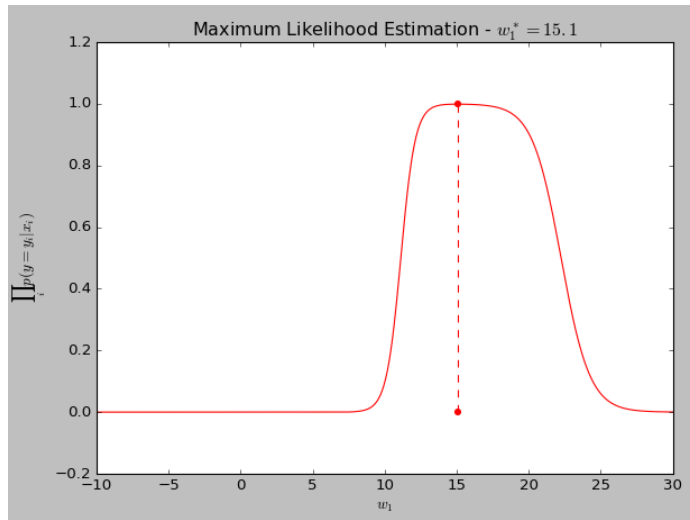
$$ll_i(w) = \ln(1^{y_i}) - y_i \ln(1 + \exp(-w^T x_i)) \\ - (1 - y_i)w^T x_i - (1 - y_i)\ln(1 + \exp(-w^T x_i))$$

Stąd:

$$ll_i(w) = -(1 - y_i)w^T x_i - \ln(1 + \exp(-w^T x_i))$$

$$LL(w) = \sum_{i=1}^m ll_i(w)$$

Porównanie przebiegów $L(w)$ i $LL(w)$



Gradient LL(w)

- Miejsce maksimum LL(w) wyznaczone jest za pomocą metod gradientowych (brak rozwiązania analitycznego)
- Dla pojedynczego punktu x_i :

$$ll_i(w) = -(1 - y_i)w^T x_i - \ln(1 + \exp(-w^T x_i))$$

$$\frac{d}{dw_j} ll_i(w) = x_{ij}y_i - x_{ij} - x_{ij} \frac{\exp(-w^T x_i)}{1 + \exp(-w^T x_i)}$$

$$\frac{d}{dw_j} ll_i(w) = x_{ij} \left(y_i - 1 - \frac{\exp(-w^T x_i)}{1 + \exp(-w^T x_i)} \right)$$

$$\frac{d}{dw_j} ll_i(w) = x_{ij} \left(y_i - \frac{1}{1 + \exp(-w^T x_i)} \right)$$

$$\frac{d}{dw_j} ll_i(w) = x_{ij}(y_i - p(y_i = 1|x_i, w))$$

- Jeżeli y_i i $p(y_i = 1|x_i, w)$ są zgodne: wkład punktu x_i do gradientu wynosi 0
- W przeciwnym przypadku gradient w kierunku w_j jest proporcjonalny do j -tej składowej wektora x_i

Gradient LL(w)

- Podsumowując:

$$\frac{d}{dw_j} LL(w) = \sum_{i=1}^m x_{ij} (y_i - p(y_i = 1 | x_i, w))$$

Dla każdego wymiaru

Suma po obserwacjach

$$\nabla LL(w) = \left[\frac{d}{dw_0} LL(w), \frac{d}{dw_1} LL(w), \dots, \frac{d}{dw_n} LL(w), \right]$$

- Algorytm gradientu prostego (gradient ascent) ma postać:
wybierz wartości początkowe $w(0)$
while !stop():
 $w(t + 1) = w(t) + \eta \nabla LL(w)$
- Zazwyczaj η jest małą stałą lub maleje wraz z numerem iteracji t
- Warunek końca: liczba iteracji, $|\nabla LL(w)| \approx 0$ lub brak poprawy funkcji celu

Cechy

- W bardziej ogólnym przypadku, zamiast wektorów $x \in \mathbb{R}^n$ rozważany jest wektor **cech** (ang. features)

$$h(x) = [h_1(x), \dots, h_N(x)]$$

- Równanie regresji liniowej przybiera postać:

$$\ln \left(\frac{p(y = 1|x)}{1 - p(y = 1|x)} \right) = w^T h(x)$$

$$p(y_i = 1|x_i, w) = \frac{1}{1 + \exp(-w^T h(x_i))}$$

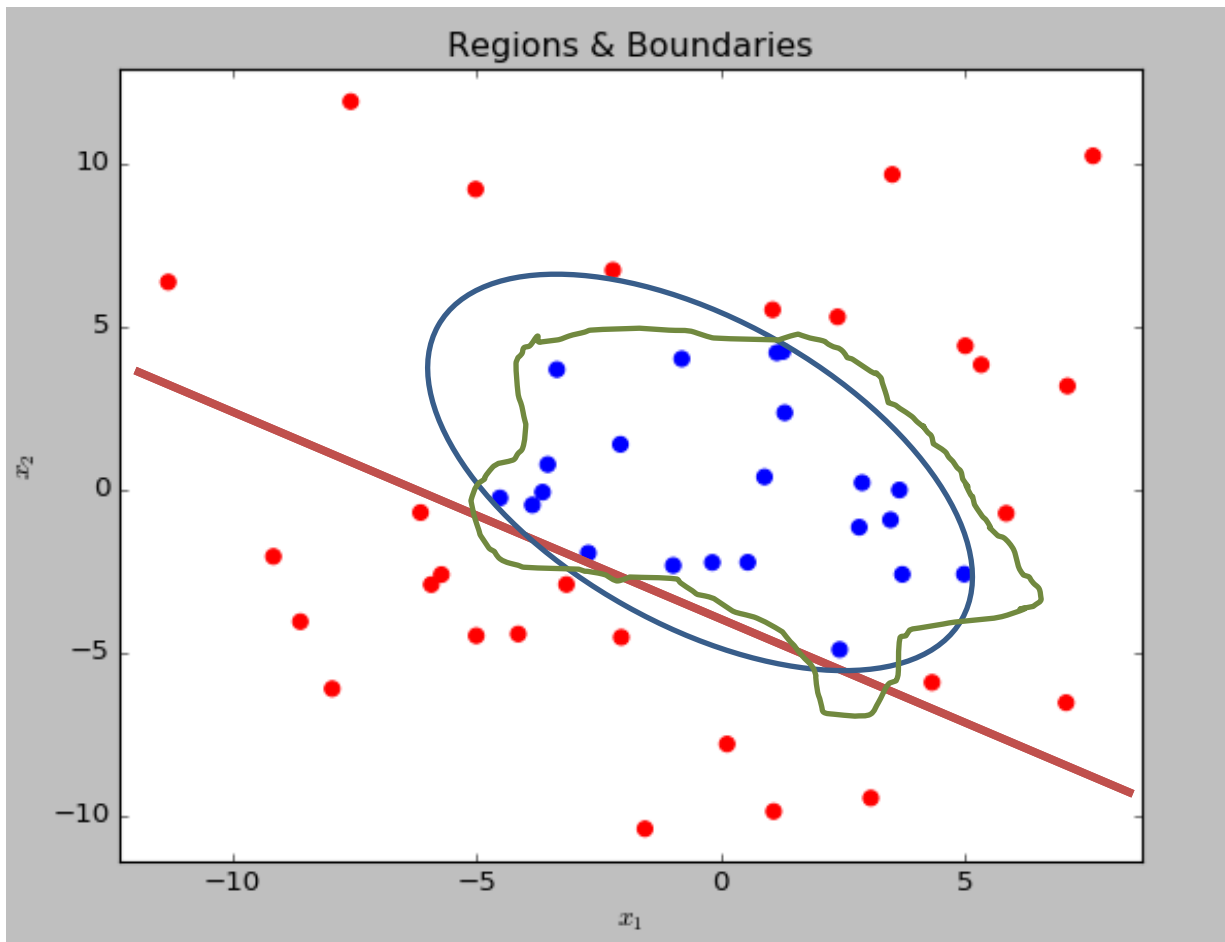
- Po uwzględnieniu cech, wzór służący do obliczenia pochodnej względem w_j przyjmuje postać:

$$\frac{d}{dw_j} LL(w) = \sum_{i=1}^m h_j(x_i) (y_i - p(y_i = 1|x_i, w))$$

- Dzięki zastosowaniu transformacji danych wejściowych do postaci cech możliwe jest uzyskanie bardziej złożonych przebiegów granic klas

Przykład

Błękitne punkty (klasa $y=1$) otoczone są punktami czerwonymi ($y=0$).



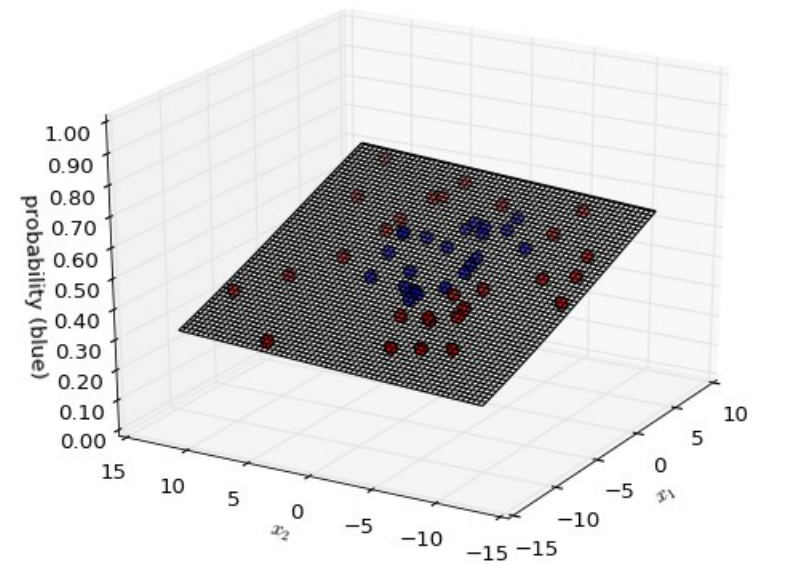
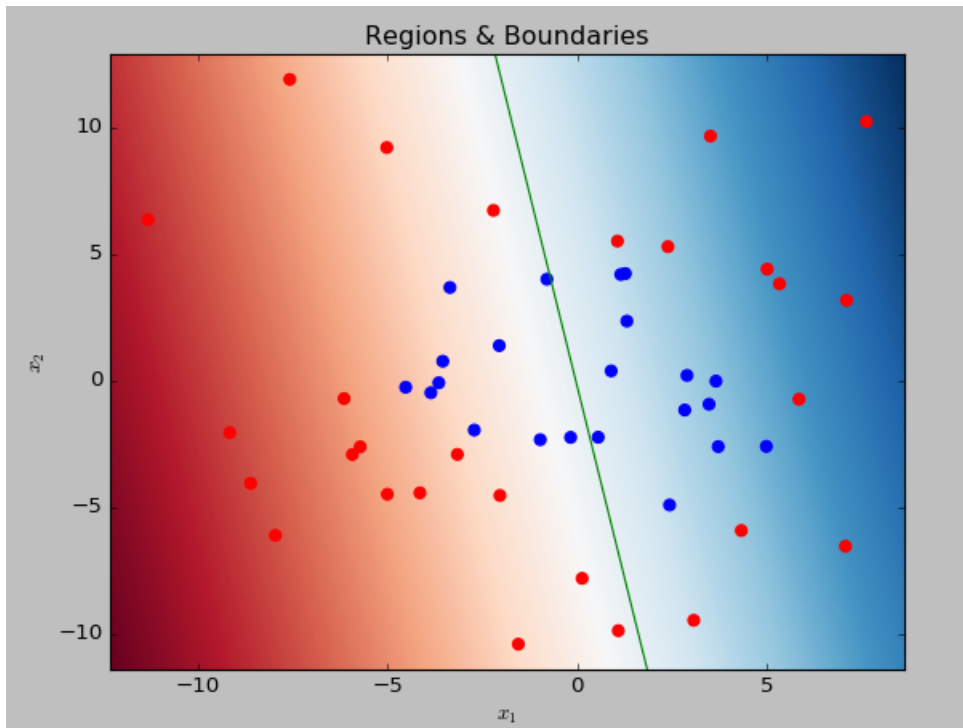
Prawdopodobnie źle?

Pożądany kształt

Raczej overfitting

Przykład

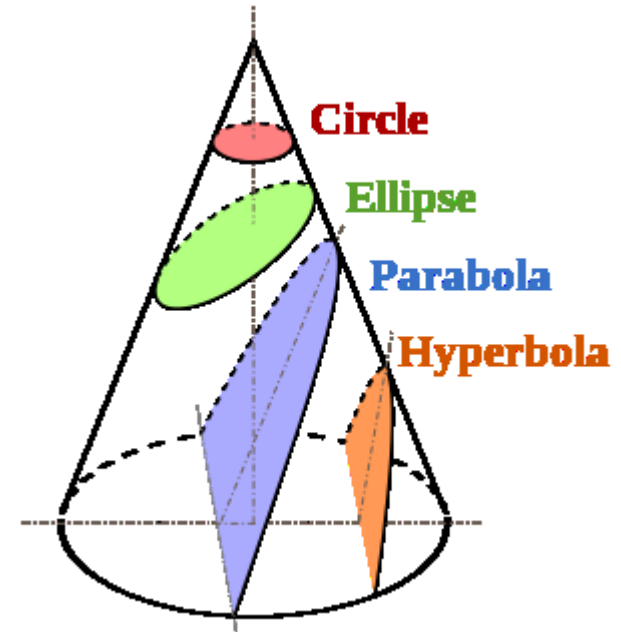
- Wybór prostej jako granicy klas nie daje dobrych rezultatów
- Brak wyraźnego przejścia w funkcji gęstości prawdopodobieństwa



Przykład

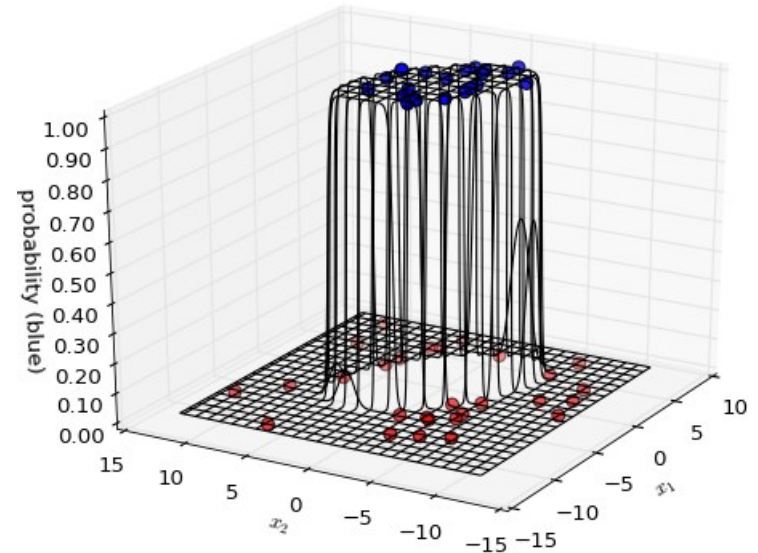
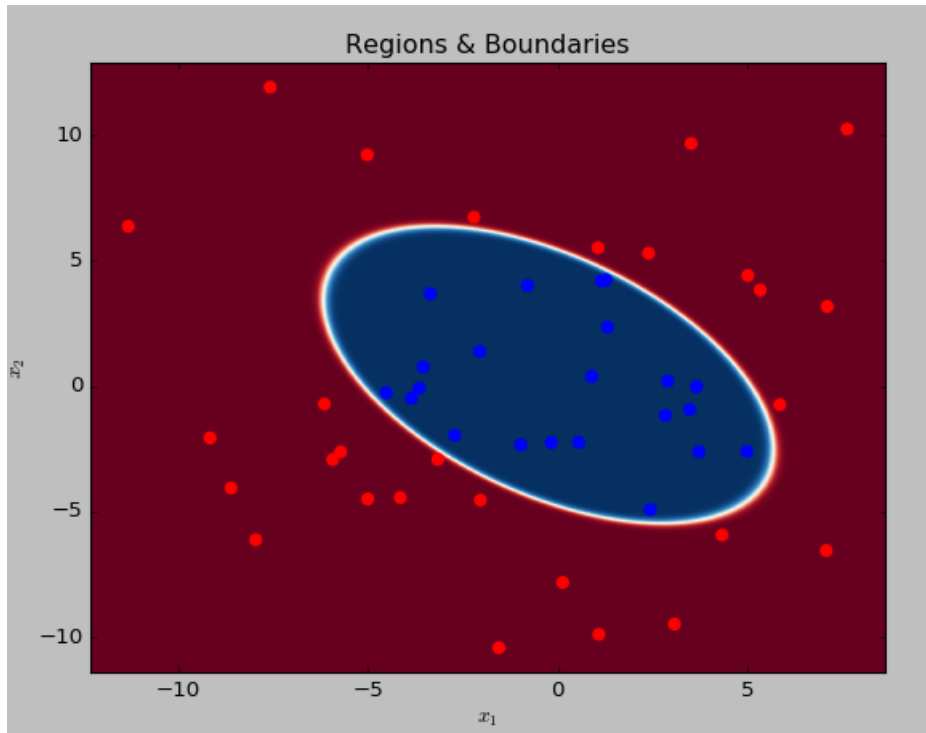
Definiujemy cechy pochodne

- $h_0(x) = 1$ (bias)
 - $h_1(x) = x_1$ (czytaj $x[1]$)
 - $h_2(x) = x_2$
 - $h_3(x) = x_1^2$
 - $h_4(x) = x_2^2$
 - $h_5(x) = x_1x_2$
- Zestaw jest wystarczający, aby zdefiniować region decyzyjny w postaci krzywej stożkowej:
$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$
 - W przypadku elipsy współczynnik B odpowiada za obrót



[Rysunek: https://en.wikipedia.org/wiki/Conic_section]

Wyniki



- Wektor wag: $w = [38.35, 0.05, 1.09, -1.47, -1.48, -1.48]$ – obecne współczynniki wyższego stopnia, w tym **obrót**
- Bardzo strome przejścia gęstości prawdopodobieństwa na granicy klas (dobre dopasowanie modelu)
- Nie zapominajmy jednak, że celem są dobre własności generalizacji

Konwersje danych do postaci numerycznej

Zarówno regresja zwykła, jak i logistyczna używa wyłącznie danych (cech) numerycznych. W zbiorze źródłowym mogą występować zmienne kategoriyczne i tekstowe, które muszą zostać przekonwertowane do postaci cech numerycznych.

- Dane kategoriyczne **binarne** można zastąpić wartościami ze zbioru $\{0,1\}$
 $\{YES, NO\} \rightarrow \{0,1\}$
- Dane kategoriyczne **porządkowe** o k wartościach można przekonwertować na zbiór liczb całkowitych z zakresu $0 \div k - 1$:
 $\{low, medium, high\}: low \rightarrow 0, medium \rightarrow 1, high \rightarrow 2$
- Dane **nominalne** o k wartościach zamienia się na k -zmiennych o wartościach 0/1.

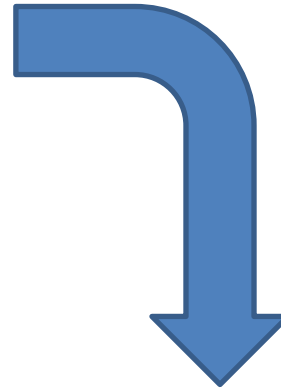
Konwersja ta nazywa się **one-hot** (jedna linia danych „gorąca”)

Color=blue		
red	green	blue
0	0	1

Przykład

Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no



Weka zapewnia odpowiednie filtry, np..
RenameNominalValues

Relation: weather-weka.filters.unsupervised.attribute.RenameNominalValues-R4-NTRUE:1,FALSE:0-weka.filters.un

No.	1: outlook=sunny Numeric	2: outlook=overcast Numeric	3: outlook=rainy Numeric	4: temperature Numeric	5: humidity Numeric	6: windy Nominal	7: play Nominal
1	1.0	0.0	0.0	85.0	85.0	0	no
2	1.0	0.0	0.0	80.0	90.0	1	no
3	0.0	1.0	0.0	83.0	86.0	0	yes
4	0.0	0.0	1.0	70.0	96.0	0	yes
5	0.0	0.0	1.0	68.0	80.0	0	yes
6	0.0	0.0	1.0	65.0	70.0	1	no
7	0.0	1.0	0.0	64.0	65.0	1	yes
8	1.0	0.0	0.0	72.0	95.0	0	no
9	1.0	0.0	0.0	69.0	70.0	0	yes
10	0.0	0.0	1.0	75.0	80.0	0	yes
11	1.0	0.0	0.0	75.0	70.0	1	yes
12	0.0	1.0	0.0	72.0	90.0	1	yes
13	0.0	1.0	0.0	81.0	75.0	0	yes
14	0.0	0.0	1.0	71.0	91.0	1	no

Konwersje danych tekstowych

- Dla danych tekstowych typową reprezentacją jest **bag-of-words**, czyli przypisanie *string* \rightarrow N



Zawody lekkoatletyczne odbyły się w Krakowie



W zawodach Polak zajął drugie miejsce

- Następnie tworzone są wektory słów – ich elementami są liczby wystąpień słów w dokumencie (czasem znormalizowane lub pomnożone przez wagi, np. reprezentujące znaczenie słów TF-IDF)
- Na ogół wektory te są bardzo rzadkie
- Jest to również konwersja typu **one-hot**: każde słowo staje się atrybutem

Przykład - Titanic

- W zbiorze danych **Titanic** występuje pole tekstowe name (nazwisko, imiona, forma grzecznościowa, itd.)
- Po konwersji filtrem Weka: **StringToWordVector** pojawiło się 1950 dodatkowych atrybutów, w tym *a, the, von, van, of, Mr, Mrs*, itd.

No.	1: pclass Numeric	2: survived Numeric	3: name String	4: sex Nominal	5: age Numeric
1	1.0	1.0	Allen, Miss. Elisabeth Walton	female	29.0
2	1.0	1.0	Allison, Master. Hudson Trevor	male	0.9167
3	1.0	0.0	Allison, Miss. Helen Loraine	female	2.0
4	1.0	0.0	Allison, Mr. Hudson Joshua Creighton	male	30.0
5	1.0	0.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0
6	1.0	1.0	Anderson, Mr. Harry	male	48.0
7	1.0	1.0	Andrews, Miss. Korpelia Theodosia	female	63.0
8	1.0	0.0	Andrews, Mr. Thomas Jr	male	39.0
9	1.0	1.0	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.0

Z analizy zbioru podobno wynika, że większe szanse na przeżycie katastrofy mieli członkowie kilkuosobowych rodzin.

70: Aline Numeric	71: Allen Numeric	72: Allis Numeric	73: Allison Numeric	74: Allum Numeric	75: Alma Numeric	76: Aloisia Numeric	77: Ar Num
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0