Regresja liniowa

Dane są zmienne $X_1, X_2, X_3, \ldots$ oraz Y. Zakładamy zależność:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + \ldots$$

gdzie parametry $a_0, a_1, a_2, \ldots$ należy oszacować na podstawie danych. W przypadku gdy mamy tylko jedną zmienną X, zależność ta przyjmuje formę:

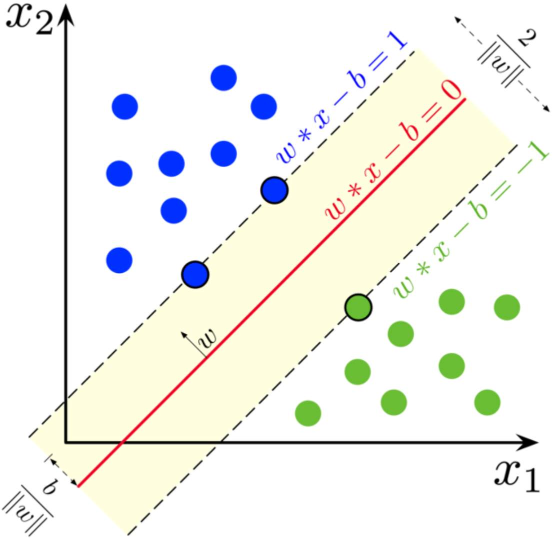
$$Y = aX + b.$$

Zauważmy, że w ogólnym przypadku $X_1, X_2, \ldots, Y$ przyjmują wartości rzeczywiste.

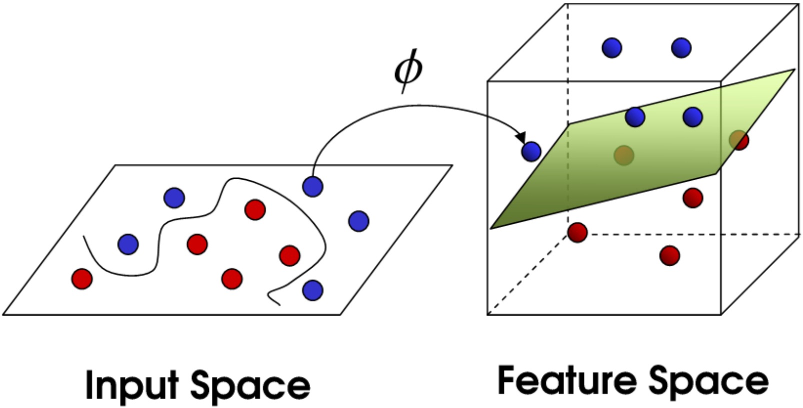
Regresja logistyczna

- W przypadku gdy Y może przyjmować tylko dwie wartości - przykładowo 0 i 1 stosowanie regresji liniowej jest nieuzasadnione.
- W tym przypadku definiujemy $p = P(Y = 1)$, $q = P(Y = 0)$.
- Oczywiście $p + q = 1$, zatem $q = 1 - p$. Celem regresji logistycznej jest oszacowanie p.
- Zakładamy następującą jego postać:
$$p = \frac{1}{1+e^{-(a_0+a_1X_1+a_2X_2+a_3X_3+\ldots)}}$$
- Jeżeli $p > 0.5$ to obserwacje klasyfikujemy do grupy 1, w przeciwnym przypadku do grupy 0.

SVM



SVM - kernelizacja

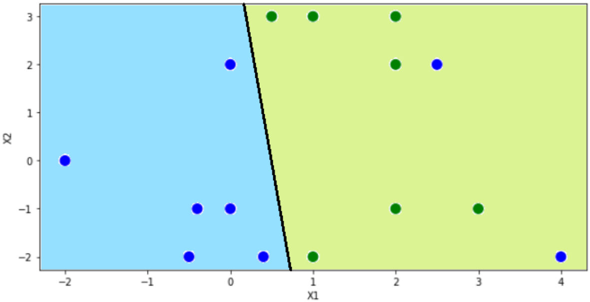


Przykład

X1	X2	Y
0	-1	0
0.4	-2	0
-0.5	-2	0
-0.4	-1	0
4	-2	0
-2	0	0
2.5	2	0
0	2	0
0.5	3	1
1	3	1
3	-1	1
1	-2	1
2	3	1
2	2	1
2	-1	1

Regresja liniowa/SVM

- Rozwiązanie z SVM:



- Rozwiązanie z regresją logistyczną:

$$p = \frac{1}{1+e^{-(-0.9011+0.5861x_1+0.4331x_2)}}$$

Przykładowo dla pierwszej obserwacji (0,-1) jest $p = \frac{1}{1+e^{-(-0.9011+0.5861\cdot 0+0.4331\cdot(-1))}} = 0.21 < 0.5$, zatem zaliczamy ją do klasy 0.