Metoda wektorów nośnych (SVM) – wstęp teoretyczny

Metoda wektorów nośnych (SVM, ang. *Support Vector Machines*) to algorytm nadzorowanego uczenia maszynowego, służący zarówno do rozwiązywania problemów klasyfikacyjnych (SVC, ang. *Support Vector Classification*), jak i regresyjnych (SVR, ang. *Support Vector Regression*). Ogólnie rzecz biorąc, metoda ta polega na wyborze hiperpłaszczyzny (2D – prostej, 3D – płaszczyzny) z maksymalnym marginesem, która najlepiej oddziela od siebie obiekty. Punkty znajdujące się na liniach granicznych wyznaczających margines nazywa się **wektorami nośnymi.** Wąski margines wnosi ryzyko błędnego prognozowania, a szeroki daje lepsze własności generalizacji modelu i mniejsze ryzyko jego przeuczenia.

SVM w ujęciu klasyfikacyjnym

Przy założeniu **liniowej separowalności** dwóch klas (podział binarny) hiperpłaszczyznę można przedstawić ogólnym wzorem:

$$\boldsymbol{wx} + b = 0$$

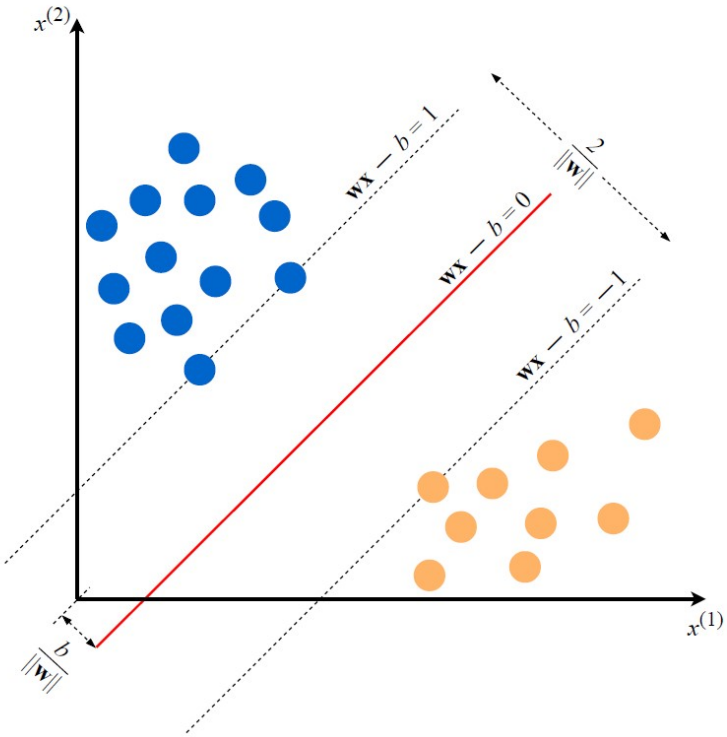
gdzie:

b – wyraz wolny,
$\boldsymbol{w}$ – wektor wag o takim samym rozmiarze jak $\boldsymbol{x}$,
$\boldsymbol{x}$ – macierz danych wejściowych.

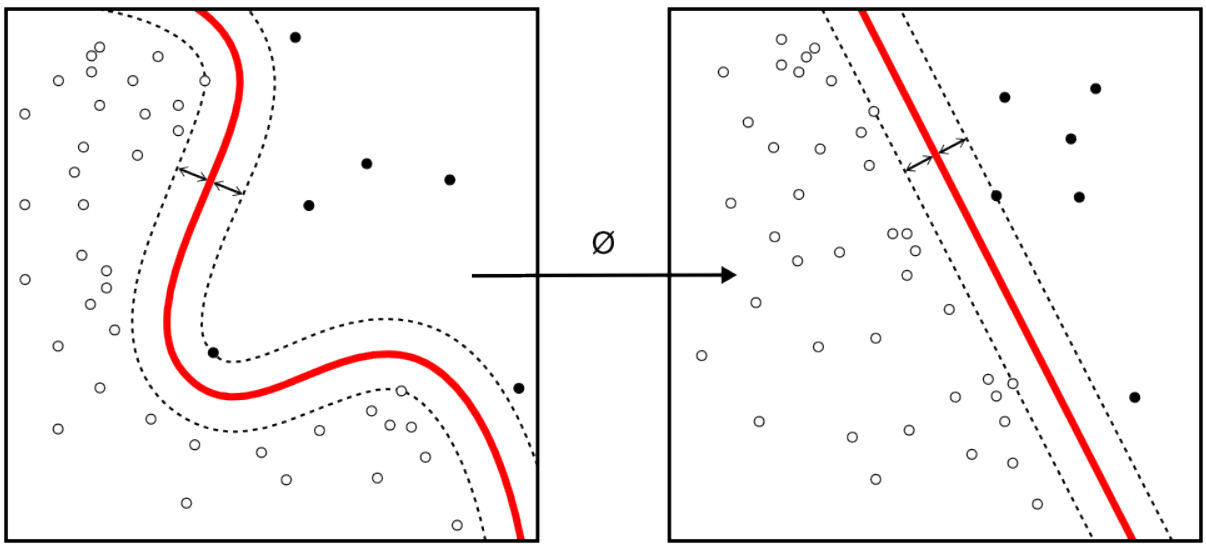
Algorytm SVC wymaga, aby odpowiedź pozytywna miała wartość 1, a odpowiedź negatywna -1. Oznacza to, że wektor odpowiedzi $\boldsymbol{y}$ dla wektora elementów $\boldsymbol{x}$ można obliczyć ze wzoru:

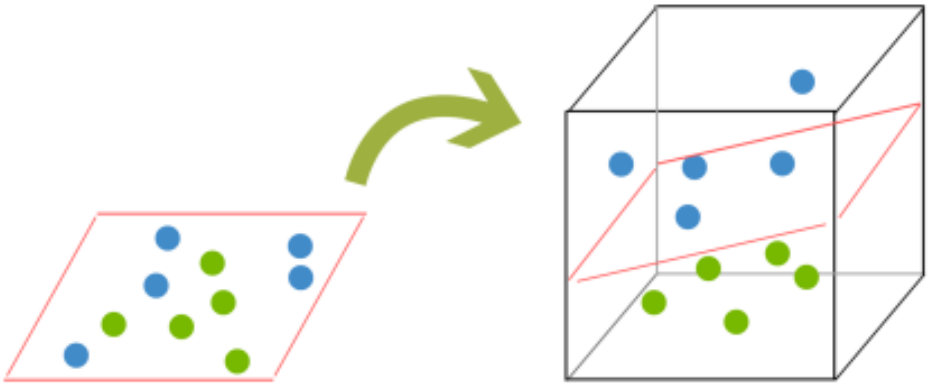
$$\boldsymbol{y} = +/-(\boldsymbol{wx} - b) = (w_1 \times x_1 + w_2 \times x_2 + \cdots + w_n \times x_n - b)$$

Znak funkcyjny zwraca wartość 1, jeśli obserwowana wartość zmiennej zależnej jest liczbą dodatnią oraz wartość -1 w przypadku, gdy jest ona liczbą ujemną. Punkty, dla których funkcja przyjmuje wartość ≥ 1 są zaklasyfikowane poprawnie, w innym przypadku są zaklasyfikowane niepoprawnie. O marginesie decyduje tzw. norma euklidesowa $||w||$ – im mniejszy wektor wagowy $\boldsymbol{w}$, tym większy margines.



Najczęściej jednak **klasy nie są liniowo separowalne** – algorytm wektorów nośnych radzi sobie z takimi wyzwaniami poprzez transformację obserwacji do przestrzeni o wyższym wymiarze, w którym możliwa jest liniowa separowalność klas, za pomocą funkcji jądra K, nazywanej inaczej sztuczką jądra (ang. *kernel trick*). Funkcja jądra jest sposobem obliczania iloczynu punktowego dwóch wektorów $\boldsymbol{x}$ i $\boldsymbol{y}$ w pewnej (wysokowymiarowej) przestrzeni cech. Do najczęściej stosowanych funkcji jądra należą: funkcje liniowe, funkcje wielomianowe, funkcje gaussowskie/radialne funkcje bazowe (RBF, ang. *radial basis function kernel*) oraz funkcje sigmoidalne.





SVM w ujęciu regresyjnym

Algorytm SVR działa na podobnej zasadzie do regresji liniowej – różnica polega na tym, że w funkcji liniowej dąży się do zminimalizowania błędu, a w regresji wektorów nośnych akceptowalne są błędy do pewnej wyznaczonej wartości. W odróżnieniu od podejścia klasyfikacyjnego, hiperpłaszczyzna oraz margines wyznaczane są w taki sposób, aby jak największa liczba punktów znalazła się między liniami granicznymi.

ZADANIE. Stwórz model klasyfikacyjny metodą wektorów nośnych wg poniższych wytycznych:

1) Wczytaj zbiór danych „penguins_size.csv” zawierający charakterystykę trzech gatunków pingwinów (Adelie, Chinstrap i Gentoo).
2) Wytnij brakujące wartości.
3) Stwórz wykresy pokazujące relacje pomiędzy zmiennymi (seaborn.pairplot).
4) Na podstawie wykresów wybierz: (i) dwie cechy oraz dwa gatunki (dwie klasy), dla których obiekty są liniowo separowalne, (ii) dwie cechy oraz dwa gatunki (dwie klasy), dla których obiekty nie są liniowo separowalne.
5) Dla obiektów liniowo separowalnych zbuduj model z wykorzystaniem liniowej funkcji jądra (kernel=”linear”) i policz dokładność, z jaką przewiduje. Liczebność zbioru walidacyjnego powinna wynosić 20 % całego zbioru danych. Pamiętaj o autoskalowaniu danych. Obiekty w przestrzeni cech przedstaw na wykresie z zaznaczeniem wektorów własnych oraz marginesem.
6) Dla obiektów liniowo nieseparowalnych zbuduj modele z wykorzystaniem:
    - liniowej funkcji jądra (kernel=”linear”),
    - funkcji wielomianowej jądra (kernel=”poly”),
    - radialnej funkcji bazowej (kernel=”rbf”).

    Dla każdego z nich oblicz dokładność. Liczebność zbioru walidacyjnego we wszystkich przypadkach powinna wynosić 20 % całego zbioru danych. Pamiętaj o autoskalowaniu danych. Obiekty w przestrzeni cech przedstaw na wykresie (dla kernel=”linear”) z zaznaczeniem wektorów własnych oraz marginesem.
7) Krótko skomentuj uzyskane wyniki.