Uczenie maszynowe:
wykład 12

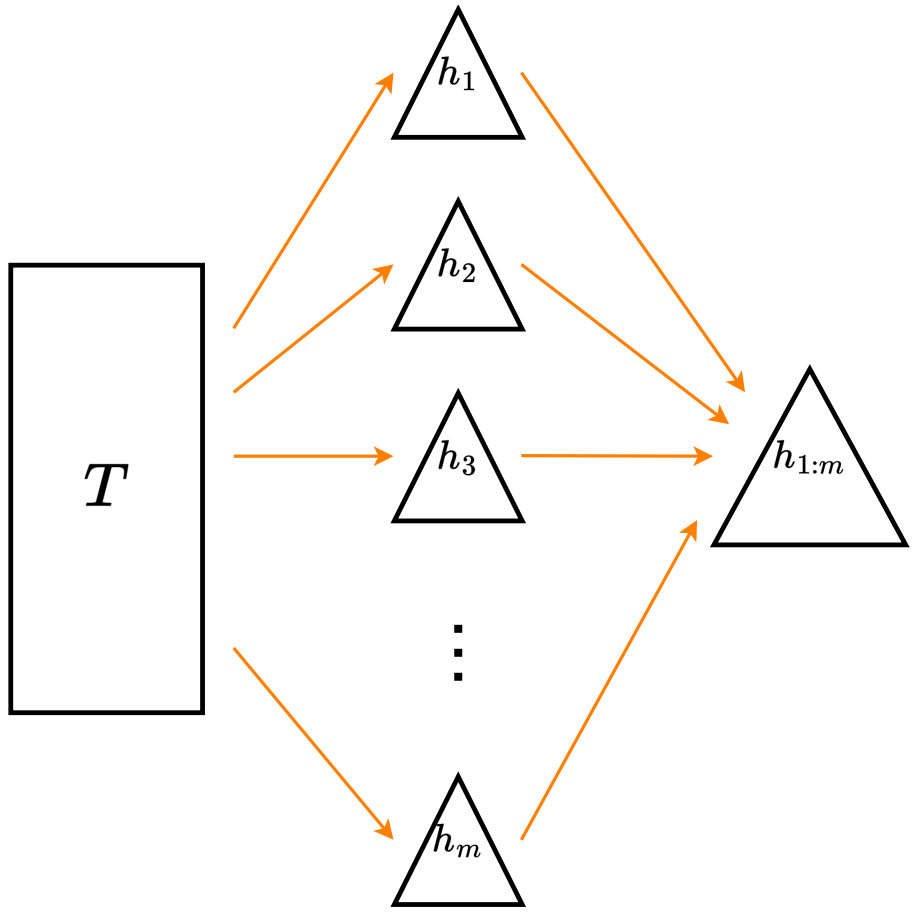
Paweł Cichosz

Koncepcja modelowania zespołowego

Motywacja: wzmocnienie siły predykcyjnej i kompensacja niedoskonałości przez łączenie modeli.

Modele bazowe: różne modele $h_1, h_2, \ldots, h_m$ dla tego samego zadania klasyfikacji lub regresji.

Łączenie predykcji: predykcja $h_{1:m}(x)$ na podstawie $h_1(x), h_2(x), \ldots, h_m(x)$.



Bagging

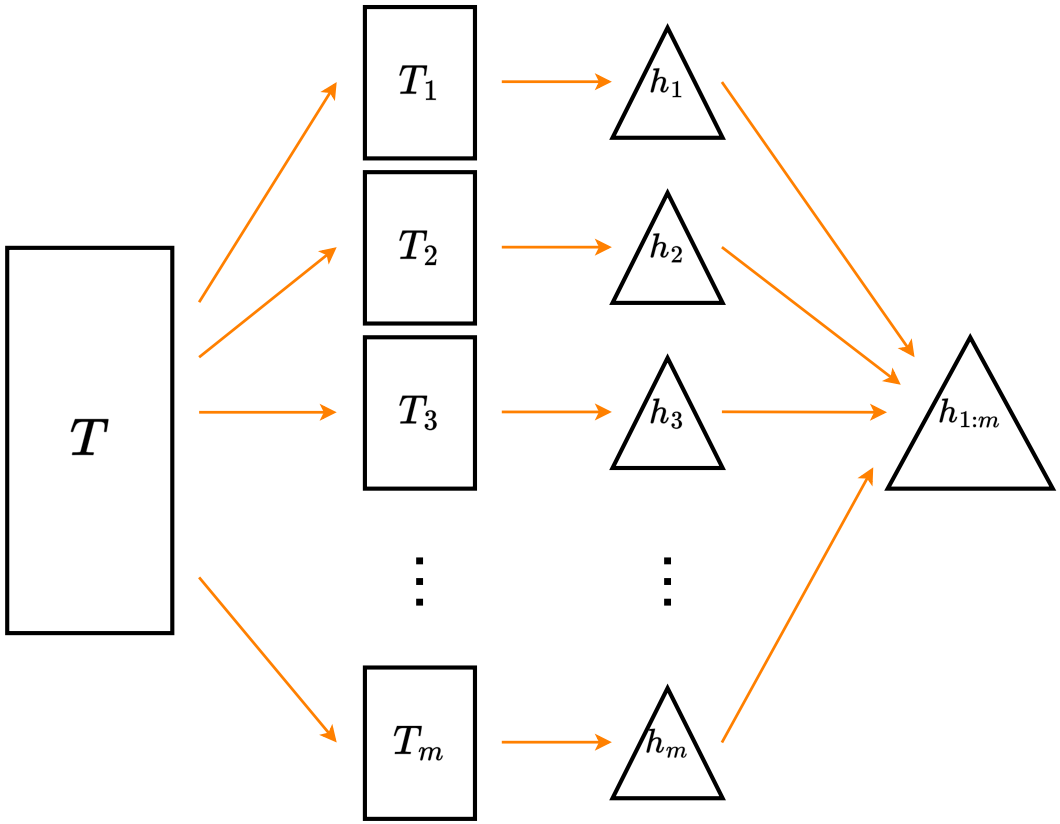
Modele bazowe: tworzone przez ten sam algorytm stosowany do *bootstrapowych prób* (losowanych ze zwracaniem) ze zbioru trenującego:

bag: ok. 63.2% przykładów $(1 - (1 - \frac{1}{N})^N \approx 1 - 1/e \approx 0.632)$.

out of bag: ok. 36.8% przykładów.

Łączenie predykcji: zwykłe głosowanie/uśrednianie.

Bagging



Właściwości: umiarkowana poprawa jeśli do tworzenia modeli bazowych jest stosowany niestabilny algorytm – wrażliwy na zaburzenia zbioru trenującego (zwykle drzewa decyzyjne/drzewa regresji).

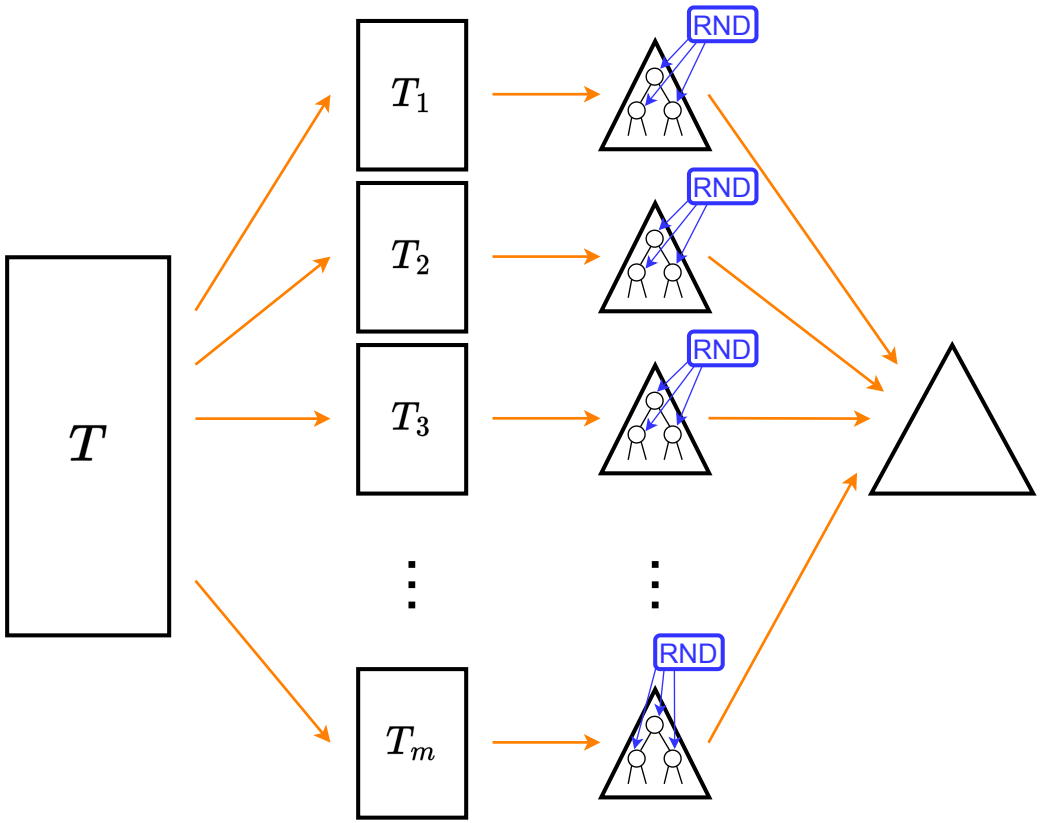
Las losowy

Modele bazowe: drzewa decyzyjne/drzewa regresji tworzone na podstawie prób *bootstrapowych* przez algorytm randomizowany w celu zwiększenia zróżnicowania:

- wybór podziału: spośród podziałów opartych na podzbiorze atrybutów niezależnie losowanym w każdym węźle (zazwyczaj losuje się $\lfloor\sqrt{n}\rfloor$ spośród wszystkich n atrybutów dla klasyfikacji i $\lfloor n/3\rfloor$ dla regresji).
- rozbudowane drzewa: późno działające kryteria stopu, brak przycinania (więcej węzłów – więcej okazji do zróżnicowania).

Łączenie predykcji: zwykłe głosowanie/uśrednianie.

Las losowy



Właściwości: zazwyczaj bardzo wysoka jakość predykcji przy niewielkim wysiłku, odporność na nadmierne dopasowanie, ograniczona wrażliwość na ustawienia parametrów, zwykle co najmniej kilkaset modeli bazowych.

Margines klasyfikacji

Margines geometryczny dla przykładu x: odległość od granicy decyzyjnej (nieujemna jeśli klasyfikowany poprawnie, ujemna jeśli klasyfikowany niepoprawnie):

$$\gamma_{\mathbf{w}}(x) = c_{-}(x)\frac{\mathbf{w} \circ \mathbf{a}(x)}{\|\mathbf{w}_{1:n}\|}$$

gdzie $c_{-}(x) = 2c(x) - 1$.

Margines funkcyjny dla przykładu x:

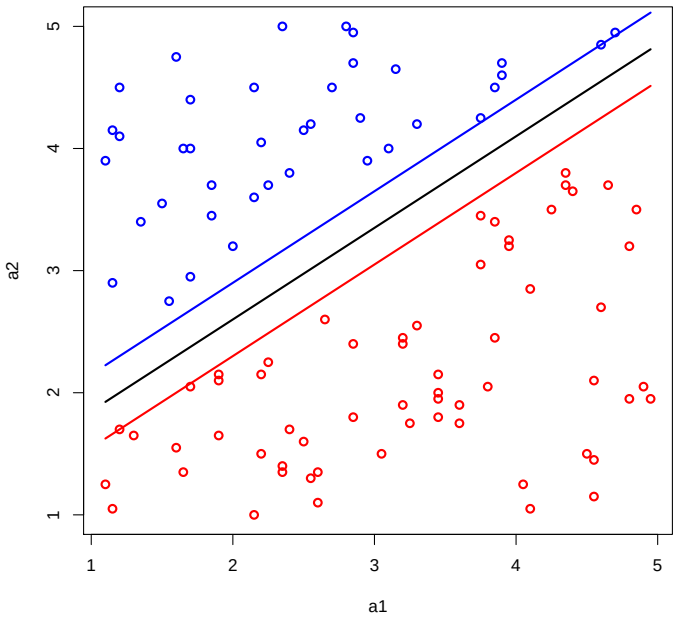
$$\hat{\gamma}_{\mathbf{w}}(x) = c_{-}(x)\mathbf{w} \circ \mathbf{a}(x)$$

Margines geometryczny/funkcyjny dla zbioru trenującego:

$$\gamma_{\mathbf{w}}(T) = \min_{x \in T} \gamma_{\mathbf{w}}(x)$$

$$\hat{\gamma}_{\mathbf{w}}(T) = \min_{x \in T} \hat{\gamma}_{\mathbf{w}}(x)$$

Margines klasyfikacji



Maksymalizacja marginesu: poszukiwanie położenia granicy decyzyjnej, które maksymalizuje margines geometryczny zbioru trenującego w celu zmniejszenia ryzyka nadmiernego dopasowania (pożądane zwłaszcza dla dużych n).

Niewrażliwość na skalowanie parametrów: pomnożenie wektora parametrów $w_1, \ldots, w_n, w_{n+1}$ przez dodatnią stałą nie zmienia położenia granicy decyzyjnej i wartości marginesu geometrycznego (ale odpowiednio skaluje wartość marginesu funkcyjnego).

Twardy margines

Założenie: liniowo separowalny zbiór trenujący, poszukiwana granica decyzyjna poprawnie separująca wszystkie przykłady.

Postać kanoniczna wektora parametrów: można ograniczyć się do wektorów parametrów spełniających warunek $\hat{\gamma}_{\mathbf{w}}(T) = 1$ (margines funkcyjny można dowolnie skalować nie zmieniając położenia granicy decyzyjnej).

Maksymalizacja marginesu geometrycznego: równoważna maksymalizacji $\frac{1}{\|\mathbf{w}_{1:n}\|}$, co z kolei jest równoważne minimalizacji $\|\mathbf{w}_{1:n}\|^2$.

Zadanie optymalizacji:

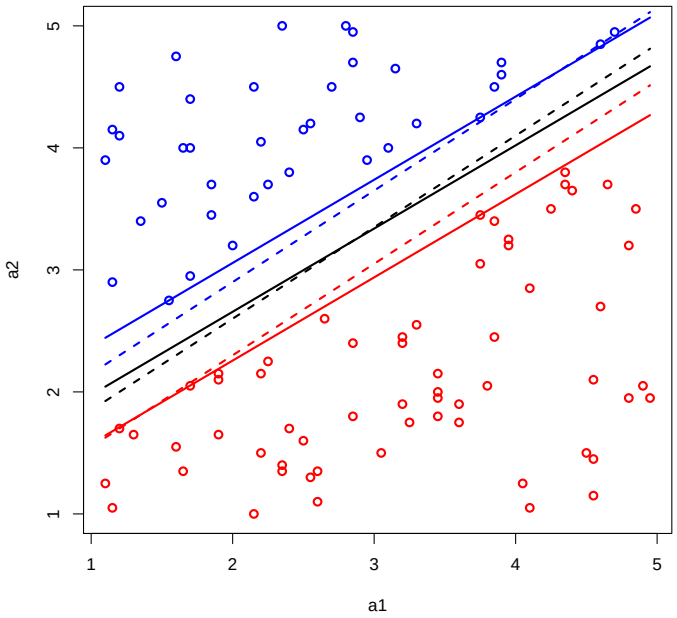
minimalizacja:

$$\frac{1}{2}\|\mathbf{w}_{1:n}\|^2$$

przy ograniczeniach:

$$(\forall x \in T) \quad c_-(x)\mathbf{w} \circ \mathbf{a}(x) \geq 1$$

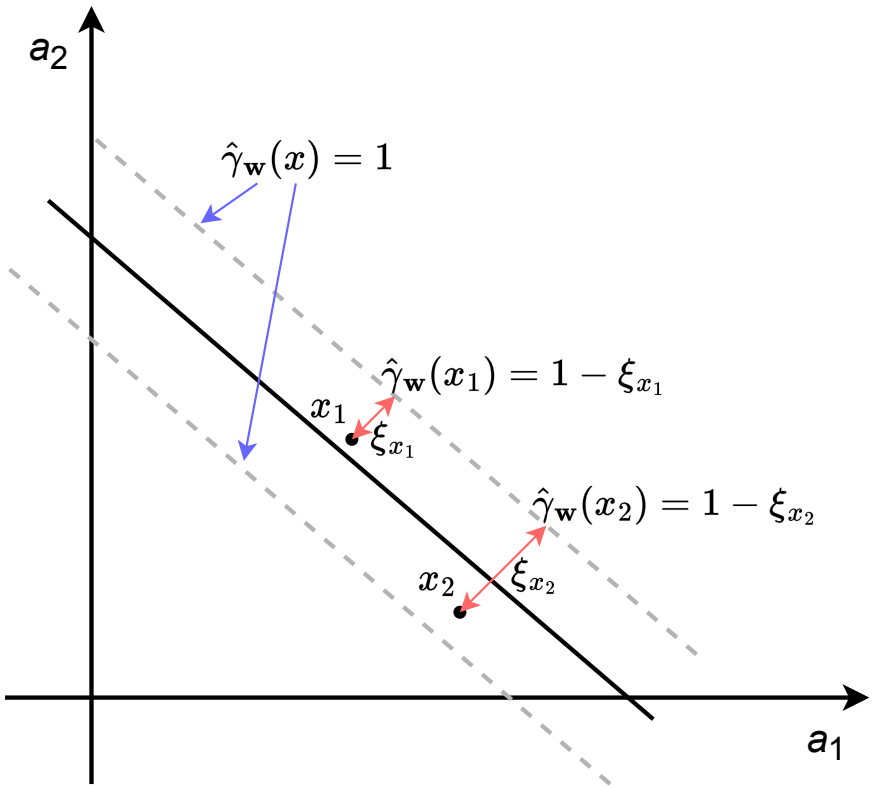
Twardy margines



Realizacja: programowanie kwadratowe lub bardziej efektywny dedykowany algorytm optymalizacji.

Wektory podpierające: przykłady, dla których $c_{-}(x)\mathbf{w} \circ \mathbf{a}(x) = 1$ („leżące na marginesie").

Miękki margines



Założenie: zbiór trenujący niekoniecznie liniowo separowalny, poszukiwana granica decyzyjna nie musi poprawnie separować wszystkich przykładów.

Zmienne luzujące: ξ_x – wielkość naruszenia ograniczenia dla przykładu x.

Miękki margines

Zadanie optymalizacji:

minimalizacja:

$$\frac{1}{2}\|\mathbf{w}_{1:n}\|^2 + \mathcal{C}\sum_x \xi_x$$

przy ograniczeniach:

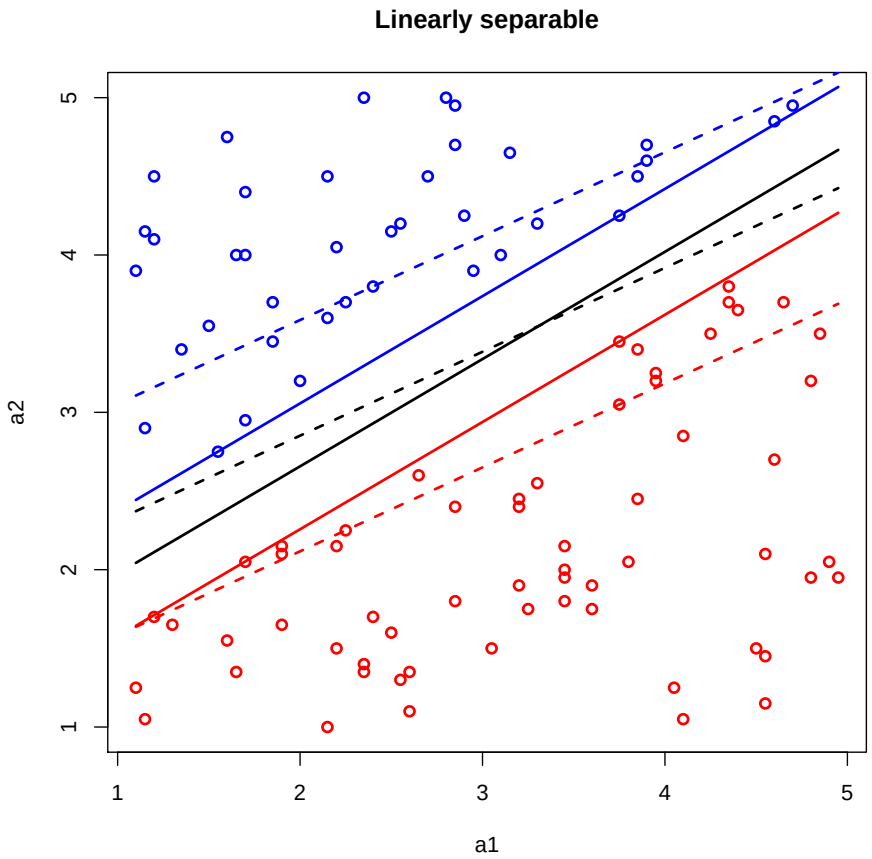
$$(\forall x \in T) \quad c_-(x)\mathbf{w} \circ \mathbf{a}(x) \geq 1 - \xi_x$$
$$(\forall x \in T) \quad \xi_x \geq 0$$

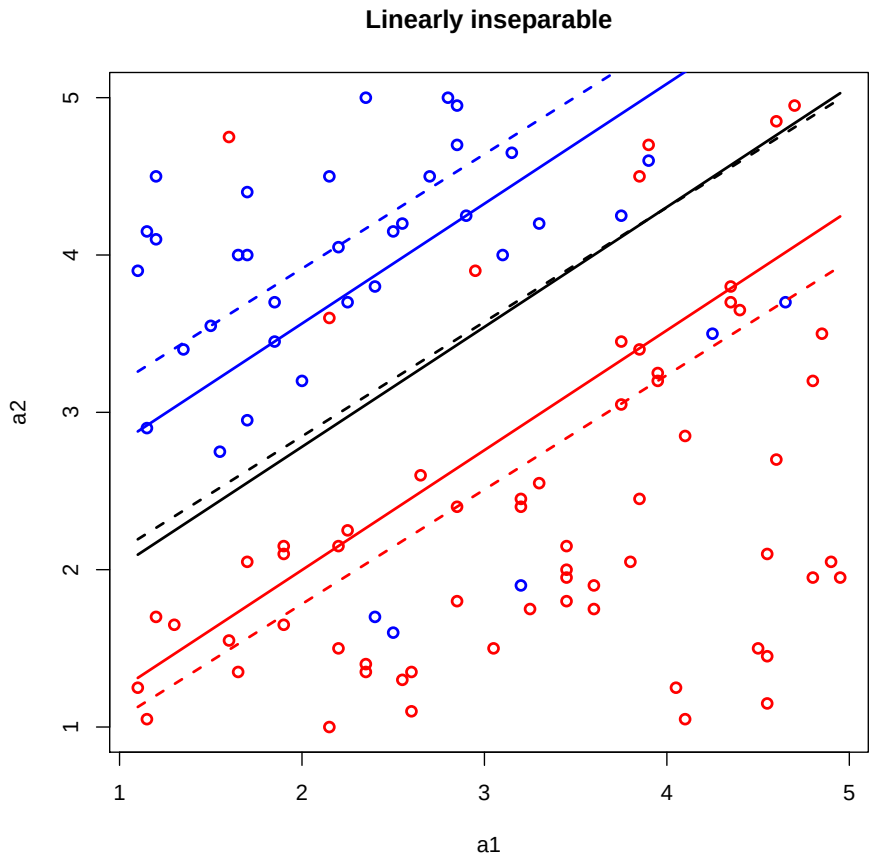
Koszt: $\mathcal{C}$ – koszt naruszenia ograniczenia (tutaj stosowane oznaczenie $\mathcal{C}$ zamiast standardowego C dla odróżnienia kosztu od zbioru klas).

Wektory podpierające: przykłady, dla których $c_-(x)\mathbf{w} \circ \mathbf{a}(x) \leq 1$ („leżące na marginesie" jeśli $c_-(x)\mathbf{w} \circ \mathbf{a}(x) = 1$).

Miękki margines

Niższe wartości $\mathcal{C}$ (linie przerywane) poszerzają margines przy zwiększeniu liczby przykładów naruszających ograniczenia.





Rozszerzenia algorytmu SVM

Funkcje jądrowe:

- niejawna nieliowa transformacja atrybutów,
- tworzenie i stosowanie modelu z użyciem przykładów wyłącznie parami jako argumentów iloczynu skalarnego,
- funkcja jądrowa stosowana do wektorów wartości pierwotnych atrybutów wyznacza iloczyn skalarny dla nowych atrybutów,
- szczegóły poza zakresem wykładu.

SVR (SVM do regresji):

- minimalizacja kwadratu normy wektora parametrów w celu ograniczenia ryzyka nadmiernego dopasowania,
- ograniczenia wymuszające dokładność predykcji,
- szczegóły poza zakresem wykładu.

Właściwości algorytmu SVM

- Należy do najbardziej skutecznych i często używanych algorytmów dla zadań klasyfikacji i regresji:
  - zwiększona odporność na nadmierne dopasowanie nawet przy dużej liczbie atrybutów,
  - możliwość reprezentowania nieliniowych zależności.
- Wymaga staranności przy doborze parametrów.
- Brak możliwości bezpośredniej interpretacji modeli.