Maszyna Wektorów Nośnych (SVM)

dr inż. Dawid Ewald

Zakład Sztucznej Inteligencji
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

22 marca 2024

Spis treści

Wprowadzenie

- Prof. V. Vapnik w 1998 r. stworzył nowe podejście do kształtowania struktury sieci neuronowej oraz definiowania problemu uczenia próbując wyeliminować znane wady sieci neuronowych typu MLP i RBF stosujące minimalizację nieliniowych funkcji błędu.

Problemy z sieciami MLP i RBF

- Minimalizowana funkcja błędu jest zwykle wielomodalna względem optymalizowanych parametrów i posiada wiele minimów lokalnych.
- Algorytm uczący nie jest w stanie skutecznie kontrolować złożoności struktury sieci neuronowej.

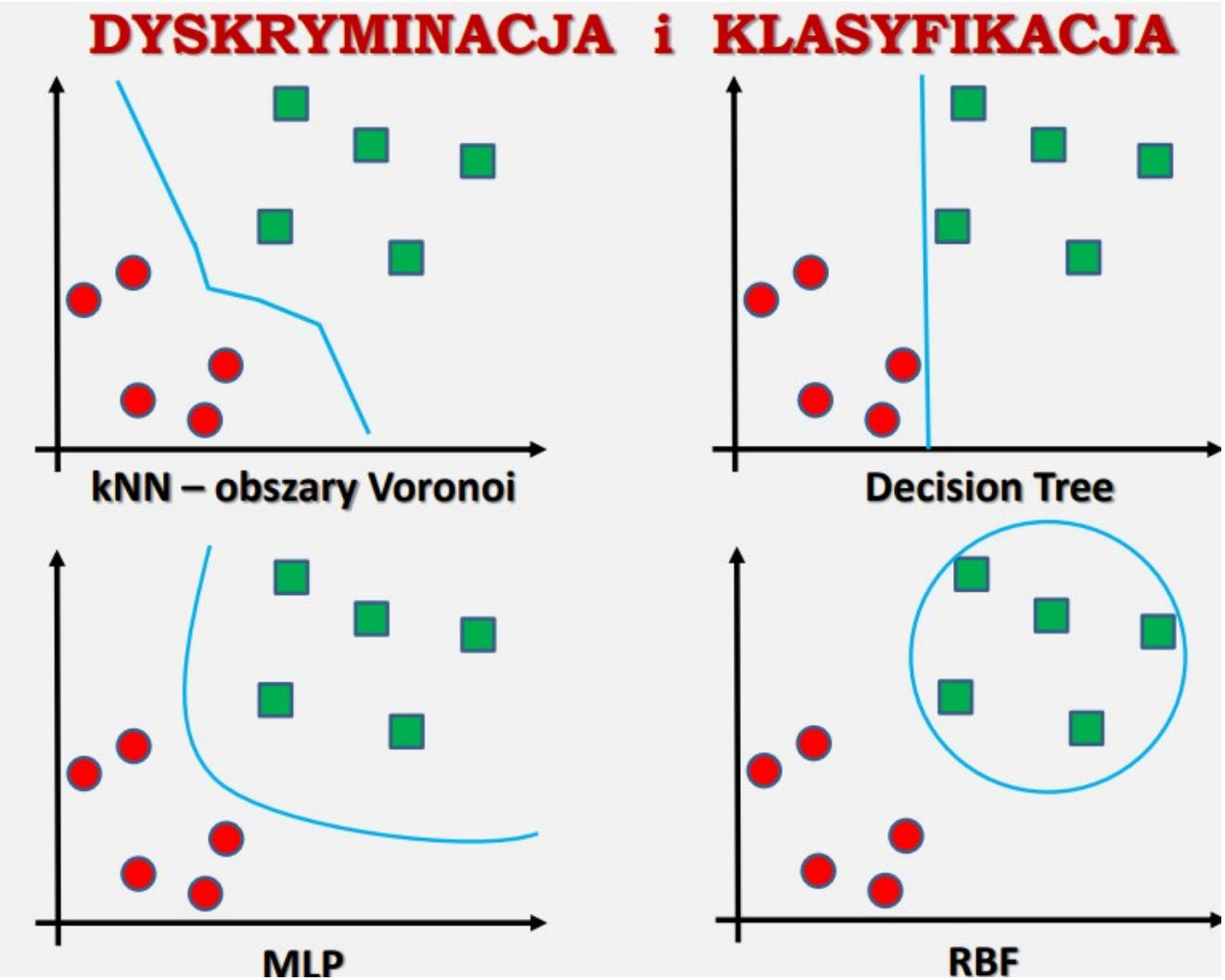
Nowe podejście

- Proces uczenia jest przedstawiony jako proces dobierania wag, w którym maksymalizowany jest margines separacji oddzielający skrajne (najbliższe) punkty w przestrzeni danych definiujących różne klasy.
- Bierze się pod uwagę tylko te najtrudniej separowalne punkty przestrzeni przy budowie modelu, które określają tzw. wektory nośne.

Sieci SVM

- Sieci SVM tworzą specyficzną dwuwarstwową strukturę neuropodobną stosującą różne rodzaje funkcji aktywacji (liniowe, wielomianowe, radialne, sigmoidalne).
- Uczenie oparte jest na programowaniu kwadratowym, które charakteryzuje się istnieniem tylko jednego minimum globalnego.
- Dedykowane głównie do zagadnień klasyfikacji, w których jedną klasę separujemy możliwie dużym marginesem od pozostałych klas.

Sieci SVM


DYSKRYMINACJA i KLASYFIKACJA
kNN – obszary Voronoi
Decision Tree
MLP
RBF

Klasyfikacja

- Klasyfikacja odnosi się do procesu przypisywania obiektów do określonych klas lub kategorii na podstawie ich cech.
- Jest to zadanie nadzorowane uczenia maszynowego, gdzie algorytm uczony jest na podstawie danych zawierających etykiety klas dla każdego obiektu.
- Celem klasyfikacji jest zbudowanie modelu, który potrafi przewidywać klasę nowych, nieznanych obserwacji na podstawie wcześniej widzianych danych uczących.

Dyskryminacja

- Dyskryminacja, w kontekście analizy danych, odnosi się do procesu identyfikowania różnic lub zależności pomiędzy różnymi grupami obserwacji.
- Może to obejmować identyfikację cech lub wzorców, które rozróżniają jedną grupę od drugiej.
- Dyskryminacja może być wykorzystywana jako część procesu klasyfikacji, gdzie celem jest znalezienie tych cech lub wzorców, które są istotne dla rozróżnienia między klasami.

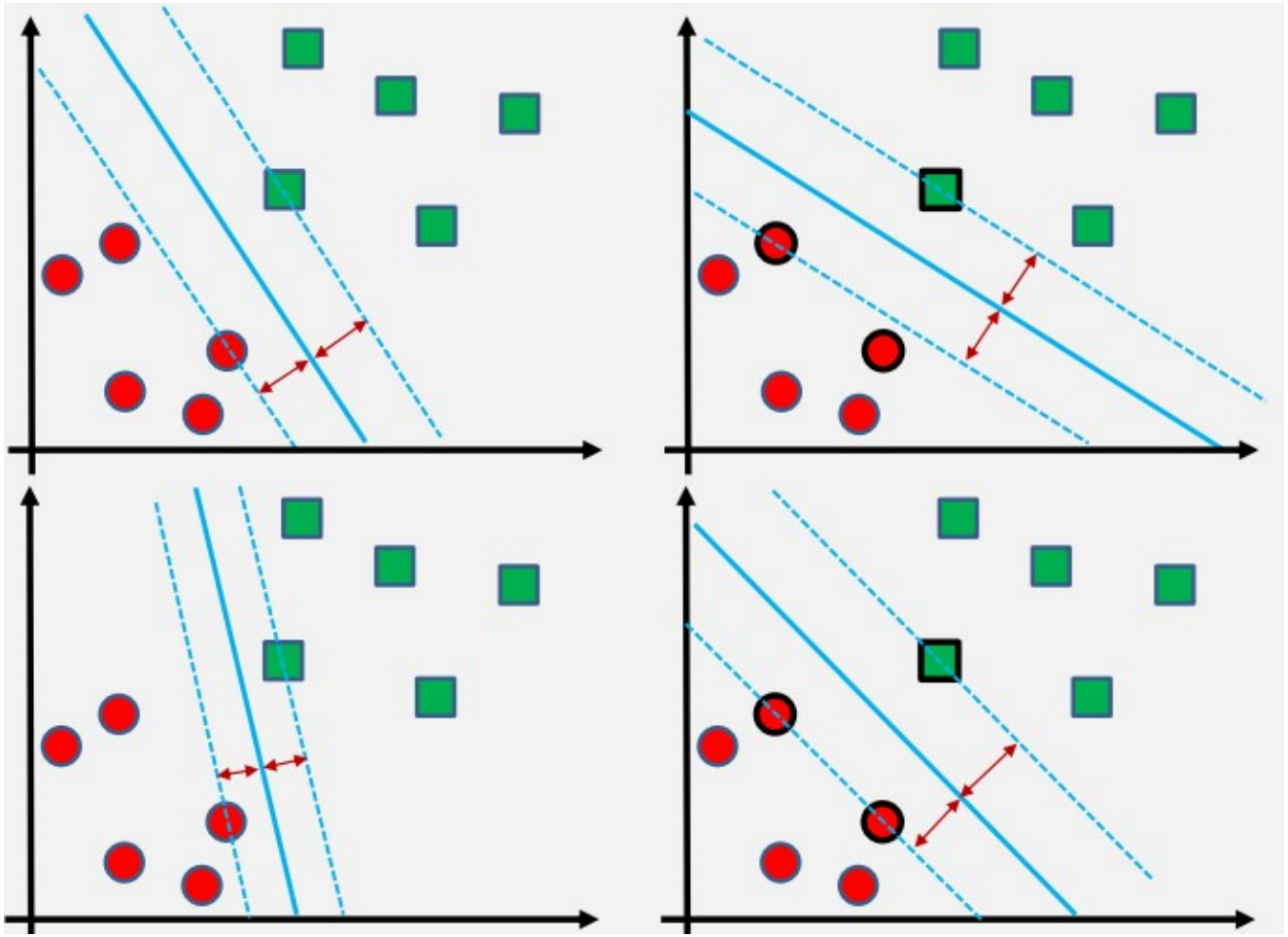
Podsumowanie

- Klasyfikacja skupia się na przewidywaniu klas dla nowych obserwacji.
- Dyskryminacja skupia się na analizie różnic między grupami danych.
- Dyskryminacja może być używana jako narzędzie w procesie klasyfikacji do identyfikowania istotnych cech lub wzorców, które pomagają w dokonywaniu trafnych predykcji klas.

Metoda SVM

- Metoda Maszyny Wektorów Nośnych (SVM) ma na celu wyznaczenie najszerszej granicy dyskryminacji spośród możliwych, których zwykle istnieje nieskończona ilość.
- Jest to istotne, ponieważ większa szerokość granicy może przekładać się na lepszą zdolność generalizacji modelu i lepszą wydajność predykcyjną.
- SVM dąży do maksymalizacji marginesu separacji pomiędzy klasami, co prowadzi do wyznaczenia granicy, która jest możliwie najbardziej oddalona od punktów danych obu klas.

Metoda SVM



Maksymalizacja marginesu oddzielającego klasy

- Aby osiągnąć jak najlepszą dyskryminację wzorców poszczególnych klas, warto zmaksymalizować margines oddzielający te klasy.
- Margines ten to odległość pomiędzy najbliższymi punktami różnych klas a hiperpłaszczyzną separującą te klasy.
- Im większy margines, tym większa szansa na poprawne oddzielenie klas i lepszą generalizację modelu.

Ograniczenie analizy do najtrudniejszych punktów przestrzeni

- Gdy mamy do czynienia z wieloma danymi, można ograniczyć analizę tylko do tych punktów, które są najtrudniejsze do zdyskryminowania.
- Koncentrując się na tych trudnych przypadkach, model może być lepiej dopasowany do rzeczywistych warunków i mieć lepszą zdolność do rozróżniania klas.

Charakterystyka modelu uwzględniającego najtrudniejsze wzorce

- Model uwzględniający najtrudniejsze wzorce powinien cechować się dobrą jakością oraz prostotą reprezentacji.
- Poprzez skupienie się na trudnych przypadkach, model może być bardziej adekwatny do rzeczywistych warunków, co może prowadzić do lepszych wyników w klasyfikacji.

Wyznaczenie optymalnej hiperpłaszczyzny dyskryminującej klasy

- Aby rozróżnić wzorce jednej klasy od pozostałych, należy wyznaczyć optymalną hiperpłaszczyznę dyskryminującą.
- Hiperpłaszczyzna ta powinna jak najlepiej oddzielać klasy, aby minimalizować błędy w klasyfikacji.

Opis algorytmu SVM

- Załóżmy, że mamy zbiór p par uczących: x_i, d_i dla $i = 1, 2, \ldots, p$, gdzie x_i to wektor danych wejściowych, a $d_i \in \{-1, +1\}$ reprezentuje dyskryminowane klasy: $d_i = +1$ oznacza klasę dyskryminowaną, zaś $d_i = -1$ oznacza pozostałe klasy.
- Przy założeniu liniowej separowalności obu klas możliwe jest określenie równania hiperpłaszczyzny separującej te wzorce: $w^T x + b = 0$, gdzie w to wektor wag, x to wektor danych wejściowych, a b to polaryzacja.
- Możemy więc zdefiniować równania decyzyjne:
  - Jeżeli $w^T x + b \geq 0$ to $d_i = +1$
  - Jeżeli $w^T x + b \leq 0$ to $d_i = -1$

Opis algorytmu SVM

- Co możemy zapisać w postaci nierówności: $d_i(w^T x + b) \geq 1$, której spełnienie przez pary punktów x_i, d_i definiuje wektory nośne (support vectors), które decydują o położeniu hiperpłaszczyzny i szerokości marginesu separacji.
- Potrzebne jest więc wyznaczenie b oraz w, żeby określić decyzję.

Przekroczenie granic separacji

- Czasami występuje konieczność zmniejszenia marginesu separacji dla problemów niecałkowicie separowalnych liniowo oraz punktów x_i, d_i leżących wewnątrz strefy marginesu separacji, co możemy zapisać za pomocą nierówności:

$$d_i(w^T x_i + b) \geq 1 - \delta_i$$

gdzie $\delta_i \geq 0$ i zmniejsza margines separacji.

- Jeśli:
  - $0 \leq \delta_i < 1$ – wtedy x_i, d_i leży po właściwej stronie hiperpłaszczyzny, więc decyzja o przynależności do klasy będzie poprawna,
  - $\delta_i = 1$ – wtedy x_i, d_i leży na hiperpłaszczyźnie, więc decyzja o przynależności do klasy będzie nieokreślona,
  - $\delta_i > 1$ – wtedy x_i, d_i leży po niewłaściwej stronie hiperpłaszczyzny, więc decyzja o przynależności do klasy będzie błędna.
- Określając granicę decyzyjną, należy więc możliwie zminimalizować wartość δ_i.

Szerokość marginesu separacji

- Szerokość marginesu separacji możemy wyznaczyć jako iloczyn kartezjański wektora wag w oraz różnicy odległości dwóch wektorów nośnych x^+ i x^- należących do przeciwnych klas:

$$\Delta = (x^+ - x^-) \cdot \frac{w}{\|w\|} = 2 \cdot r \cdot \frac{w}{\|w\|} = 2 \cdot r \cdot x_{SV}$$

  gdzie r to odległość wektorów nośnych od hiperpłaszczyzny.
- Chcąc więc zmaksymalizować margines separacji pomiędzy wektorami nośnymi różnych klas $\Delta = 2 \cdot r \cdot \frac{w}{\|w\|}$, trzeba zminimalizować w, co jest równoważne minimalizacji wyrażenia $\frac{1}{2} \cdot \|w\|^2$ przy pewnych ograniczeniach liniowych wynikających ze zdefiniowanej nierówności decyzyjnej.
- W takich przypadkach stosujemy mnożniki Lagrange'a i minimalizujemy funkcję Lagrange'a.

Funkcja Lagrange'a dla maksymalizacji marginesu separacji

- Możemy więc teraz określić funkcję Lagrange'a dla problemu maksymalizacji marginesu separacji:

$$\min_{\mathrm{w},\mathrm{b},\boldsymbol{\alpha},\boldsymbol{\delta}} \frac{1}{2}\mathrm{w}^T\mathrm{w} + \xi\sum_{i=1}^{p}\delta_i$$

przy zdefiniowanych ograniczeniach:

$$d_i(\mathrm{w}^T\mathrm{x}_i + b) \geq 1 - \delta_i$$

$$\delta_i \geq 0$$

gdzie ξ to waga, z jaką traktowane są błędy testowania w stosunku do marginesu separacji, decydującą o złożoności sieci neuronowej, dobieraną przez użytkownika w sposób eksperymentalny, np. metodą walidacji krzyżowej.

- Otrzymujemy więc następującą funkcję Lagrange'a:

$$L(\mathrm{w},\mathrm{b},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\mu}) = \frac{1}{2}\mathrm{w}^T\mathrm{w} + \xi\sum_{i=1}^{p}\delta_i - \sum_{i=1}^{p}\alpha_i(d_i(\mathrm{w}^T\mathrm{x}_i + b) - 1 + \delta_i) - \sum_{i=1}^{p}\mu_i\delta_i$$

gdzie $\boldsymbol{\alpha}$ jest wektorem mnożników Lagrange'a o wartościach

Funkcja Lagrange'a dla maksymalizacji marginesu separacji

Rozwiązanie minimalizacji funkcji Lagrange'a polega na określeniu punktu siodłowego, czyli wyznaczenia pochodnych cząstkowych względem mnożników.

Minimalizacja funkcji Lagrange'a

- Warunki optymalnego rozwiązania wyznaczone są zależnościami:

$$\frac{\partial \mathcal{L}}{\partial \mathsf{w}} = 0 \rightarrow \mathsf{w} = \sum_{i=1}^{p} \alpha_i \delta_i \mathsf{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial \mathsf{b}} = 0 \rightarrow \sum_{i=1}^{p} \alpha_i \delta_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\delta}} = 0 \rightarrow \boldsymbol{\mu}_i = \xi - \alpha_i$$

Podstawienie do funkcji Lagrange'a

- Możemy teraz podstawić wyznaczone zależności do funkcji Lagrange'a:

$$\mathcal{L}(\mathrm{w}, \mathrm{b}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\mu}) = \frac{1}{2}\sum_{i=1}^{p} \alpha_i \delta_i \mathrm{x}_i^T \mathrm{x}_i + \xi \sum_{i=1}^{p} \delta_i - \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \delta_i \mathrm{x}_i^T \mathrm{x}_j$$
$$- \xi \sum_{i=1}^{p} \alpha_i \delta_i \mathrm{x}_i^T \mathrm{x}_i + \sum_{i=1}^{p} \alpha_i - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j \delta_i \delta_j \mathrm{x}_i^T \mathrm{x}_j$$
$$= \sum_{i=1}^{p} \alpha_i - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j \delta_i \delta_j \mathrm{x}_i^T \mathrm{x}_j$$

Problem dualny

- W punkcie siodłowym iloraz mnożnika Lagrange'a δ_{SVM} i odpowiedniego ograniczenia związanego μ_{SVM} z wektorem nośnym x_{SVM} jest równy zeru ($\delta_{SVM}\mu_{SVM} = 0$), gdyż $\mu_{SVM} = 0$, więc zależność:

$$w^T x_i + b \geq 1 - \mu_i$$

w punkcie wektora nośnego sprowadza się do:

$$w^T x_{SVM} + b = \pm 1$$

co pozwala wyznaczyć wartość b:

$$b = \pm 1 - w^T x_{SVM}$$

Otrzymany problem dualny

- Otrzymaliśmy więc problem dualny

$$\max_{\boldsymbol{\alpha}} \mathcal{Q}(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \alpha_j \delta_i \delta_j \mathrm{x}_i^T \mathrm{x}_j$$

przy ograniczeniach dla $i = 1, 2, \ldots, p$ zdefiniowanych następująco:

$$0 \leq \alpha_i \leq \rho$$

$$\sum_{i=1}^{p} \alpha_i \delta_i = 0$$

Rozwiązanie problemu dualnego pozwala znaleźć poszukiwaną hiperpłaszczyznę:

$$\mathrm{y}(\mathrm{x}) = \sum_{i=1}^{p} \alpha_i \delta_i \mathrm{x}_i^T \mathrm{x} + b$$

Wnioski

- Zmienna dopełniająca δ_i ani mnożniki Lagrange'a nią związane nie pojawiają się w sformułowaniu problemu dualnego.
- Mnożniki muszą spełniać jedynie podstawowy warunek mówiący, iż iloczyn mnożników i wartości funkcji ograniczenia dla każdej pary danych uczących jest równy zeru. Jeśli więc ograniczenie spełnione jest z nadmiarem dla wektorów nienośnych, wtedy mnożniki te muszą być równe zeru. Niezerowe wartości mnożników występują zaś dla wektorów nośnych.

Wektory nośne w SVM

- Niezerowe wartości mnożników określają wektory nośne, których ilość oznaczamy jako $N_{SVM} \leq p$, a więc równanie sieci liniowej SVM o wagach optymalnych wyznacza hiperpłaszczyznę zależną tylko od wektorów nośnych:

$$y(x) = \sum_{i=1}^{N_{SVM}} \alpha_i \delta_i x_i^T x + b$$

Nieliniowe przekształcenie danych

- Większość problemów klasyfikacji nie posiada jednak właściwości liniowej separowalności. Potrzebne jest więc nieliniowe rzutowanie danych oryginalnych w inną przestrzeń funkcyjną, gdzie wzorce staną się separowalne liniowo i będzie można zastosować hiperpłaszczyznę separującą SVM.
- Warunkiem jest zastosowanie transformacji nieliniowej o odpowiednio wysokim wymiarze K przestrzeni cech $K \geq N$.

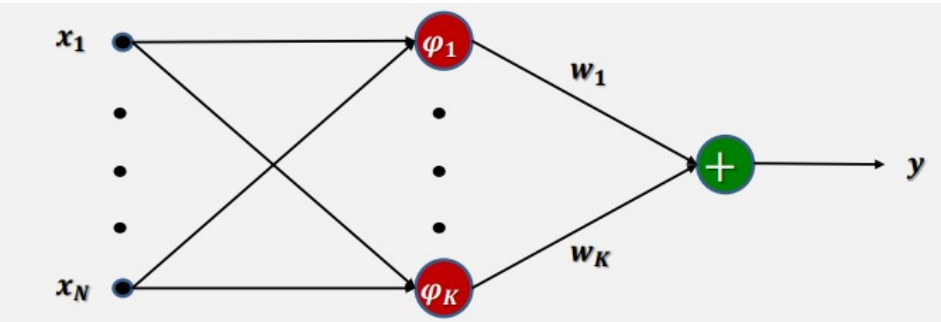
Nieliniowa sieć SVM

Dla zadań nieseparowalnych liniowo rzutujemy każdy wzorzec z jego N-wymiarowej przestrzeni cech do K-wymiarowej przestrzeni cech $\boldsymbol{\xi}_j \mathrm{x}$, gdzie $j = 1, 2, \ldots, K$. W efekcie tego nieliniowego przekształcenia równanie hiperpłaszczyzny określone będzie wzorem:

$$\mathrm{y}(\mathrm{x}) = \mathrm{w}^T \boldsymbol{\xi}(\mathrm{x}) + b = \sum_{j=1}^{K} w_j \xi_j(\mathrm{x}) + b$$

Nieliniowa sieć SVM

- gdzie w_i oznaczają wagi prowadzące od neuronu o nieliniowej funkcji aktywacji ξ_j na wektorze danych wejściowych x do wyjściowego neuronu liniowego.
- Otrzymujemy więc dwuwarstwową strukturę sieci neuronowej zawierającą jedną warstwę ukrytą:



Nieliniowa sieć SVM

Rozwiązanie problemu pierwotnego uzyskujemy więc poprzez zastąpienie zmiennej x_i przez $\boldsymbol{\xi}_i x$. Otrzymujemy więc:

$$\max_{\boldsymbol{\alpha}} \mathcal{Q}(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \alpha_j d_i d_j K(x_i, x_j)$$

gdzie K nazywamy funkcją jądra (*kernel function*), zdefiniowaną następująco:

$$K(x_i, x_j) = \boldsymbol{\xi}^T x_i \boldsymbol{\xi} x_j$$

Rozwiązanie problemu nieliniowej sieci SVM

Rozwiązanie problemu sprowadza się do wyznaczenia wartości wag sieci:

$$\mathrm{w} = \sum_{i=1}^{p} \alpha_i d_i \boldsymbol{\xi} \mathrm{x}_i$$

$$b = \pm 1 - \mathrm{w}^T \boldsymbol{\xi} \mathrm{x}_i$$

Otrzymując ostatecznie sygnał wyjściowy dla nieliniowej sieci SVM w postaci:

$$\mathrm{yx} = \mathrm{w}^T \boldsymbol{\xi} \mathrm{x} + b = \sum_{i=1}^{N_{SV}} \alpha_i d_i K(\mathrm{x}_i, \mathrm{x}) + b = 0$$

Sygnał wyjściowy dla nieliniowej sieci SVM

Otrzymując ostatecznie sygnał wyjściowy dla nieliniowej sieci SVM w postaci:

$$\mathrm{y}\mathrm{x} = \mathrm{w}^T \boldsymbol{\xi}\mathrm{x} + b = \sum_{i=1}^{N_{SV}} \alpha_i d_i K(\mathrm{x}_i, \mathrm{x}) + b = 0$$

Na kandydatów na funkcje jądra K możemy wybrać funkcje spełniające warunek twierdzenia Mercera, np. funkcje gaussowskie, wielomianowe, sklejane, a nawet sigmoidalne przy pewnych ograniczeniach.

Nieliniowe funkcje jądra sieci SVM

Do najczęściej stosowanych funkcji jądra należą:

- Funkcje liniowe:

$$K(\mathrm{x}_i, \mathrm{x}) = \mathrm{x}^T \mathrm{x}_i + \gamma$$

- Funkcje wielomianowe:

$$K(\mathrm{x}_i, \mathrm{x}) = (\mathrm{x}^T \mathrm{x}_i + \gamma)^p$$

- Funkcje gaussowskie:

$$K(\mathrm{x}_i, \mathrm{x}) = \exp\left(-\frac{\|\mathrm{x} - \mathrm{x}_i\|^2}{2\gamma^2}\right)$$

- Funkcje sigmoidalne:

$$K(\mathrm{x}_i, \mathrm{x}) = \tanh(\beta \mathrm{x}^T \mathrm{x}_i + \gamma)$$

Gdzie β, γ to stałe współczynniki liczbowe, a p to stopień wielomianu.

Sieć SVM z radialnymi funkcjami bazowymi (RBF)

Sieć SVM z radialnymi funkcjami bazowymi jest bardzo podobna do sieci radialnej RBF, aczkolwiek sposób jej tworzenia i wyznaczania wag różni się. Podobnie, stosując funkcje sigmoidalne, otrzymujemy dwuwarstwową sieć MLP.
Aby zastosować sieci SVM do większej ilości klas niż dwie, konieczne jest zbudowanie kilku sieci SVM, które dyskryminują wzorce każdej z klas od pozostałych lub pomiędzy parą każdych dwóch klas, a następnie wyniki są sumowane.

Algorytmy rozwiązania zadania dualnego dla dużych zbiorów danych

- Niezależnie od zastosowanego jądra i rodzaju zadania główny problem obliczeniowy w sieciach SVM sprowadza się do rozwiązania zadania programowania kwadratowego z ograniczeniami liniowymi.
- Problemem staje się duża ilość danych uczących, co związane jest z nieraz ogromną ilością optymalizowanych zmiennych – tutaj mnożników Lagrange'a. Pojawiają się problemy z pamięcią i złożonością obliczeniową, co eliminuje możliwość zastosowania klasycznych metod programowania kwadratowego, np. MINOS, OSL, LOQO czy Matlab.

Algorytmy rozwiązania zadania dualnego dla dużych zbiorów danych

- Stosuje się dekompozycję zbioru uczącego na szereg podzbiorów oraz strategię aktywnych ograniczeń wynikających z równości, zaniedbując te nieaktywne ze znakiem silniej nierówności. Dzięki temu w kolejnych iteracjach następuje przemieszczanie części wzorców ze zbioru ograniczeń aktywnych do nieaktywnych.
- Wykorzystuje się również różne wersje algorytmu programowania sekwencyjnego SMO lub BSVM Platta oraz suboptymalną metodę SVMLight Joachimsa.

Wykłady online o SVM

- Open MIT Lectures about SVM:
  `https://www.youtube.com/watch?v=_PwhiWxHK8o`
- Caltech Lectures about SVM:
  `https://www.youtube.com/watch?v=eHsErlPJWUU`