LABORATORIUM SIECI NEIRONOWE MASZYNY WEKTORÓW PODPIERAJĄCYCH

Danuta Gaweł

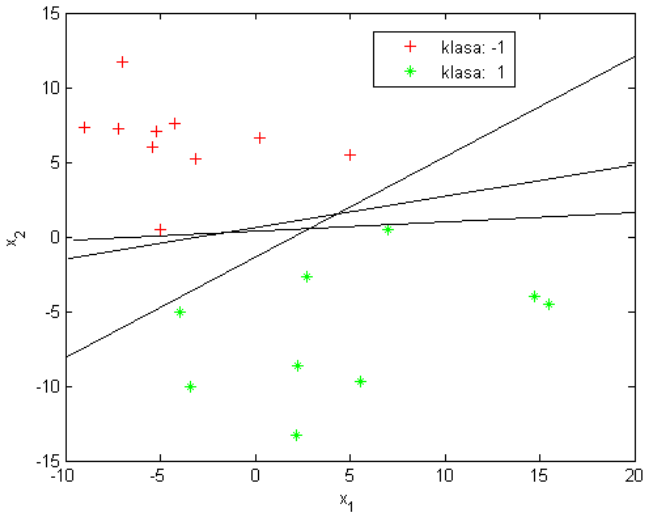
15 maja 2013

1 Maszyny Wektorów Podpierających

Maszyna wektorów podpierających (SVM – Support Vector Machine) to klasyfikator, którego celem jest znalezienie hiperpłaszczyzny separującej dwie grupy obiektów z maksymalnym marginesem [1].

Aby wyjaśnić powyższe stwierdzenie konieczne jest wprowadzenie oznaczeń opisujących każdy z obiektów należących do zbioru poprzez zbiór par $\{(x_i, y_i)\}_{i=1,2,\ldots,N}$, gdzie x_i to punkty w przestrzeni n-wymiarowej $x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \in R^n$, a y_i to etykieta mówiąca o przynależności i-tego obiektu do konkretnej klasy ($y_i = \{-1, 1\}$) [1].

W najprostszym przypadku, gdy klasy są separowane liniowo można znaleźć nieskończenie wiele płaszczyzn separujących dwie klasy obiektów (rysunek 1).



Rysunek 1: Przykład grup obiektów separowanych liniowo wraz z przykładowymi płaszczyznami separującymi

Chcemy jednak, aby płaszczyzna separująca była maksymalnie oddalona od obiektów różnych klas. Spowoduje to, że prawdopodobieństwo popełnienia błędnej klasyfikacji punktu na podstawie jego położenia względem hiperpłaszczyzny będzie mniejsze.

Określmy zatem płaszczyznę poprzez jej wektor normalny (prostopadły do niej) $[w] = (w_1, w_2, \ldots, w_n)$ i wyraz wolny b. Równanie tej płaszczyzny można wyrazić jako[1]:

$$b + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \tag{1}$$

lub jako

$$b+ < \mathbf{w}, \mathbf{x} > . \tag{2}$$

Gdzie $< \mathbf{w}, \mathbf{x} >$ oznacza iloczyn skalarny dwóch wektorów $\mathbf{w}$ i $\mathbf{x}$.
Do określenia maksymalnej odległości płaszczyzny od punktów należących do różnych klas wykorzystywana jest odległość d punktu x od płaszczyzny opisanej przez wzór (1) lub (2). Odległość tę można obliczyć na podstawie wzoru [2]:

$$d(\mathbf{x}, \mathbf{w}, b) = \frac{|b+ < \mathbf{w}, \mathbf{x} > |}{||\mathbf{w}||} \tag{3}$$

gdzie $||\mathbf{w}||$ oznacza długość wektora $\mathbf{w}$ czyli normę. Zatem [2]:

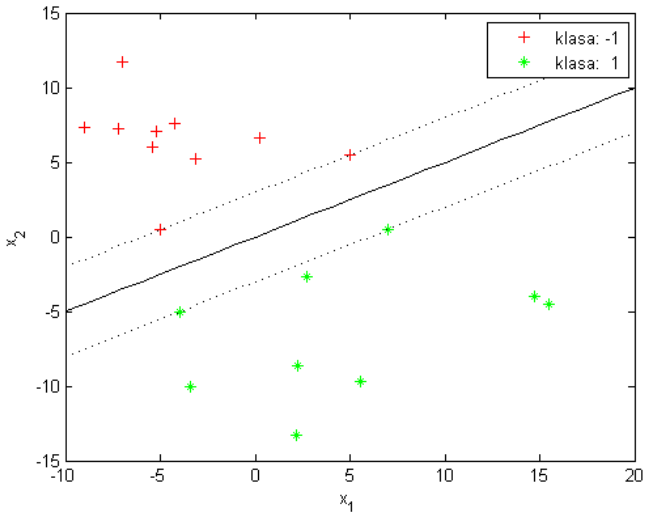
$$||\mathbf{w}|| = \sqrt{w_1^2 + w_2^2 + \ldots + w_n^2} \tag{4}$$

Wspomnianym wcześniej marginesem separacji oznaczanym literą τ nazywa się odległość płaszczyzny od najbliższego względem niej punktu(obiektu). Zatem dla ustalonej płaszczyzny wielkość ta wyrażona jest wzorem [2]:

$$\tau(\mathbf{w}, b) = \min_{i=1,2,\ldots,N} \frac{y_i(b+ < \mathbf{w}, x >)}{||\mathbf{w}||} \tag{5}$$

Znalezienie płaszczyzny dla której margines separacji jest maksymalny jest zadaniem optymalizacyjnym względem τ przy ograniczeniach [2]:

$$\forall i y_i(b+ < \mathbf{w}, x_i >) \geqslant \tau(\mathbf{w}, b)||\mathbf{w}|| \tag{6}$$



Rysunek 2: Przykład grup obiektów separowanych liniowo wraz z hiperpłaszczyzną separującą i maksymalnym marginesem separacji

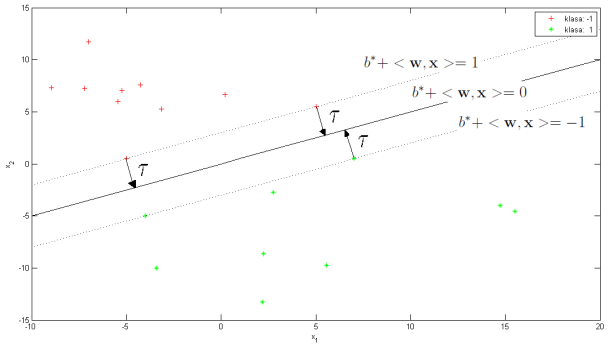
Znalezienie płaszczyzny spełniającej powyższe warunki nie gwarantuje jednak uzyskania pojedynczego rozwiązania ponieważ wymnożenie równania płaszczyzny (2) obustronnie poprzez tę samą dowolną liczbę nie spowoduje zmiany położenia tejże płaszczyzny, a jedynie przeskalowanie wartości $\mathbf{w}^*$ i b^* reprezentujących znalezione rozwiązanie [2]. Zakłada się, że

$$||\mathbf{w}||\tau = 1 \tag{7}$$

Zmieniają się zatem ograniczenia zadania optymalizacyjnego [2]:

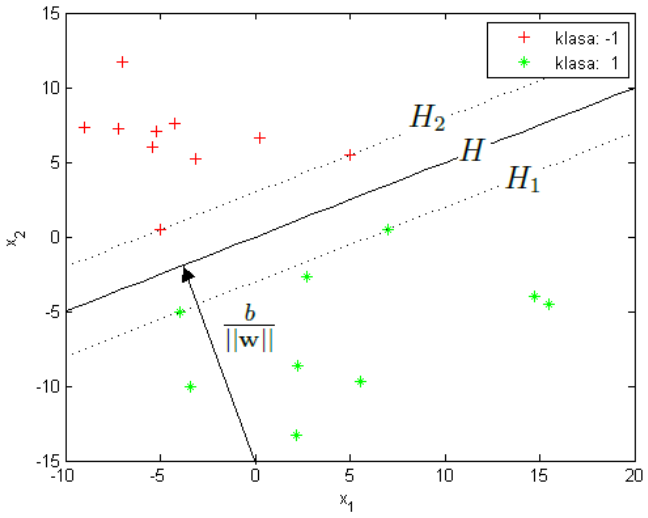
$$\forall i \; y_i(b+ < \mathbf{w}, x_i >) \geqslant 1 \tag{8}$$

Dodatkowo z równania (7) wynika, że im mniejsza jest długość wektora $\mathbf{w}$ tym większy jest margines τ. Zatem ze zbioru płaszczyzn spełniających (8) należy wybrać tę której $\mathbf{w}$ ma najmniejszą normę [2].



Rysunek 3: Przykład grup obiektów separowanych liniowo wraz z hiperpłaszczyzną separującą i maksymalnym marginesem separacji [2]

Odległość hiperpłaszczyzny od środka układu można obliczyć jako [1]: $\frac{b}{||\mathbf{w}||}$ (rys.4). Podobnie odległości dwóch pozostałych prostych wyznaczających margines separacji od początku układu współrzędnych to odpowiednio [1]: $\frac{|1-b|}{||\mathbf{w}||}$ i $\frac{|-1-b|}{||\mathbf{w}||}$. Zatem odległość pomiędzy H_1 i H_2 (oznaczenia jak na rysunku 4) wynosi $\frac{|2|}{||\mathbf{w}||}$.



Rysunek 4: Przykład grup obiektów separowanych liniowo wraz z hiperpłaszczyzną separującą i maksymalnym marginesem separacji [1]

Aby margines separacji był jak największy należy minimalizować funkcję $||\mathbf{w}||$. Jednak ze względu na dalsze obliczenia zamiast minimalizować $||\mathbf{w}||$ minimalizuje się $Q(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2$. W tym celu wykorzystuje się metodę mnożników Lagrange'a. Funkcja Lagrange'a w tym przypadku ma postać [1]:

$$Q(\mathbf{w}, b, \lambda) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} (\lambda_i \{y_i(b + < \mathbf{w}, x_i > -1)\}) \tag{9}$$

Punktami podparcia będziemy nazywać punkty danych x_i dla których odpowiadające $\lambda_i^* > 0$, gdzie λ_i^* to optymalne rozwiązania (9).

Reguła dyskryminacyjna ma zatem postać [1]:

$$f(x) = sgn[\sum_{\text{wektory podpierające}} y_i \lambda_i^0 (< x_i, x >) + b_0] \tag{10}$$

gdzie λ_i^0 to optymalne mnożniki Lagrange'a
b_0 to stała spełniająca warunki konieczne na istnienie punktu siodłowego czyli [2]:

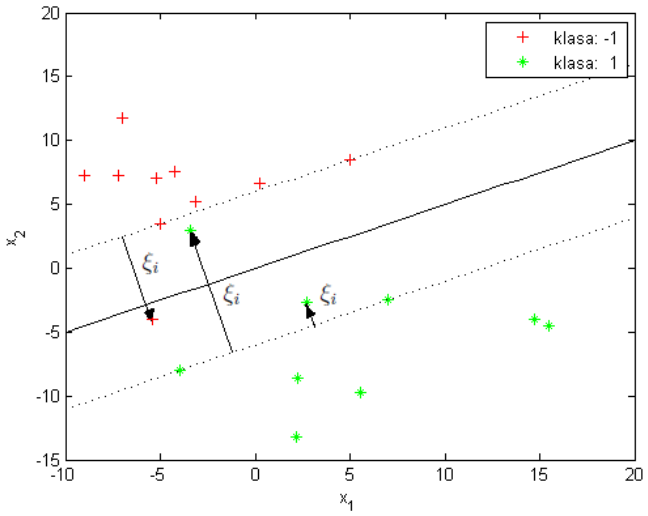
$$\frac{\partial Q}{\partial w_1} = 0, \frac{\partial Q}{\partial w_2} = 0, ..., \frac{\partial Q}{\partial w_n} = 0 \tag{11}$$

$$\frac{\partial Q}{\partial b} = 0 \tag{12}$$

$$\forall i \ \frac{\partial Q}{\partial \lambda_i} = 0 \tag{13}$$

dla $\forall i \ \lambda_i \geqslant 0$

W przypadku bardziej skomplikowanym, gdy badane grupy nie są separowalne liniowo znajdowana jest płaszczyzna na podstawie której możliwa jest jak najbardziej poprawna klasyfikacja (dopuszcza się występowanie punktów „po złej stronie" hiperpłaszczyzny separującej rys. 5). Płaszczyzna taka znajduje się możliwie jak najdalej od skupisk typowych dla danej klasy.



Rysunek 5: Przykład grup obiektów nieseparowanych liniowo wraz z hiperpłaszczyzną separującą separacji[2]

W takim przypadku należy zatem zmienić ograniczenia [1]:

$$y_i(< \mathbf{w}, x_i > +b) \geqslant 1 - \xi_i \forall i \ \xi \geqslant 0 \tag{14}$$

Gdzie ξ (zmienna luźna) to odległość punktu znajdującego się wewnątrz pewnego „marginesu" od odpowiedniej płaszczyzny separacji (rys. 5), lub inaczej jest to odległość na jakiej dany punkt leży w pasie „marginesowym" [2].

Jeżeli dany punkt leży poza tym „marginesem" po „odpowiedniej" stronie płaszczyzny separującej x_i to $\xi_i = 0$. Jeżeli punkt leży pomiędzy płaszczyzną H i H_1 lub H_2, ale po „dobrej" stronie płaszczyzny to $\xi_i \in (0, \tau >$. Jeżeli natomiast x_i leży po „nieodpowiedniej" stronie płaszczyzny separującej to $\xi_i > \tau$ [2].

Zadanie optymalizacyjne sprowadza się w tym przypadku do znalezienia maksymalnego „marginesu" takiego żeby suma ξ_i była jak najmniejsza.

Minimalizujemy zatem wyrażenie [2]:

$$Q(\mathbf{w}, \xi_1, \xi_2, \ldots, \xi_N) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \xi_i \tag{15}$$

Przy ograniczeniach $\forall i \; y_i(b+ < \mathbf{w}, x_i >) + \xi_i \geqslant 1$, $\xi_i \geqslant 0$. C jest współczynnikiem kary [1], przyjmowaną dodatnią stałą wyrażającą kompromis pomiędzy dużym „marginesem", a wielkością sumy błędów ξ_i. Zatem im C zostanie ustalone większe, tym wpływ drugiego członu równania (15) będzie większy i rozwiązanie będzie preferowało mniejszy „margines". W przeciwnym wypadku, gdy C będzie mniejsze, większy będzie „margines" nawet za cenę dużych błędów ξ_i [2].

Oznacza to zatem, że mała wartość C spowoduje, że końcowe położenie płaszczyzny będzie bardziej zależało od położenia bardziej typowych punktów w klasach niż od błędów ξ_i punktów odstających lub mniej typowych [2].

Czasem jednak, gdy zbiory nie są separowane liniowo stosuje się podniesienie wymiarowości za pomocą której znajdowana jest krzywoliniowa granica klasyfikacji [2] (w bogatszej przestrzeni funkcja klasyfikacyjna jest liniowa).

W metodzie tej wektory x zostają zastąpione funkcjami wektorowymi $h(x)$ [1]. Pozostałe kroki metody pozostają takie same, zatem funkcja dyskryminacyjna przyjmuje postać [1]:

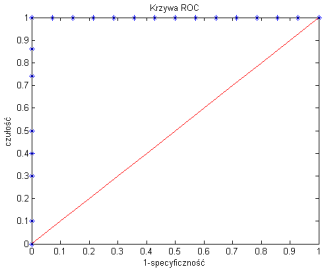
$$f(x) = sgn(\sum_{\text{wektory podpierające}} y_i \lambda_i^0 (< h(x_i), h(x) >) + b_0) \tag{16}$$

gdzie $< h(x_i), h(x) >= \mathbf{K}(x_i, x)$ to funkcja jądrowa (ang. Kernel function).

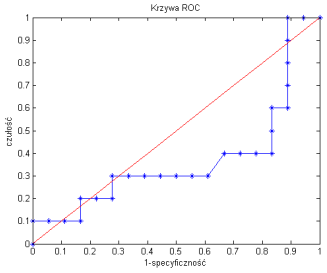
2 Krzywa ROC

Najlepszym klasyfikatorem jest ten dla którego wartość czułości i specyficzności jest największa. Na wartości te można jednak wpływać poprzez zmianę parametrów funkcji decyzyjnej powodując tym samym zmianę klasyfikacji obserwacji zbioru testowego. Wykreślenie zależności pomiędzy tymi wskaźnikami dla zmienianych parametrów funkcji decyzyjnej poprzez zaznaczenie na osi odciętych wartości $1 -$ specyficzność, a na osi rzędnych wartości czułości pozwala na uzyskanie wskaźnika jakości klasyfikatora niezależnego od przyjętych parametrów funkcji decyzyjnej.

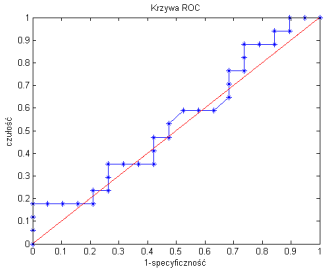
Przykładowe krzywe ROC zostały przedstawione na rysunkach 6, 7 i 8.



Rysunek 6: Idealna krzywa ROC



Rysunek 7: Przykładowa krzywa ROC



Rysunek 8: Przykładowa krzywa ROC

Idealna krzywa ROC została przedstawiona na rysunku (6). W przypadku gdy krzywa ROC pokrywa się z przekątną (zaznaczoną na rysunkach czerwoną linią) klasyfikator nadaje obserwacjom testowym losowe etykiety (najgorszy wynik). Przykład tego typu krzywej ROC został przedstawiony na rysunku (8).

2.1 Pole pod krzywą ROC (AUC)

AUC (*ang. Area Under Curve*) jest to pole pod krzywą ROC. AUC może przyjmować wartości z przedziału od 0 do1. Im większa wartość tego wskaźnika tym lepszy jest klasyfikator. Wykorzystywanie wartości pola pod krzywą ROC jest wygodniejszym sposobem porównywania wyników oceny jakości klasyfikacji aniżeli porównywanie samych krzywych ROC.

3 Zadania do wykonania na zajęciach

1. Wczytanie danych do przestrzeni roboczej Matlaba
2. Wykorzystanie jednej z poznanych na poprzednich zajęciach metod do stworzenia zbiorów uczącego i testowego
3. Klasyfikacja danych i obliczenie wskaźników jakości klasyfikacji
4. Stworzenie wykresu ROC dla stworzonego klasyfikatora
5. Obliczenie wartości AUC
6. Przetestowanie wpływu funkcji jądrowej na wynik klasyfikacji

W sprawozdaniu proszę zamieścić kod programu wraz z komentarzami oraz wykresy z zaznaczonymi punktami należącymi do zbioru danych wraz z wektorami podpierającymi i znalezioną hiperpłaszczyną.
Przydatne funkcje: svmtrain, svmclassify

Literatura

[1] Maszyny wektorów podpierających, R.Powalski,Uniwersytet Warszawski, Wydział Nauk Ekonomicznych (www.mimuw.edu.pl~dorotaj3iie_sm2_2009-2010svm.pdf 19.02.2012r.)

[2] Klasyfikatory SVM, P. Klęsk (http:wikizmsi.zut.edu.pluploadsee2Svm.pdf 19.02.2012r.)