Techniki eksploracji danych

Krzysztof Gajowniczek

Rok akademicki: 2020/2021

Section 1

Maszyna wektorów nośnych

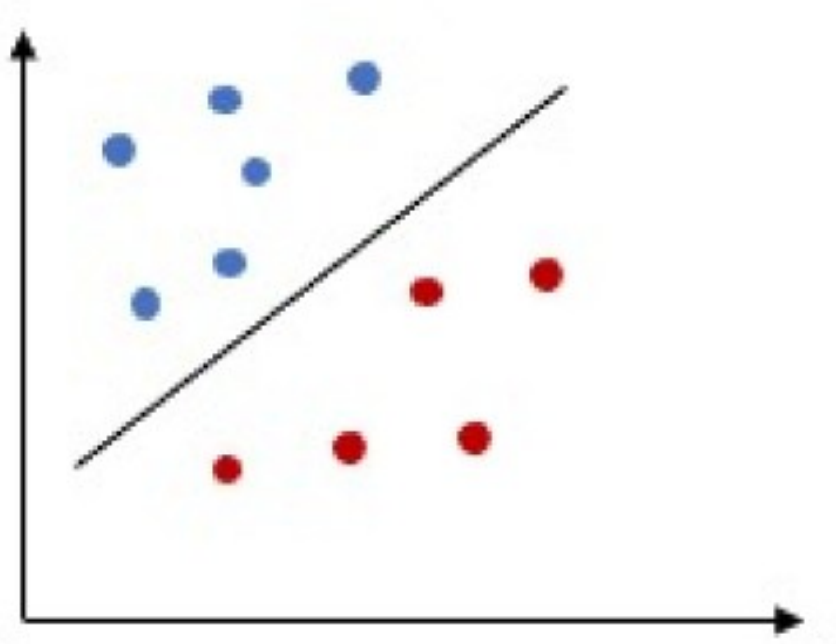
Subsection 1

Idea

- **Maszyna wektorów nośnych, maszyna wektorów podpierających** - abstrakcyjny koncept maszyny, która działa jak klasyfikator, a której nauka ma na celu wyznaczenie hiperpłaszczyzny rozdzielającej z maksymalnym marginesem przykłady należące do dwóch klas.
- Istnieją rozszerzenia pierwotnej idei na przypadek klasyfikacji wieloklasowej oraz regresji.

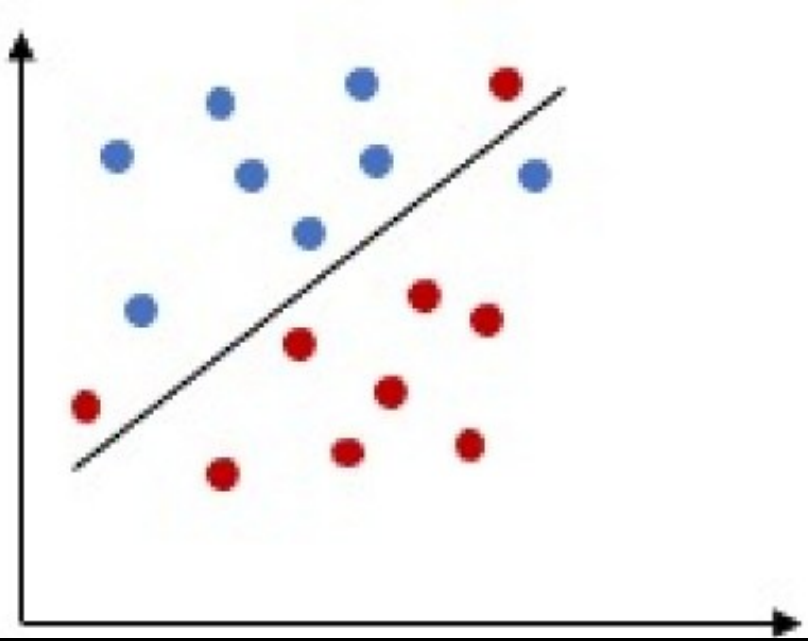
Liniowa SVM – twardy margines

- Tutaj budujemy naszą początkową koncepcję SVM, klasyfikując idealnie oddzielony zbiór danych (klasyfikacja liniowa).



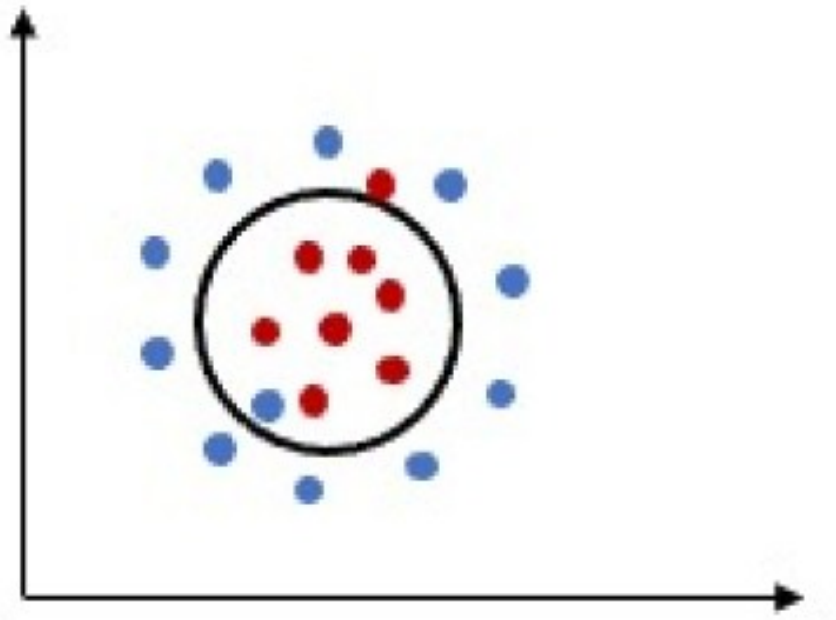
Liniowa SVM – miękki margines

- Rozszerzymy naszą koncepcję, aby zsklasyfikować zbiór danych, w którym występują wartości nietypowe.
- W takim przypadku wszystkie punkty danych nie mogą być oddzielone linią prostą, będą pewne punkty niepoprawnie sklasyfikowane.
- Jest to podobne do dodawania regularyzacji do modelu regresji.



Nieliniowa SVM

- Wreszcie wprowadzamy nieliniową SVM za pomocą metod jądrowych.



Subsection 2

Hiperpłaszczyzna

- Do oddzielenia danych, które są w dwóch wymiarach (mają 2 cechy x_1 i x_2), potrzeba użyć linii.
- Podobnie potrzeba płaszczyzny $2D$, aby oddzielić dane w 3 wymiarach.

Wzór

$$h(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x} + \theta_0 = \sum_{j=1}^{p} \theta_j x^{(j)} + \theta_0$$

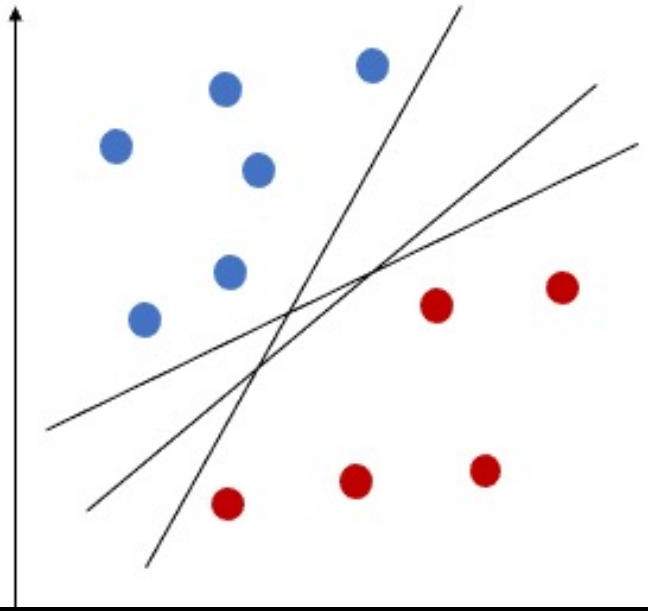
- Klasyfikację przeprowadza się za pomocą hiperpłaszczyzny rozdzielającej.
- Znak $h(\boldsymbol{x}_i)$ wskazuje, czy etykieta wyjściowa to $+1$ czy -1, a wielkość określa, jak daleko $\boldsymbol{x}_i$ leży od hiperpłaszczyzny.
- Wiemy, że gdy obserwacja $\boldsymbol{x}_i$ leży na hiperpłaszczyźnie to:

$$h(\boldsymbol{x}_i) = \boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0 = 0$$

- co daje:

$$y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) > 0 \begin{cases} h(\boldsymbol{x}_i) < 0 & \text{jeżeli } y_i = -1 \\ h(\boldsymbol{x}_i) > 0 & \text{jeżeli } y_i = +1 \end{cases}$$

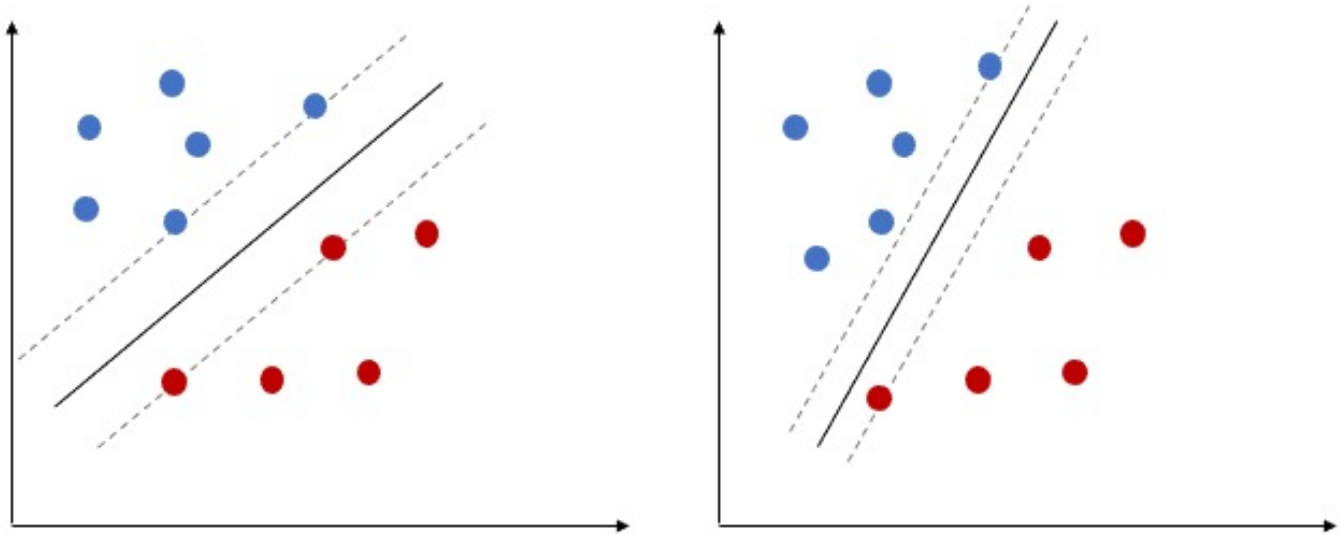
- Powyższe równanie pomoże znaleźć jakość klasyfikatora, jednak musimy znaleźć najlepszą hiperpłaszczyznę.
- Jeśli dane można doskonale oddzielić hiperpłaszczyzną, wówczas może istnieć nieskończona liczba możliwych hiperpłaszczyzn, ponieważ dana rozdzielająca hiperpłaszczyzna może być przesuwana / obracana.
- Dlatego musimy znaleźć sposób, aby zdecydować, którą hiperpłaszczyznę użyć.



Subsection 3

Margines

- Margines można zdefiniować za pomocą minimalnej odległości (odległość normalna) od każdej obserwacji do danej oddzielającej hiperpłaszczyzny.



- Rozmiar marginesu określa moc klasyfikatora, dlatego preferowany jest najszerszy margines.

- Biorąc pod uwagę przykład szkoleniowy $\{\boldsymbol{x}_i, y_i\}$, margines będzie wyglądał następująco:

$$y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) = \gamma_i$$

- Zamiast wartości nieco większej niż 0, definiuję się wartość marginesu za pomocą γ.

- Teraz skorzystajmy z pomocy geometrii, aby znaleźć równanie marginesu.
- Obliczymy odległość od każdego punktu $\{\boldsymbol{x}_i, y_i\}$ do hiperpłaszczyzny.
- Minimalna odległość od punktu $\{\boldsymbol{x}_i, y_i\}$ do hiperpłaszczyzny to odległość normalna, wynosząca γ_i.
- Wiemy już, że $\boldsymbol{\theta}$ jest prostopadła (pod kątem 90 stopni) do oddzielającej hiperpłaszczyzny h.
- Stąd kierunek $\boldsymbol{\theta}$ można uzyskać za pomocą wektora jednostkowego $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}$.
- Ponieważ odcinek $(h, \{\boldsymbol{x}_i, y_i\})$ będzie mieć ten sam kierunek co $\boldsymbol{\theta}$, możemy obliczyć zdefiniowany punkt O na h jako:

$$O = \boldsymbol{x}_i - \gamma_i \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}$$

- Ponieważ punkt O leży na oddzielającej hiperpłaszczyźnie, możemy zapisać następujące równanie jako:

$$\boldsymbol{\theta}^T \left(\boldsymbol{x}_i - \gamma_i \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \right) + \theta_0 = 0$$

$$\boldsymbol{\theta}^T \boldsymbol{x}_i - \gamma_i \frac{\boldsymbol{\theta}^T \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} + \theta_0 = 0$$

$$\gamma_i \frac{\boldsymbol{\theta}^T \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} = \boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0$$

- Możemy wyrazić $\boldsymbol{\theta}$ za pomocą następującego równania:

$$\|\boldsymbol{\theta}\| = \sum_{j=1}^{p} \theta_j^2 = \sqrt{\boldsymbol{\theta}^T \boldsymbol{\theta}}$$

$$\boldsymbol{\theta}^T \boldsymbol{\theta} = \|\boldsymbol{\theta}\|^2$$

- Dlatego możemy napisać:

$$\gamma_i \frac{\|\boldsymbol{\theta}\|^2}{\|\boldsymbol{\theta}\|} = \boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0$$

$$\gamma_i \|\boldsymbol{\theta}\| = \boldsymbol{\theta}^T \boldsymbol{x}_i + \boldsymbol{\theta}^T \theta_0$$

$$\gamma_i = \left(\boldsymbol{x}_i \frac{\boldsymbol{\theta}^T}{\|\boldsymbol{\theta}\|} + \frac{\theta_0}{\|\boldsymbol{\theta}\|} \right)$$

- Wreszcie, możemy połączyć pozytywne i negatywne przykłady w jednym równaniu:

$$\gamma_i = y_i \left(\boldsymbol{x}_i \frac{\boldsymbol{\theta}^T}{\|\boldsymbol{\theta}\|} + \frac{\theta_0}{\|\boldsymbol{\theta}\|} \right)$$

- Biorąc pod uwagę zbiór danych treningowych $\{\boldsymbol{x}_i, y_i\}$ i oddzielającą hiperpłaszczyznę, możemy znaleźć odległość między każdym punktem a hiperpłaszczyzną jako:

$$\gamma_i = \frac{y_i h(\boldsymbol{x}_i)}{\|\boldsymbol{\theta}\|} = \frac{y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0)}{\|\boldsymbol{\theta}\|}$$

- We wszystkich n punktach definiujemy margines klasyfikatora liniowego jako minimalną odległość punktu od oddzielającej hiperpłaszczyzny:

$$\gamma_i^* = \min_{\boldsymbol{x}_i}\{\frac{y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0)}{\|\boldsymbol{\theta}\|}\}$$

- Wszystkie punkty, które osiągają tę minimalną odległość, nazywane są wektorami nośnymi $\{\boldsymbol{x}^*, y^*\}$ dla hiperpłaszczyzny:

$$|\boldsymbol{\theta}^T \boldsymbol{x} + \theta_0| = 1$$

- więc:

$$\gamma_i^* = \frac{y_i^*(\boldsymbol{\theta}^T \boldsymbol{x}^* + \theta_0)}{\|\boldsymbol{\theta}\|} = \frac{1}{\|\boldsymbol{\theta}\|}$$

- Oznacza to, że minimalna odległość między dwiema klasami będzie wynosić co najmniej $\frac{2}{\|\boldsymbol{\theta}\|}$

- Dla każdego wektora nośnego $\boldsymbol{x}_i^*$ mamy $y_i^* h(\boldsymbol{x}^*) = 1$, a dla wszystkich innych punktów, które nie są wektorem nośnym, możemy zdefiniować jedno złożone równanie:

$$y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) \geq 1$$

- Zatem funkcja celu dla liniowej SVM z twardym marginesem ma za zadanie znalezienie optymalnej hiperpłaszczyzny h^*:

$$h^* = \max_h \{\gamma_h^*\} = \max_{\boldsymbol{\theta},\theta_0} \{\frac{1}{\|\boldsymbol{\theta}\|}\}$$

- przy ograniczeniach:

$$y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) \geq 1$$

- Zamiast maksymalizować margines $\frac{1}{\|\boldsymbol{\theta}\|}$ możemy również zminimalizować $\boldsymbol{\theta}$.
- Jednak $\boldsymbol{\theta} = 1$ jest niewypukłym problemem optymalizacji, co oznacza, że może istnieć wiele lokalnych optymalizacji.

- Stąd możemy zastosować następujące sformułowanie minimalizujące (które ma charakter wypukły):

$$h^* = \min_{\boldsymbol{\theta},\theta_0}\{\frac{\|\boldsymbol{\theta}^2\|}{2}\}$$

- przy ograniczeniach:

$$y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) \geq 1$$

- Ta funkcja celu jest pierwotnym problemem optymalizacji wypukłej i można ją rozwiązać za pomocą ograniczeń liniowych przy użyciu standardowych algorytmów.

- Do tej pory zakładaliśmy, że zbiór danych można rozdzielić liniowo, co tak naprawdę nie ma miejsca w prawdziwym scenariuszu.
- Stąd spójrzmy na nieco bardziej skomplikowany przypadek.
- Nadal pracujemy nad liniową SVM, jednak tym razem niektóre klasy nakładają się na siebie w taki sposób, że idealne rozdzielenie jest niemożliwe, ale dane nadal można rozdzielić liniowo.

- Możemy rozwiązać ten problem wprowadzając nową zmienną ξ, a następnie przedefiniowując nasze ograniczenie nierówności jako:

$$y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) \geq 1 - \xi_i$$

- Funcka celu przyjmuje postać:

$$h^* = \min_{\boldsymbol{\theta},\theta_0} \{ \frac{\|\boldsymbol{\theta}^2\|}{2} + C \sum_{i=1}^{n} \xi_i^k \}$$

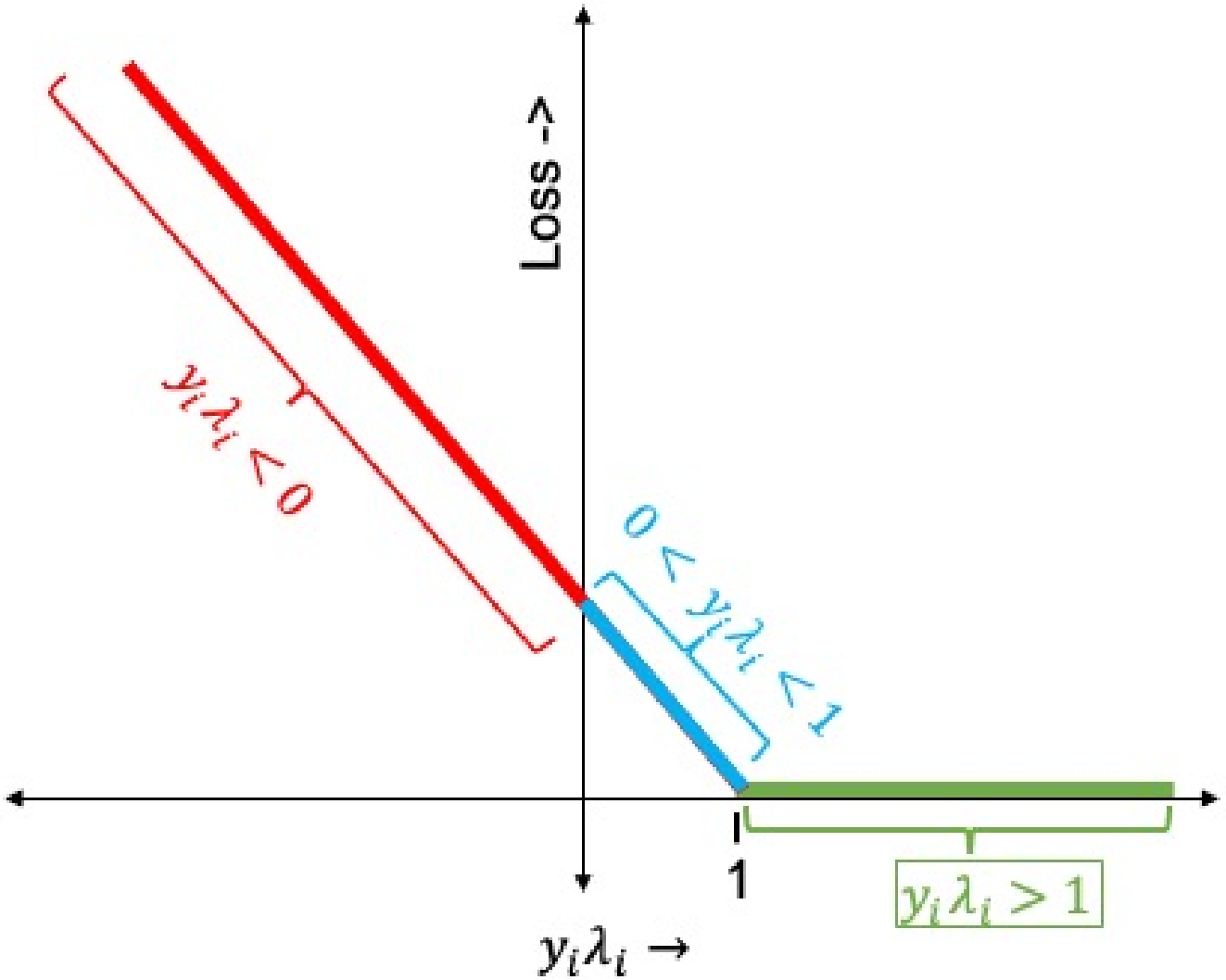
- gdzie C i k to stałe, które równoważą koszt błędnej klasyfikacji.

- HyperParameter C jest również nazywany stałą regularyzacją.
- Jeśli $C \to 0$, to strata wynosi zero i staramy się zmaksymalizować margines.
- Jeśli $C \to \infty$, to margines nie ma żadnego wpływu, a funkcja celu stara się po prostu zminimalizować stratę.
- k jest zwykle ustawione na 1 lub 2.
- Jeśli $k = 1$, to strata nazywana jest stratą zawiasową (hinge loss), a jeśli $k = 2$, to strata kwadratowa.

Subsection 4

Optymalizacja

- Wykorzystanie algorytmu gradientowego jest możliwe wtedy kiedy wykorzystujemy hinge loss.
- Zdefiniujmy λ_i jako decyzję $(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0)$, wtedy:
- Jeśli $(y_i \lambda_i) > 1$, to klasyfikator prawidłowo przewiduje znak (oba mają ten sam znak), a $\boldsymbol{x}_i$ jest daleko od marginesu, stąd nie ma kary / straty.
- Jeśli y_i i λ_i mają ten sam znak, ale $\lambda_i < 1$, wtedy $\boldsymbol{x}_i$ jest pomiędzy marginesem a hiperpłaszczyzną. Zdarzenie, mimo że klasyfikator prawidłowo przewiduje znak, będzie pewna kara / przegrana. Co więcej, kara rośnie, gdy $\boldsymbol{x}_i$ zbliża się do hiperpłaszczyzny.
- Jeśli y_i i λ_i mają różne znaki, wówczas kara będzie duża i gdy $\boldsymbol{x}_i$ odsunie się od granicy po złej stronie, kara / strata będzie rosła liniowo.


Loss ->
$y_i\lambda_i < 0$
$0 < y_i\lambda_i < 1$
1
$y_i\lambda_i > 1$
$y_i\lambda_i \rightarrow$

- Ponieważ nie potrzebujemy żadnej straty, gdy $(y_i \lambda_i) > 1$, możemy zmienić ograniczenie nierówności na ograniczenie równości, przepisując równanie $y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) \geq 1 - \xi_i$ w następujący sposób:

$$\xi_i = \max(0, 1 - y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0))$$

- Możemy włączyć to bezpośrednio do samej funkcji celu i obliczyć funkcję straty jako:

$$L = \frac{\|\boldsymbol{\theta}^2\|}{2} + C \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0))$$

- Aby znaleźć minimum funkcji, musimy wziąć pochodną funkcji:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta} - C \sum_{i=1,\xi_i \geq 1}^{n} y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial \theta_0} = -C \sum_{i=1,\xi_i \geq 1}^{n} y_i$$

Section 2

Literatura

- *Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.*