Przegląd modeli i narzędzi uczenia maszynowego – wersja robocza

Przemysław Jaśko

Uniwersytet Ekonomiczny w Krakowie

14 stycznia 2017

Estymacja parametrów modeli regresji z wykorzystaniem regularyzacji

Modele regresji liniowej z regularyzacją

Rozważmy liniowy model regresji:

$$\mathbf{y} = \mathbf{X}\beta + \xi$$

przy założeniach:

- $\xi \sim \mathbf{N}^n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, gdzie $\mathbf{0}$ jest n-elementowym wektorem zer, $\mathbf{I}_n$ macierzą jednostkową stopnia n, a $\sigma^2 > 0$ dodatnim skalarem
- $E(\xi|\mathbf{X}) = E(\xi) = \mathbf{0}$
- $Cov(\xi|\mathbf{X}) = Cov(\xi) = \sigma^2\mathbf{I}_n$

Stąd: $\hat{\mathbf{y}} = E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta$
Normę L_p $(p = 1, 2, \ldots)$ wektora $\mathbf{z} = [z_1 \ldots z_M]'$ definiuje się następująco:

$$||\mathbf{z}||_p = \left[\sum_{m=1}^{M} |z_m|^p\right]^{\frac{1}{p}}$$

Dla $p = \infty$ mamy: $||\mathbf{z}||_\infty = \max_{1 \leqslant m \leqslant M} |z_m|$.

Kryteria optymalizacyjne dla regresji z regularyzacją

Regresja LASSO (*least absolute shrinkage and selection operator*):

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$$

przy czym hiperparametr $\lambda \geqslant 0$
Regresja grzbietowa (*ridge regression*):

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2$$

przy czym hiperparametr $\lambda \geqslant 0$
Regresja z regularyzacją *elasticnet*:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2$$

przy czym hiperparametry $\lambda_1, \lambda_2 \geqslant 0$

Podane problemy optymalizacyjne można wyrazić równoważnie jako problemy optymalizacyjne z warunkami ograniczającymi w postaci nierówności:
Regresja LASSO:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in D} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$$

gdzie

$$D = \{\boldsymbol{\beta} \in \mathbb{R}^k : ||\boldsymbol{\beta}||_1 \leqslant t\}$$

Istnieje jednoznaczna zależność pomiędzy hiperparametrami λ a t.
Regresja grzbietowa:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in D} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$$

gdzie

$$D = \{\boldsymbol{\beta} \in \mathbb{R}^k : ||\boldsymbol{\beta}||_2^2 \leqslant t\}$$

Istnieje jednoznaczna zależność pomiędzy hiperparametrami λ a t.

- W modelach regresji z regularyzacją dokonuje się uprzedniej standaryzacji wartości zmiennych objaśniających (wartości kolumn macierzy $\mathbf{X}$), żeby parametry będące składowymi wektora parametrów β, charakteryzowały się tymi samymi rzędami wielkości. Zabieg ten ma celu zapobiegnięcie sytuacji, w której wartość normy wektora β zostaje zdominowana przez parametry odnoszące się do zmiennej objaśniającej (zmiennych objaśniających) przyjmującej wartości charakteryzujące się niższym rzędem wielkości względem tych dla pozostałych zmiennych.
- Estymator regresji grzbietowej (regularyzacja L_2) zadany jest analitycznie.
- W regresji LASSO (regularyzacja L_1) celem określenia wartości wektora oszacowań parametrów $\hat{\beta}$ wykorzystuje się procedury numeryczne.

Dla regresji grzbietowej funkcja kryterium (z regularyzacją L_2), przy $\lambda \geqslant 0$ wyrażona jest następująco:
$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} =$
$\mathbf{y}'\mathbf{y} - 2\mathbf{X}'\mathbf{y}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{y}'\mathbf{y} - 2\mathbf{X}'\mathbf{y}\boldsymbol{\beta} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)\boldsymbol{\beta}$,
gdzie $\mathbf{I}_k$ jest macierzą jednostkową stopnia k
Jako, że funkcja celu jest wypukła w $\mathbb{R}^k$, WKW (warunkiem koniecznym i wystarczającym) istnienia minimum globalnego funkcji kryterium jest zerowanie się jej gradientu:

$$-2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)\boldsymbol{\beta} = 0 \Leftrightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)^{-1}\mathbf{X}'\mathbf{y}$$

LASSO (regularyzacja L_1)

- procedura zmierza w kierunku wyzerowania wybranych elementów wektora oszacowań parametrów $\hat{\boldsymbol{\beta}}$

Regresja grzbietowa (regularyzacja L_2)

- w przypadku liniowo zależnych kolumn macierzy zmiennych objaśniających zachodzi brak identyfikowalności parametrów dla estymatora KMNK (macierz $\mathbf{X}'\mathbf{X}$ jest osobliwa). W przypadku wprowadzenia do funkcji kryterium regularyzacji L_2, problem znika, gdyż macierz $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)$ jest dodatnio określona dla dowolnego $\lambda > 0$.

Nieparametryczne modele regresji

Nieparametryczne modele regresji

- regresja LOESS / LOWESS
- regresja jądrowa Nadarayi–Watsona
- regresja lokalna wielomianowa
- regresja splajnowa / model MARS (*Multivariate Adaptive Regression Splines*)

Modele regresji lokalnej: LOESS, LOWESS

- LOWESS (*LOcally WEighted Scatterplot Smoothing*)
- LOESS (*LOcal regrESSion*)

Dla kolejno rozważanych wektorów $\mathbf{x}$ buduje się lokalny model regresji y w oparciu o podzbiór obserwacji z próby, dla których $\mathbf{x}_i$ należy do pewnego otoczenia $\mathbf{x}$. Parametry modelu lokalnego $\beta(\mathbf{x})$ szacuje się ważoną metodą najmniejszych kwadratów (WMNK) z funkcją wag, której wartość dla danej obserwacji próby $\mathbf{x}_i$ zależy od jej odległości od $\mathbf{x}$.
W LOESS stosuje się funkcję wag (*tri-cube weight function*):

$$w(u) = (1 - |u|^3)^3 I[|u| < 1]$$

gdzie I jest funkcją wskaźnikową
Przyjmuje się, że $u = \frac{d(\mathbf{x},\mathbf{x}_i)}{d_{\max}(\mathbf{x})}$, gdzie d jest funkcją odległości,

$$d_{\max}(\mathbf{x}) = \begin{cases} d_{(\lfloor \alpha n \rfloor)}(\mathbf{x}) & \text{, gdy } 0 < \alpha \leqslant 1 \\ \alpha^{\frac{1}{k}} d_{(n)}(\mathbf{x}) & \text{, gdy } \alpha > 1 \end{cases}$$

, natomiast $d_{(r)}(\mathbf{x})$ jest r-tą statystyką porządkową dla odległości $d(\mathbf{x}, \mathbf{x}_i)$, $i = 1, \ldots, n$.

Tak więc funkcja wag dla pary $(\mathbf{x}, \mathbf{x}_i)$ wyrażona jest przez:

$$w(\mathbf{x}, \mathbf{x}_i) = \left(1 - \left|\frac{d(\mathbf{x}, \mathbf{x}_i)}{d_{\max}(\mathbf{x})}\right|^3\right)^3 I\left[\left|\frac{d(\mathbf{x}, \mathbf{x}_i)}{d_{\max}(\mathbf{x})}\right| < 1\right]$$

gdzie I jest funkcją wskaźnikową
Niech $\{\mathbf{x_1}, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ będzie n-elementową próbą wektorów $k \times 1$ zmiennych objaśniających oraz $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1\, \tilde{\mathbf{x}}_2\, \ldots\, \tilde{\mathbf{x}}_n]'$, przy czym za wektory $\tilde{\mathbf{x}}_i$, $i = 1, \ldots, n$ przyjmuje się najczęściej:

- $\tilde{\mathbf{x}}_i = [1|\mathbf{x}_i']'$ – lokalna regresja liniowa
- $\tilde{\mathbf{x}}_i = [1|\mathbf{x}_i'|(\mathbf{x}_i^2)']'$ – regresja lokalna wielomianowa drugiego stopnia
- $\tilde{\mathbf{x}}_i = [1|\mathbf{x}_i'|(\mathbf{x}_i^2)'|(\mathbf{x}_i^3)']'$ – regresja lokalna wielomianowa trzeciego stopnia.

Analogicznie określa się wektor $\tilde{\mathbf{x}}$ w oparciu o $\mathbf{x}$.

Macierz wag dla wektora $\mathbf{x}$ zadana jest następująco:

$$\mathbf{W}(\mathbf{x}) = \mathrm{diag}\{w(\mathbf{x}, \mathbf{x}_i),\ i = 1, \ldots, n\}$$

Estymator WMNK modelu lokalnego dla $\mathbf{x}$ określa:

$$\hat{\beta}(\mathbf{x}) = [\tilde{\mathbf{X}}'\mathbf{W}(\mathbf{x})\tilde{\mathbf{X}}]^{-1}\tilde{\mathbf{X}}'\mathbf{W}(\tilde{\mathbf{X}})\mathbf{y}$$

Wartość dopasowanej krzywej regresji dla wektora $\mathbf{x}$ wyznacza się jako:

$$\hat{y}(\mathbf{x}) = \tilde{\mathbf{x}}'\hat{\beta}(\mathbf{x})$$

W celu wyznaczenia (przybliżonego) przebiegu krzywej regresji za $\mathbf{x}$ przyjmuje się wartości ze zbioru określonego przez odpowiednio gęstą siatkę punktów należących do $\mathcal{X} \subset \mathbb{R}^k$, który to zbiór wyznacza się biorąc pod uwagę przedział zmienności wartości $\mathbf{x}_i$ obserwowanych w próbie.

Model regresji jądrowej Nadarayi–Watsona (*Nadaraya–Watson kernel regression*)

Funkcją jądrową nazywa się funkcję $K : \mathbb{R} \to [0, +\infty)$ spełniającą następujące warunki:

- $K(u) \geqslant 0$ (nieujemność)
- $\int_{-\infty}^{+\infty} K(u)\,\mathrm{d}u = 1$
- $\forall u \in \mathbb{R} :\ K(u) = K(-u)$ (symetria)

Dodatkowo zakłada się, że zero jest słabym maksimum globalnym funkcji K, tzn. $\forall u \in \mathbb{R} :\ K(u) \leqslant K(0)$.
Niech $h > 0$ będzie tzw. parametrem wygładzania oraz $u = \frac{y - y_i}{h} \Leftrightarrow \mathrm{d}u = \frac{\mathrm{d}y}{h}$, stąd

$$\int_{-\infty}^{+\infty} K(u)\,\mathrm{d}u = 1 \Leftrightarrow \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{y - y_i}{h}\right)\,\mathrm{d}y = 1$$

Tym samym $\frac{1}{h}K\left(\frac{y-y_i}{h}\right)$ jest jądrem dla $\frac{y-y_i}{h}$.

Z definicji funkcji jądrowej wynika, że:

$$\int_{-\infty}^{+\infty} uK(u)\,\mathrm{d}u = \int_{0}^{+\infty}[-uK(-u) + uK(u)]\,\mathrm{d}u \overset{K(-u)=K(u)}{=}$$
$$-\int_{0}^{+\infty} uK(u)\,\mathrm{d}u + \int_{0}^{+\infty} uK(u)\,\mathrm{d}u = 0$$

Dla $u = \frac{y-y_i}{h}$ zachodzi:

$$\int_{-\infty}^{+\infty} uK(u)\,\mathrm{d}u = \frac{1}{h}\int_{-\infty}^{+\infty} \frac{y-y_i}{h} K(\frac{y-y_i}{h})\,\mathrm{d}y = 0 \Leftrightarrow$$
$$\int_{-\infty}^{+\infty} y\frac{1}{h}K(\frac{y-y_i}{h})\,\mathrm{d}y = y_i \underbrace{\frac{1}{h}\int_{-\infty}^{+\infty} K(\frac{y-y_i}{h})\,\mathrm{d}y}_{=1} = y_i$$

Z wykorzystaniem funkcji jądrowych określa się estymator jądrowy funkcji gęstości dla rozkładu jednowymiarowej zmiennej X:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Estymator jądrowy łącznej funkcji gęstości dla rozkładu dwuwymiarowej zmiennej (X, Y) zadany jest przez:

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right)$$

Oszacowanie funkcji gęstości rozkładu warunkowego y względem $X = x$ wyznacza się jako:

$$\hat{f}(y|x) = \frac{\hat{f}(x, y)}{\hat{f}(x)} = \frac{1}{h} \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

Uogólniając, estymator jądrowy łącznej funkcji gęstości rozkładu k-wymiarowej zmiennej $(X_1, X_2, \ldots, X_k)$ dany jest przez:

$$\hat{f}(x_1, x_2, \ldots, x_k) = \frac{1}{nh^k} \sum_{i=1}^{n} \prod_{j=1}^{k} K\left(\frac{x_j - x_{ij}}{h}\right)$$

Wartość oczekiwana zmienna zależnej Y warunkowa względem wartości (wektora) zmiennych zależnych $X = x$ określona jest przez:

$$E(Y|X = x) = \int_{-\infty}^{+\infty} yf(y|x)\,\mathrm{d}y = \int_{-\infty}^{+\infty} y\frac{f(x,y)}{f(x)}\,\mathrm{d}y$$

Regresja Nadarayi–Watsona jest nieparametrycznym oszacowaniem $E(Y|X = x)$.
Regresja jądrowa Nadarayi–Watsona:
$E(Y|X = x) \approx \hat{y} = \int_{-\infty}^{+\infty} y\frac{\hat{f}(x,y)}{\hat{f}(x)}\,\mathrm{d}y =$
$\int_{-\infty}^{+\infty} y\frac{\frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)K\left(\frac{y-y_i}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}\,\mathrm{d}y =$
$\frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\int_{-\infty}^{+\infty} y\frac{1}{h}K\left(\frac{y-y_i}{h}\right)\mathrm{d}y}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$
Ostatnia równość wynika z wcześniej wykazanej:
$\int_{-\infty}^{+\infty} y\frac{1}{h}K(\frac{y-y_i}{h})\,\mathrm{d}y = y_i$

Mamy więc dla regresji NW:

$$E(Y|X=x) \approx \hat{y} = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

Przykłady funkcji jądrowych:

- jądro gaussowskie: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$
- jądro Epanecznikowa: $K(u) = \frac{3}{4}(1-u^2) I_{\{|u| \leqslant 1\}}$, gdzie I jest funkcją wskaźnikową
- jądro trójkątne: $K(u) = (1-|u|) I_{\{|u| \leqslant 1\}}$
- jądro jednostajne: $K(u) = \frac{1}{2} I_{\{|u| \leqslant 1\}}$

Regresja kwantylowa Koenkera

Oszacowanie wartości kwantyla $q^{(\tau)}$, $0 < \tau < 1$ rozkładu zmiennej y, w oparciu o próbę $\mathbf{y} = \{y_1, \ldots, y_n\}$ uzyskuje się jako rozwiązanie problemu minimalizacji:

$$\hat{q}^{(\tau)} = \text{argmin}_{q \in \mathbb{R}} \sum_{i=1}^{n} \{(1-\tau)\,[y_i - q]_- + \tau\,[y_i - q]_+\}$$

przy czym funkcje:
$[z]_+ = z$, gdy $z > 0$ oraz $[z]_+ = 0$ w przeciwnym razie
$[z]_- = -z$, gdy $z < 0$ oraz $[z]_- = 0$ w przeciwnym razie

Regresja kwantylowa Koenkera

Zakładając, że wartość kwantyla $q_i^{(\tau)}$, $0 < \tau < 1$ rozkładu warunkowego zmiennej zależnej $(y_i|\mathbf{x}_i)$ jest zależna od liniowej kombinacji wartości zmiennych objaśniających: $q_i^{(\tau)} = \sum_{j=1}^{k} \beta_j^{(\tau)} x_{ij}$, $i = 1, 2, \ldots, n$, wartość oszacowania $\hat{q}_i^{(\tau)}$ uzyskuje się wykorzystując rozwiązanie następującego problemu minimalizacji:

$$\hat{\boldsymbol{\beta}}^{(\tau)} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^{n} \{(1-\tau) \left[y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right]_- + \tau \left[y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right]_+ \}$$

Dla $\tau = 0,5$ odnosimy się do mediany, a powyższe kryterium jest równoważne kryterium LAD (*Least Absolute Deviation*):

$$\operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right|$$

Estymator LAD jest mniej efektywnym, lecz bardziej odpornym niż KMNK estymatorem parametrów modelu regresji liniowej.

Model SVM w klasyfikacji wzorcowej

SVM (*Support Vector Machine*)

Rozpatrzmy problem klasyfikacji wzorcowej z binarną zmienną zależną, taką, że:

$$y_i \in \{-1, 1\}$$

Dla zbiorów danych $(\mathbf{y}, \mathbf{X})$ możliwe są trzy przypadki separowalności klas $y \in \{-1, 1\}$ w przestrzeni wartości zmiennych objaśniających $\mathbf{X}$:

- **pełna liniowa separowalność klas** – istnieje podział przestrzeni $\mathbf{X}$ w oparciu o hiperpłaszczynę H na podzbiory odnoszące się do klas, skutkujący brakiem błędnych przewidywań dotyczących przynależności klasowej obiektów
- **niepełna liniowa separowalność klas** – istnieje podział przestrzeni $\mathbf{X}$ w oparciu o hiperpłaszczynę H na podzbiory odnoszące się do klas, jednak wiąże się z nim możliwość występowania błędnych przewidywań dotyczących przynależności klasowej obiektów

- **nieliniowa separowalność klas** – w wyjściowej przestrzeni wartości zmiennych objaśniających nie istnieje hiperpłaszczyzna H umożliwiająca separację obiektów odmiennych klas, w przekształconej przestrzeni o wyższym wymiarze (możliwie nieskończenie wielowymiarowej) powstałej jako odpowiednie nieliniowe odwzorowanie $\mathbf{X}$ istnieje możliwość określenia hiperpłaszczyzny separującej klasy (formalna postać odwzorowania nie musi być znana, w związku z możliwością zastosowania tzw. tricku jądrowego $<$*kernel trick*$>$)

Model SVM przy pełnej liniowej separowalności klas

Hiperpłaszczyzna H separująca klasy w przestrzeni wartości zmiennych objaśniających $\mathbf{X}$ zadana jest przez:

$$\mathbf{w}'\mathbf{x} + b = 0$$

gdzie $\mathbf{w}$ jest wektorem normalnym hiperpłaszczyzny separującej
Ograniczenia dla przypadku pełnej liniowej separowalności zbioru danych są następujące:
Dla $i = 1, 2, \ldots, n$:

$$\begin{cases} \mathbf{w}'\mathbf{x}_i + b \geqslant 1, & \text{gdy } y_i = 1 \\ \mathbf{w}'\mathbf{x}_i + b \leqslant -1, & \text{gdy } y_i = -1 \end{cases}$$

Co jest równoważne:
$y_i(\mathbf{w}'\mathbf{x}_i + b) - 1 \geqslant 0$, dla każdego $i = 1, 2, \ldots, n$

$H_1 : \mathbf{w}'\mathbf{x}_i + b = 1$

odległość hiperpłaszczyzny H_1 od początku układu współrzędnych $(0, 0)$: $\frac{|1-b|}{||\mathbf{w}||}$

$H_2 : \mathbf{w}'\mathbf{x}_i + b = -1$

odległość hiperpłaszczyzny H_2 od początku układu współrzędnych $(0, 0)$: $\frac{|-1-b|}{||\mathbf{w}||}$

Odległość pomiędzy hieprpłaszczyznami H_1 i H_2:
Niech: $c_1, c_2 \in \mathbb{R}$, $\mathbf{w}$ to wektor normalny dla hiperpłaszczyzn H_1, H_2,
$\mathbf{x}_1 = c_1 \cdot \mathbf{w} \in H_1$ oraz $\mathbf{x}_2 = c_2 \cdot \mathbf{w} \in H_2$, w tej sytuacji wartość $||\mathbf{x}_2 - \mathbf{x}_1||$ będzie odpowiadać odległości H_1 od H_2, stąd:
$||\mathbf{x}_2 - \mathbf{x}_1|| = ||(c_2 - c_1) \cdot \mathbf{w}|| = |c_2 - c_1|||\mathbf{w}||$
oraz
$\mathbf{w}'\mathbf{x}_1 + b - 1 = 0$
$\mathbf{w}'\mathbf{x}_2 + b + 1 = 0$
Odejmując od siebie równania uzyskujemy:
$\mathbf{w}'(\mathbf{x}_2 - \mathbf{x}_1) + 2 = 0 \Leftrightarrow \mathbf{w}'(\mathbf{x}_2 - \mathbf{x}_1) = -2 \Leftrightarrow \mathbf{w}'\mathbf{w} \cdot (c_2 - c_1) = -2 \Leftrightarrow ||\mathbf{w}||^2 \cdot (c_2 - c_1) = -2 \Leftrightarrow c_2 - c_1 = \frac{-2}{||\mathbf{w}||^2} \Leftrightarrow |c_2 - c_1| = \frac{2}{||\mathbf{w}||^2}$
W związku z powyższym:
$||\mathbf{x}_2 - \mathbf{x}_1|| = |c_2 - c_1|||\mathbf{w}|| = \frac{2}{||\mathbf{w}||^2} \cdot ||\mathbf{w}|| = \frac{2}{||\mathbf{w}||}$
Odległość pomiędzy hiperpłaszczyznami H_1 i H_2 zadana przez $\frac{2}{||\mathbf{w}||}$ podlega maksymalizacji przy nałożonych ograniczeniach:
$y_i(\mathbf{w}'\mathbf{x}_i + b) - 1 \geqslant 0,\ i = 1, 2, \ldots, n.$

Tak więc mamy następujący problem optymalizacyjny:

$$\max_{\mathbf{w}\in\mathbb{R}^k} \frac{2}{||\mathbf{w}||}$$

równoważnie

$$\min_{\mathbf{w}\in\mathbb{R}^k} \frac{||\mathbf{w}||^2}{2}$$

pod warunkiem $y_i(w'\mathbf{x}_i + b) - 1 \geqslant 0,\ i = 1, 2, \ldots, n$
Jest to problem optymalizacyjny z kwadratową funkcją celu oraz warunkami ograniczającymi w postaci nierówności.
Dla powyższego problemu **funkcja Lagrange'a** określona jest następująco:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}'\mathbf{w} - \sum_{i=1}^{n} \alpha_i[(\mathbf{w}'\mathbf{x}_i + b)y_i - 1]$$

przy czym $\alpha_i \geqslant 0,\ i = 1, \ldots, n$
Funkcja Lagrange'a L podlega minimalizacji względem $\mathbf{w}, b$ oraz maksymalizacji względem $\boldsymbol{\alpha}$

Z twierdzenia Karusha–Kuhna–Tuckera (KKT) warunkiem koniecznym istnienia punktu siodłowego dla funkcji Lagrange'a L jest:

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = 0 \\ \frac{\partial L}{\partial b} = 0 \\ \alpha_i[(\mathbf{w}'\mathbf{x}_i + b)y_i - 1] = 0,\ i = 1, \ldots, n \end{cases}$$

Tak więc:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

Zerowanie się gradientu funkcji Lagrange'a prowadzi do warunków:

$$\begin{cases} \mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Po podstawieniu w funkcji L za zmienną $\mathbf{w}$ powyższego wyrażenia oraz uwzględnieniu drugiego z równań, uzyskujemy podlegającą maksymalizacji względem $\boldsymbol{\alpha}$ funkcję W:

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x_i}' \mathbf{x}_j$$

przy warunkach ograniczających:

$$\alpha_i \geqslant 0,\ i = 1, \ldots, n \ \wedge \ \sum_{i=1}^{n} \alpha_i y_i = 0$$

Postać funkcji W uzyskujemy z wykorzystaniem wspomnianych podstawień poprzez:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}'\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left[(\mathbf{w}'\mathbf{x}_i + b)y_i - 1\right] =$$
$$\frac{1}{2}\left(\sum_{i=1}^{n} y_i\alpha_i\mathbf{x}_i\right)'\left(\sum_{i=1}^{n} y_i\alpha_i\mathbf{x}_i\right) - \sum_{i=1}^{n} \alpha_i\left(\sum_{i=1}^{n} y_i\alpha_i\mathbf{x}_i\right)'\mathbf{x}_i y_i -$$
$$b\underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} + \sum_{i=1}^{n} \alpha_i =$$
$$\frac{1}{2}\sum_{i=1}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i'\mathbf{x}_j - \sum_{i=1}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i'\mathbf{x}_j + \sum_{i=1}^{n} \alpha_i =$$
$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i'\mathbf{x}_j = W(\boldsymbol{\alpha})$$

Jako rozwiązanie problemu warunkowej maksymalizacji funkcji W uzyskujemy wartości wektora mnożników Lagrange'a $\boldsymbol{\alpha}^* = [\alpha_1^*, \ldots, \alpha_n^*]'$, przy czym $\alpha_i^* > 0$, gdy $i \in SV$ oraz $\alpha_i = 0$, gdy $i \notin SV$, gdzie SV jest zbiorem indeksów wektorów zbioru danych, będących wektorami wspierającymi.

Biorąc pod uwagę, że $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$ oraz przyjmując α_i^* jako wartości α_i, $i = 1, \ldots, n$, hiperpłaszczyzna separująca $H : \mathbf{w}'\mathbf{x} + b = 0$ zadana jest przez:

$$H : \sum_{i \in SV} y_i \alpha_i^* (\mathbf{x}_i' \mathbf{x}) + b^* = 0$$

gdzie $b^* = \frac{1}{2}[\mathbf{w}'\mathbf{x}^*(1) + \mathbf{w}'\mathbf{x}^*(-1)]$, natomiast $\mathbf{x}^*(1)$, $\mathbf{x}^*(-1)$ to odpowiednio dowolny wektor wspierający klasy 1 oraz -1
Przyporządkowanie obiektu do klasy – reguła dyskryminacyjna:

$$\hat{y}|\mathbf{x} = \operatorname{sgn}\left(\sum_{i \in SV} y_i \alpha_i^* (\mathbf{x}_i' \mathbf{x}) + b^* \right)$$

Model SVM przy niepełnej liniowej separowalności klas

Ograniczenia dla przypadku pełnej liniowej separowalności zbioru danych są następujące:
Dla $i = 1, 2, \ldots, n$:

$$\begin{cases} \mathbf{w}'\mathbf{x}_i + b \geqslant 1 - \xi_i, & \text{gdy } y_i = 1 \\ \mathbf{w}'\mathbf{x}_i + b \leqslant -1 + \xi, & \text{gdy } y_i = -1 \\ \xi_i \geqslant 0 \end{cases}$$

Funkcja kryterium podlegająca minimalizacji:

$$\frac{||\mathbf{w}||^2}{2} + C \sum_{i=1}^{n} \xi_i$$

gdzie hiperparametr $C > 0$ nazywany jest współczynnikiem kary

Model SVM przy nieliniowej separowalności klas

Nieliniowe odwzorowanie przestrzeni $\mathbf{X}$ w przestrzeń o wyższym wymiarze, dla którego przyjmujemy oznaczenie $\mathbf{h}$, przyporządkowuje $\mathbf{x} \in \mathbb{R}^k$ wektor $\mathbf{h}(\mathbf{x})$, którego elementy są nieliniowymi przekształceniami elementów wektora wartości pierwotnych zmiennych objaśniających. W przekształconej przestrzeni poszukuje się hiperpłaszczyzny separującej klasy obiektów.

Jednym ze stosowanych przekształceń nieliniowych jest odwzorowanie wielomianowe p-tego stopnia. I tak przekształcenie wielomianowe drugiego stopnia $(p = 2)$ $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}) \dots h_{\tilde{P}}(\mathbf{x})]'$ zadane jest przez:

$$h_r(\mathbf{x}) = \begin{cases} \mathbf{x}_r & \text{dla } r = 1, \dots, k \\ \mathbf{x}_{r-k}^2 & \text{dla } r = k+1, \dots, 2k \\ \mathbf{x}_{\tau(r)} x_{\kappa(r)} & \text{dla } r = 2k+1, \dots, \tilde{p} \end{cases}$$

przy czym funkcje τ oraz κ odpowiedni pierwszy i drugi element z dwuelementowych kombinacji na zbiorze indeksów związanych z pierwotnymi zmiennymi objaśniającymi, liczba tych kombinacji równa jest $\binom{k}{2} = \frac{k(k-1)}{2}$, stąd $\tilde{p} = 2p + \frac{k(k-1)}{2} = \frac{p(p+3)}{2}$

Hiperpłaszczyzna separująca klasy w przekształconej przestrzeni „nowych" zmiennych objaśniających wyrażona jest następująco:

$$H : \sum_{i \in SV} y_i \alpha_i^* (\mathbf{h}(\mathbf{x}_i)' \mathbf{h}(\mathbf{x})) + b^* = 0$$

Dla przekształcenia wielomianowego iloczyn skalarny $\mathbf{h}(\mathbf{x}_i)'\mathbf{h}(\mathbf{x})$ można równoważnie wyrazić za pomocą funkcji jądrowej:

$$K(\mathbf{x}, \mathbf{x}_i) = (1 + \mathbf{x}_i'\mathbf{x})^p$$

Wtedy reguła dyskryminacyjne może być przedstawiona z wykorzystaniem przyjętej funkcji jądrowej:

$$\hat{y}|\mathbf{x} = \text{sgn}\left(\sum_{i \in SV} y_i \alpha_i^* K(\mathbf{x}_i'\mathbf{x}) + b^*\right)$$

W związku z powyższym okazuje się, że formalna postać przekształcenia $\mathbf{h}$ nie musi być znana, gdyż wystarczy wskazać postać funkcji jądrowej $K(\mathbf{x}_i, \mathbf{x})$.
Stosowanymi funkcjami jądrowymi $K(\mathbf{x}_i, \mathbf{x})$ są m.in.:

- uogólnione jądro wielomianowe p-tego stopnia: $(\gamma \mathbf{x}_i'\mathbf{x} + c_0)^p$
- jądro radialne: $\exp(-\gamma||\mathbf{x} - \mathbf{x}_i||^2)$
- jądro sigmoidalne: $\text{tgh}(\gamma \mathbf{x}_i'\mathbf{x} + c_0)$