Programmierblatt 3.

Besprechungswoche: **04.12. – 08.12.2023**

Der Fokus dieses Blattes liegt auf einem beliebten Optimierungsproblem, nämlich dem Klassifizierungsproblem, welches im maschinellen Lernen auftritt. Es ist zum Beispiel möglich, (16×16)-Pixel grosse Schwarzweissbilder als 256-dimensionale Objekte aufzufassen. Diese Bilder sollen dann in verschiedene Klassen eingeteilt werden. Wir werden auf diesem Blatt eine Möglichkeit dazu, die *Support-Vector-Maschine*, üblicherweise mit *SVM* abekürzt, und die zugrundeligenden Algorithmen kennenlernen.

Klassifizierungsprobleme

Gegeben seien Punkte $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{R}^d$ mit zugehörigen Labels $(y_1, \dots, y_N) \in \{-1, 1\}^N$. Dabei sei bekannt, dass die Punkte den zwei verschiedenen Klassen

$$\Omega_+ := \{\mathbf{x}_i \,:\, 1 \le i \le N,\, y_i = +1\}, \qquad \Omega_- := \{\mathbf{x}_i \,:\, 1 \le i \le N,\, y_i = -1\}$$

zugeordnet werden können. Das Ziel ist nun, eine Funktion $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ zu finden, welche der Beziehung $f(\mathbf{x}) > 0$ für alle $\mathbf{x} \in \Omega_+$ und $f(\mathbf{x}) < 0$ für alle $\mathbf{x} \in \Omega_-$ genügt. Im einfachsten Fall genügt dafür eine linear-affine Abbildung, das heisst $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - w_0$, wobei der Separator $\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^d \,:\, f(\mathbf{x}) = 0\}$ eine affine Hyperebene ist.

Offensichtlich wird das Klassifizierungsproblem gelöst, falls $y_i(\mathbf{w}^\top \mathbf{x}_i - w_0) \ge 0$ für $i = 1, \dots, N$ gilt. Sind die zwei Klassen Ω_+ und Ω_- linear separabel, so existieren allerdings mehrere Lösungen. Ein möglicher Ansatz ist daher zu fordern, dass alle Punkte mindestens einen festen Abstand zum Separator $\mathcal{S}$ besitzen sollen. Um die Funktion f so stabil wie möglich zu machen, wird dabei der Vektor $\mathbf{w}$ minimiert. Konkret soll das *Optimierungsproblem unter Nebenbedingungen*

$$\underset{\mathbf{w}}{\text{minimiere}}\ \frac{1}{2}\sum_{i=1}^{d} w_i^2, \quad \text{so dass} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - w_0) \ge 1, \quad i = 1, \dots, N, \tag{1}$$

gelöst werden.

Aktive-Mengen-Strategie

Wir schieben nun eine kurze Erläuterung zur Lösung von Optimierungsproblemen unter Nebenbedingungen ein. Dazu betrachten wir das Problem

$$\underset{\mathbf{x}}{\text{minimiere}}\ \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}, \quad \text{so dass} \quad \mathbf{C}^\top \mathbf{x} = \boldsymbol{\gamma}, \quad \mathbf{D}^\top \mathbf{x} \ge \boldsymbol{\delta}, \tag{2}$$

wobei die letzte Ungleichung komponentweise zu verstehen sei. Wir setzen hier voraus, dass die Matrix $\mathbf{A}$ symmetrisch und positiv definit ist. Der Satz von Karush, Kuhn und Tucker besagt dann, dass für jedes $\mathbf{x}^\star$, welches (2) löst, ein Vektor von Lagrange-Parametern $(\boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ existiert, so dass

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star) = \mathbf{0}, \quad \mathbf{C}^\top \mathbf{x}^\star = \boldsymbol{\gamma}, \quad \mathbf{D}^\top \mathbf{x}^\star \ge \boldsymbol{\delta}, \quad \boldsymbol{\lambda}^\star \ge \mathbf{0}, \quad (\boldsymbol{\lambda}^\star)^\top (\mathbf{D}^\top \mathbf{x}^\star - \boldsymbol{\delta}) = \mathbf{0}. \tag{3}$$

Dabei bezeichnet $\mathcal{L}$ die *Lagrange-Funktion*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) := \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} - \boldsymbol{\mu}^\top (\mathbf{C}^\top \mathbf{x} - \boldsymbol{\gamma}) - \boldsymbol{\lambda}^\top (\mathbf{D}^\top \mathbf{x} - \boldsymbol{\delta}).$$

Die letzte Gleichung in (3) impliziert offensichtlich, dass im Falle $\lambda_i^\star > 0$ automatisch $[\mathbf{D}^\mathsf{T}\mathbf{x}^\star]_i = \delta_i$ gelten muss. In diesem Fall bezeichnen wir die i-te Ungleichheitsnebenbedingung als *aktiv*; die Menge der aktiven Ungleichheitsindizes bezeichnen wir mit Λ_A. Ist dann $\mathbf{D}_A := [\mathbf{d}_i]_{i\in\Lambda_A}$, das heisst, die Spalten aus $\mathbf{D}$ zu aktiven Indizes, so entspricht (3) dem System

$$\begin{bmatrix} \mathbf{A} & -\mathbf{C} & -\mathbf{D}_A \\ -\mathbf{C}^\mathsf{T} & \mathbf{0} & \mathbf{0} \\ -\mathbf{D}_A^\mathsf{T} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^\star \\ \boldsymbol{\mu}^\star \\ \boldsymbol{\lambda}^\star|_{\Lambda_A} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -\boldsymbol{\gamma} \\ -\boldsymbol{\delta}|_{\Lambda_A} \end{bmatrix}. \tag{4}$$

Das Minimierungsproblem (2) kann mit folgender *Aktiven-Mengen-Strategie* gelöst werden:

Algorithmus Aktive-Mengen-Strategie (vgl. [1])

Input: Zulässiger Startvektor $\mathbf{x}_0$, Toleranz `tol` > 0
Output: Lösung des Minimierungsproblems (2)

1: bestimme die zum Startzeitpunkt aktiven Indizes Λ_A
2: **for** $k = 0, 1, \ldots$ **do**
3: Löse

$$\begin{bmatrix} \mathbf{A} & -\mathbf{C} & -\mathbf{D}_A \\ -\mathbf{C}^\mathsf{T} & \mathbf{0} & \mathbf{0} \\ -\mathbf{D}_A^\mathsf{T} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \boldsymbol{\mu} \\ \boldsymbol{\lambda}|_{\Lambda_A} \end{bmatrix} = \begin{bmatrix} \mathbf{b} - \mathbf{A}\mathbf{x}_k \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

4: **if** $\|\mathbf{p}\| <$ `tol` **then**
5: **if** $\lambda_i \geq 0$ für alle $i \geq 1$ **then**
6: breche ab und gebe die Lösung $\mathbf{x}^\star := \mathbf{x}_k$ aus
7: **else**
8: setze $j := \arg\min_{i\in\Lambda_A} \lambda_i$
9: setze $\Lambda_A := \Lambda_A \setminus \{j\}$
10: setze $\mathbf{x}_{k+1} = \mathbf{x}_k$
11: **end if**
12: **else**
13: setze $\alpha := \min \left\{ 1, \min\limits_{\substack{i\notin\Lambda_A \\ \mathbf{d}_i^\mathsf{T}\mathbf{p}<0}} \frac{\delta_i - \mathbf{d}_i^\mathsf{T}\mathbf{x}}{\mathbf{d}_i^\mathsf{T}\mathbf{p}} \right\}$
14: **if** $\exists j \notin \Lambda_A$ so dass $\mathbf{d}_j^\mathsf{T}\mathbf{p} < 0$ und $\alpha = \frac{\delta_j - \mathbf{d}_j^\mathsf{T}\mathbf{x}}{\mathbf{d}_j^\mathsf{T}\mathbf{p}}$ **then**
15: setze $\Lambda_A := \Lambda_A \cup \{j\}$
16: **end if**
17: setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha\mathbf{p}$
18: **end if**
19: **end for**

Aufgabe 1. Schreiben Sie eine Funktion

```
function xs = activeSet(A, b, C, D, delta, x0, tol, maxIter),
```

welche ein Optimierungsproblem der Form (2) mithilfe der Aktiven-Mengen-Strategie löst. Für die Lösung des Gleichungssystems auf Zeile 3 benutzen Sie den \-Solver. Zur Behandlung der Nebenbedingungen empfiehlt sich ein *logical*-Array, welches an der i-ten Stelle den Wert `true` enthält, falls die i-te Nebenbedingung aktiv ist. Testen Sie Ihre Funktion mit dem Problem aus Blatt 9, Aufgabe 4. Stellen Sie dabei sicher, dass Ihr Startvektor $\mathbf{x}_0$ zulässig ist.

Duales Problem

Da die Systemmatrix zum Optimierungsproblem (1) singulär ist, müssen wir für dessen Lösung auf das duale Problem zurückgreifen. Wir betrachten deshalb das *Sattelpunktproblem* für die Lagrange-Funktion

$$(\mathbf{w}^\star, \boldsymbol{\lambda}^\star) = \sup_{\boldsymbol{\lambda}} \inf_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}), \qquad \mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{N} \lambda_i \big(y_i(\mathbf{w}^\mathsf{T} \mathbf{x}_i - w_0) - 1 \big).$$

Dies entspricht dem *dualen Problem*

$$\underset{\boldsymbol{\lambda}}{\text{minimiere}} \ \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\mathsf{T} \mathbf{x}_j - \sum_{i=1}^{N} \lambda_i, \quad \text{so dass} \quad \boldsymbol{\lambda}^\mathsf{T} \mathbf{y} = 0, \quad \boldsymbol{\lambda} \geq \mathbf{0}. \tag{5}$$

Dieses Problem ist ein Optimierungsproblem der Form (2), welches mit der Aktiven-Mengen-Strategie gelöst werden kann. Die gesuchten Parameter w_0 und $\mathbf{w}$ folgen dann aus den Beziehungen

$$\mathbf{w} = \sum_{i \in \Lambda_A} \lambda_i y_i \mathbf{x}_i, \qquad y_i(\mathbf{w}^\mathsf{T} \mathbf{x}_i - w_0) = 1 \text{ für } i \in \Lambda_A.$$

Die Punkte $\mathcal{V} := \{\mathbf{x}_i \,:\, i \in \Lambda_A\}$ werden *Support-Vectors* genannt.

Aufgabe 2. Schreiben Sie ein Skript, welches für die Punkte `xs_linear` und die Labels `ys_linear` den linearen Separator berechnet. Plotten Sie die zwei Klassen und markieren Sie zusätzlich die Support-Vectors. Zeichnen Sie ausserdem die Höhenlinien der Funktion $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x} - w_0$ für die Werte -1, 0 und 1. Benutzen Sie als Parameter `tol = 1e-5`.

Hinweis. *Die Matrix* $\mathbf{A}$ *kann schlecht konditioniert oder sogar singulär sein. Verwenden Sie daher für die Berechnungen die Matrix* $\mathbf{A} + \texttt{tol} \cdot \mathbf{I}$. *Für das Zeichnen der Höhenlinien ist die Funktion* `contour` *hilfreich.*

Hilbert-Räume

Für die weiteren Anwendungen werden wir in Hilbert-Räumen arbeiten. Ein Raum $(\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H})$ mit einem Skalarprodukt $\langle \cdot, \cdot \rangle_\mathcal{H}$ heisst *Hilbert-Raum*, wenn er vollständig ist in Bezug auf die durch das Skalarprodukt induzierte Norm $\|\cdot\|_\mathcal{H} := \sqrt{\langle \cdot, \cdot \rangle_\mathcal{H}}$. Das heisst, jede Cauchy-Folge bezüglich dieser Norm besitzt einen Grenzwert in $\mathcal{H}$. Das einfachste Beispiel ist der euklidische Raum $\mathbb{R}^d$ mit dem Standard-Skalarprodukt

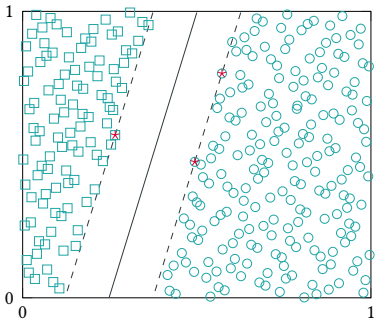
$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{i=1}^{d} u_i v_i.$$

Ein weiteres Beispiel ist der Raum der skalarwertigen, quadratisch integrierbaren Funktionen $L^2(\Omega)$,[1] mit dem L^2-Skalarprodukt

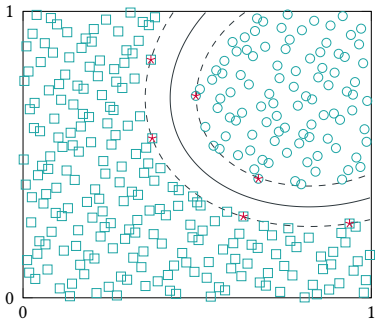
$$\langle f, g \rangle_{L^2(\Omega)} = \int_\Omega f(\mathbf{x}) g(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

Hingegen ist der Raum der stetigen Funktionen $C(\Omega)$ *kein* Hilbert-Raum.

[1] Genauer gesagt handelt es sich um Äquivalenzklassen quadratisch integrierbarer Funktionen.



Separator im linearen Fall



Separator im nichtlinearen Fall

Kernel-SVM

Es ist möglich, dass die zwei Mengen Ω_+ und Ω_- nicht linear trennbar sind. Abhilfe kann eine geeignete Transformation $\phi : \mathbb{R}^d \to \mathcal{H}$ in einen Hilbert-Raum $\mathcal{H}$ schaffen, wobei $\phi(\Omega_+)$ und $\phi(\Omega_-)$ linear trennbar in $\mathcal{H}$ sein sollen.

Dann betrachten wir das Minimierungsproblem

$$\underset{\mathbf{w}}{\text{minimiere}}\ \frac{1}{2}\|\phi(\mathbf{w})\|_{\mathcal{H}}^2, \quad \text{so dass} \quad y_i\big(\big\langle \phi(\mathbf{w}),\, \phi(\mathbf{x}_i)\big\rangle_{\mathcal{H}} - w_0\big) \geq 1, \quad i = 1, \dots, N.$$

Dabei ist die Abbildung ϕ nicht explizit bekannt. Die einzige Anforderung an ϕ ist, dass die Beziehung

$$\big\langle \phi(\mathbf{x}),\, \phi(\mathbf{y})\big\rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{y}) \quad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

gelten soll, wobei $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ eine geeignete, positiv definite Funktion sei, welche *Kernfunktion* genannt wird. Ein Beispiel für eine solche Funktion ist der *Gauß-Kern*

$$k_{\text{Gauß}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-|\mathbf{x}-\mathbf{y}|^2}{2\sigma^2}},$$

wobei der Parameter σ Korrelationslänge heisst.

Das duale Kernproblem ist dann entsprechend gegeben durch

$$\underset{\boldsymbol{\lambda}}{\text{minimiere}}\ \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{N} \lambda_i, \quad \text{so dass} \quad \boldsymbol{\lambda}^\mathsf{T}\mathbf{y} = 0, \quad \boldsymbol{\lambda} \geq \mathbf{0}. \tag{6}$$

Daraus folgen w_0 und $\phi(\mathbf{w})$ aus den Beziehungen

$$\phi(\mathbf{w}) = \sum_{i \in \Lambda_A} \lambda_i y_i \phi(\mathbf{x}_i), \qquad y_i \left(\sum_{j \in \Lambda_A}^{N} \lambda_j y_j k(\mathbf{x}_j, \mathbf{x}_i) - w_0 \right) = 1 \text{ für } i \in \Lambda_A,$$

und der Separator ist gegeben durch

$$\mathcal{S} := \left\{ \mathbf{x} \in \mathbb{R}^d \,:\, \big\langle \phi(\mathbf{w}),\, \phi(\mathbf{x})\big\rangle_{\mathcal{H}} = w_0 \right\} = \left\{ \mathbf{x} \in \mathbb{R}^{d'} \,:\, \sum_{i \in \Lambda_A} \lambda_i y_i k(\mathbf{x}_i, \mathbf{x}) = w_0 \right\}.$$

Aufgabe 3. Schreiben Sie ein Skript, welches für die Punkte `xs_kernel` und die Labels `ys_kernel` den Kern-Separator berechnet. Plotten Sie die zwei Klassen und Support-Vectors analog zu Aufgabe 2 und zeichnen Sie zusätzlich die Höhenlinien der Funktion

$$f(\mathbf{x}) = \sum_{i \in \Lambda_A} \lambda_i y_i k(\mathbf{x}_i, \mathbf{x}) - w_0$$

für die Werte -1, 0 und 1. Als Kernfunktion verwenden Sie den Gauß-Kern mit den Korrelationslängen $\sigma \in \{0.125, 0.5, 2\}$.

Soft-Margins

Für die Berechnung des Separators $\mathcal{S}$ haben wir bis jetzt gefordert, dass alle Punkte einen festen Abstand zum Rand besitzen. Liegt nun ein Punkt ausserordentlich nahe am Separator, was beispielsweise bei gestörten Daten der Fall sein kann, ist die SVM nicht mehr stabil. Dies kann umgangen werden, indem im Minimierungsproblem zugelassen wird, dass die Punkte näher am Rand oder sogar in der falschen Klasse sein können.

Anstelle von (1) wird für $\boldsymbol{\xi} \geq \mathbf{0}$ das Problem

$$\underset{\mathbf{w}}{\text{minimiere}}\ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \quad \text{so dass} \quad y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i - w_0) \geq 1 - \xi_i, \quad i = 1, \dots N,$$

gelöst, wobei $C > 0$ eine Regularisierungskonstante sei. Dies entspricht dem dualen Problem

$$\underset{\boldsymbol{\lambda}}{\text{minimiere}}\ \frac{1}{2}\sum_{i,j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^\mathsf{T}\mathbf{x}_j - \sum_{i=1}^{N}\lambda_i \quad \text{so dass} \quad \mathbf{y}^\mathsf{T}\boldsymbol{\lambda} = 0, \quad \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{C}, \tag{7}$$

mit den entsprechenden Anpassungen für die Kern-SVM. Dieser Ansatz wird *soft margin* genannt und kann auch mit der Aktiven-Mengen-Strategie gelöst werden, indem man die zusätzlichen Bedingungen $-\lambda_i \geq -C$ einführt.

Aufgabe 4. Wiederholen Sie Aufgabe 2, aber fügen Sie noch den Punkt $\mathbf{x}_{N+1} = (0.38, 0.05)^\mathsf{T}$ mit dem Label $y_{N+1} = -1$ hinzu. Wiederholen Sie Aufgabe 3, aber fügen Sie noch den Punkt $\mathbf{x}_{N+1} = (0.5, 0.5)^\mathsf{T}$ mit dem Label $y_{N+1} = -1$ hinzu. Lösen Sie anschliessend für beide neuen Datensätze das regularisierte Problem (7) und stellen Sie die Punkte analog zu den Aufgaben 2 und 3 grafisch dar. Als Regularisierungskonstante wählen Sie in beiden Fällen $C = 10$.

Literatur

[1] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 2nd edition, 2006.