AGML, Sommersemester 2020 Vorlesung 8 (5 Seiten)

Support Vector Machines und Konvexität

Thomas Kesselheim **Vorschau** Letzte Aktualisierung: 18. Mai 2020

Wie auch in den vergangenen Vorlesungen werden wir uns heute wieder mit linearer Klassifikation beschäftigen. Wir erinnern uns, dass die Hypothesenklasse $\mathcal{H}$ definiert ist als die Menge von Funktionen der Form $h_{\mathbf{w},u}\colon \mathbb{R}^d \to \{-1,+1\}$ für $\mathbf{w} \in \mathbb{R}^d$ und $u \in \mathbb{R}$ und

$$h_{\mathbf{w},u}(\mathbf{x}) = \begin{cases} +1 \text{ falls } \langle \mathbf{w}, \mathbf{x} \rangle \geq u \\ -1 \text{ sonst} \end{cases} .$$

Hierbei beschreibt $\langle \mathbf{w}, \mathbf{x} \rangle$ das Skalarprodukt der Vektoren $\mathbf{w}$ und $\mathbf{x}$. Wir nehmen auch wieder an, dass uns eine Trainingsmenge S von Datenpunkten mit Labels $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ gegeben ist.

Wie wir in der letzten Vorlesung gesehen haben, können wir in Polynomialzeit eine Hypothese berechnen, die alle Datenpunkte in S korrekt klassifiziert, sofern dies möglich ist. Gleichzeitig gibt es unter der Annahme P $\neq$ NP keinen Polynomialzeitalgorithmus, der die maximale mögliche Anzahl von Punkten korrekt klassifiziert.

Beide Probleme werden wir heute erneut betrachten. Wir werden Probleme formulieren, deren Ziel es ist eine „möglichst gute“ Hypothese zu berechnen, und die gleichzeitig Polynomialzeitalgorithmen zulassen. Die Algorithmen selbst werden wir dann in den kommenden Vorlesungen besprechen.

1 Hard-SVM-Problem

Das Ziel beim Hard-SVM-Problem ist es, eine Hypothese $h_{\mathbf{w},u}$ zu finden, die alle Datenpunkte in S richtig klassifiziert unter der Annahme, dass das möglich ist. In anderen Worten sollen die positiven von den negativen Punkten linear separierbar sein. Zusätzlich sollen die Datenpunkte möglichst deutlich klassifiziert werden. Das bedeutet, dass der Abstand von der Hyperebene, die durch $\mathbf{w}$ und u definiert wird, möglichst groß sein soll. Anders formuliert soll die Hypothese auch noch möglichst lange korrekt bleiben, selbst wenn die Punkte in ihrer Umgebung verschoben werden.

Leiten wir nun zunächst eine Formel für den Abstand von einer Hyperebene her. Zur Erinnerung: Der Abstand zweier Punkte $\mathbf{v}$ und $\mathbf{v}'$ ist definiert als die Norm der Differenz der Vektoren $\|\mathbf{v} - \mathbf{v}'\|$. Wir betrachten im Folgenden nur die euklidische Norm, definiert als $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.

Lemma 8.1. *Der Abstand eines Punktes $\mathbf{x}$ von einer Hyperebene definiert durch $(\mathbf{w}, u)$ ist $\frac{1}{\|\mathbf{w}\|} |\langle \mathbf{w}, \mathbf{x} \rangle - u|$.*

Beweis. Wir definieren einen Punkt $\mathbf{v} = \mathbf{x} - c\mathbf{w}$ mit $c = \frac{1}{\|\mathbf{w}\|^2}(\langle \mathbf{w}, \mathbf{x} \rangle - u)$. Nun werden wir nachweisen, dass $\mathbf{v}$ (i) in der Hyperebene liegt, (ii) den besagten Abstand von $\mathbf{x}$ hat und (iii) kein Punkt der Hyperebene näher an $\mathbf{x}$ liegt.

Für (i) setzen wir die Definition $\mathbf{v}$ ein und erhalten

$$\langle \mathbf{w}, \mathbf{v} \rangle = \langle \mathbf{w}, \mathbf{x} - c\mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{x} \rangle - c \langle \mathbf{w}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{x} \rangle - c\|\mathbf{w}\|^2 = u \ .$$

Also erfüllt $\mathbf{v}$ die Hyperebenegleichung.

Für (ii) nutzen wir ebenfalls die Definition von $\mathbf{v}$ und elementare Umformungen. Dies gibt uns

$$\|\mathbf{x}-\mathbf{v}\| = \|c\mathbf{w}\| = |c|\|\mathbf{w}\| = \left|\frac{1}{\|\mathbf{w}\|^2}(\langle \mathbf{w},\mathbf{x}\rangle - u)\right| \|\mathbf{w}\| = \frac{1}{\|\mathbf{w}\|}\,|\langle \mathbf{w},\mathbf{x}\rangle - u| \ .$$

Für (iii) betrachten wir nun irgendeinen anderen Punkt $\mathbf{v}'$ auf der Hyperebene. Das Quadrat dessen Abstands zu $\mathbf{x}$ berechnet sich zu

$$\|\mathbf{x}-\mathbf{v}'\|^2 = \|\mathbf{x}-\mathbf{v}+\mathbf{v}-\mathbf{v}'\|^2 = \|\mathbf{x}-\mathbf{v}\|^2+\|\mathbf{v}-\mathbf{v}'\|^2+2\left\langle \mathbf{x}-\mathbf{v},\mathbf{v}-\mathbf{v}'\right\rangle \geq \|\mathbf{x}-\mathbf{v}\|^2+2\left\langle \mathbf{x}-\mathbf{v},\mathbf{v}-\mathbf{v}'\right\rangle \ .$$

Es bleibt also nur zu zeigen, dass $\langle \mathbf{x}-\mathbf{v},\mathbf{v}-\mathbf{v}'\rangle \geq 0$. Aufgrund der Definition von $\mathbf{v}$ ist $\mathbf{x}-\mathbf{v} = c\mathbf{w}$ also

$$\left\langle \mathbf{x}-\mathbf{v},\mathbf{v}-\mathbf{v}'\right\rangle = \left\langle c\mathbf{w},\mathbf{v}-\mathbf{v}'\right\rangle = c\left(\langle \mathbf{w},\mathbf{v}\rangle - \left\langle \mathbf{w},\mathbf{v}'\right\rangle\right) = c(-u+u) = 0 \ .$$

Hierbei haben wir ausgenutzt, dass sowohl $\mathbf{v}$ als auch $\mathbf{v}'$ auf der Hyperebene liegen. □

Wir wollen nun eine Hyperebene finden, die alle Punkte $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_m,y_m)$ korrekt klassifiziert und außerdem unter diesen Hyperebenen den minimalen Abstand zu den Punkten maximiert. Dies können wir nun als ein Optimierungsproblem aufschreiben

$$\begin{aligned} \text{maximiere } & \min_i \frac{1}{\|\mathbf{w}\|}\,|\langle \mathbf{w},\mathbf{x}_i\rangle - u| & \\ \text{unter den Nebenbedingungen } & \langle \mathbf{w},\mathbf{x}_i\rangle - u \geq 0 & \text{falls } y_i = 1 \\ & \langle \mathbf{w},\mathbf{x}_i\rangle - u < 0 & \text{falls } y_i = -1 \end{aligned}$$

Eine optimale Lösung zu dieser Formulierung zu finden ist nicht einfach. Die Nebenbedingungen sind zwar linear, aber die Zielfunktion ist kompliziert. Deshalb schreiben wir das Problem leicht um.

Zunächst einmal stellen wir fest, dass wir mittels der Nebenbedingungen die Betragsstriche in der Zielfunktion eliminieren können. Egal ob $y_i = 1$ oder $y_i = -1$, gilt immer $|\langle \mathbf{w},\mathbf{x}_i\rangle - u| = y_i(\langle \mathbf{w},\mathbf{x}_i\rangle - u)$. So lautet unser Problem nun

$$\begin{aligned} \text{maximiere } & \min_i \frac{1}{\|\mathbf{w}\|} y_i(\langle \mathbf{w},\mathbf{x}_i\rangle - u) & \\ \text{unter den Nebenbedingungen } & \langle \mathbf{w},\mathbf{x}_i\rangle - u \geq 0 & \text{falls } y_i = 1 \\ & \langle \mathbf{w},\mathbf{x}_i\rangle - u < 0 & \text{falls } y_i = -1 \end{aligned}$$

Betrachten wir nun eine optimale Lösung $(\mathbf{w},u)$, stellen wir fest, dass niemals $\langle \mathbf{w},\mathbf{x}_i\rangle - u = 0$ für ein i sein wird, weil wir ansonsten u leicht erhöhen könnten. Dies würde den Zielfunktionswert nur verbessern und die Lösung würde weiter gültig bleiben. Außerdem erfüllt jede Lösung mit positivem Zielfunktionswert automatisch alle Nebenbedingungen. Somit vereinfacht sich das Problem dahingehend $\mathbf{w}$ und u zu finden, sodass

$$\min_i \frac{1}{\|\mathbf{w}\|} y_i(\langle \mathbf{w},\mathbf{x}_i\rangle - u)$$

maximiert wird.

Gegeben eine optimale Lösung $(\mathbf{w}, u)$, sei nun $\gamma = \min_i y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u)$. Betrachte $\mathbf{w}' = \frac{1}{\gamma}\mathbf{w}$, $u' = \frac{1}{\gamma}u$. Wir stellen fest, dass

$$\frac{1}{\|\mathbf{w}'\|} y_i(\langle \mathbf{w}', \mathbf{x}_i \rangle - u') = \frac{\gamma}{\|\mathbf{w}\|} y_i \left(\left\langle \frac{\mathbf{w}}{\gamma}, \mathbf{x}_i \right\rangle - \frac{u}{\gamma} \right) = \frac{1}{\|\mathbf{w}\|} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u)$$

für alle i. Also hat $(\mathbf{w}', u')$ denselben Zielfunktionswert wie $(\mathbf{w}, u)$, ist also auch eine optimale Lösung. Wir können also genauso gut auch $(\mathbf{w}', u')$ suchen. Weil bei dieser Lösung $\min_i y_i(\langle \mathbf{w}', \mathbf{x}_i \rangle - u') = 1$, ist dies gleichbedeutend mit

$$\begin{aligned} &\text{minimiere } \|\mathbf{w}'\|^2 \\ &\text{unter den Nebenbedingungen } y_i(\langle \mathbf{w}', \mathbf{x}_i \rangle - u') \geq 1 && \text{für alle } i \end{aligned}$$

Diese Formulierung heißt *Hard-SVM*. In der Tat werden wir Algorithmen kennenlernen, die ein solches Optimierungsproblem lösen können.

2 Soft-SVM-Problem

Die Ergebnisse in Abschnitt 1 setzen voraus, dass die Punkte linear separierbar sind. Das heißt, dass es eine Hypothese gibt, die alle Punkte in der Menge S korrekt klassifiziert. Um den Trainingsfehler zu minimieren, müsste man eine Hypothese finden, die möglichst viele Datenpunkte korrekt klassifiziert. In der Notation, die wir nun eingeführt haben, bedeutet dies, dass die Bedingung $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + u) \leq 1$ für möglichst wenige i nicht erfüllt ist. Wie wir bereits in der letzten Vorlesung gesehen haben, ist dies jedoch NP-schwer.

Der *Soft-SVM*-Ansatz ist daher ein anderer. Wir führen bei jeder Nebenbedingung eine Variable ξ ein, wie weit sie verletzt ist. Das heißt, wir fordern nun noch, dass $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u) \leq 1 - \xi_i$. Es ist nun auch das Ziel, den durchschnittlichen Fehler zu minimieren.

Die neue Formulierung lautet somit

$$\begin{aligned} &\text{minimiere } \lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \\ &\text{unter den Nebenbedingungen } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u) \geq 1 - \xi_i && \text{für alle } i \\ &\qquad\qquad \xi_i \geq 0 && \text{für alle } i \end{aligned}$$

Hierbei drückt $\lambda \geq 0$ eine Gewichtung aus: $\|\mathbf{w}\|^2$ ist der Term, der ursprünglich ausgedrückt hat, dass der Abstand möglichst groß sein soll; $\frac{1}{m}\sum_{i=1}^{m} \xi_i$ ist der durchschnittliche Fehler, der misst, wie weit Punkte jeweils auf der falschen Seite der Hyperebene sind.

Dieses Problem können wir noch umformulieren. Wir nutzen aus, dass in einer optimalen Lösung immer $\xi_i = \max\{0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u)\}$ sein wird. Damit ist es äquivalent,

$$\lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u)\}$$

zu minimieren.

3 Konvexe Optimierung

Das Hard-SVM- und das Soft-SVM-Problem lassen sich wie folgt darstellen. Wir möchten eine Funktion $f\colon S \to \mathbb{R}$ minimieren, wobei $S \subseteq \mathbb{R}^n$ die Menge aller zulässigen Lösungen darstellt.

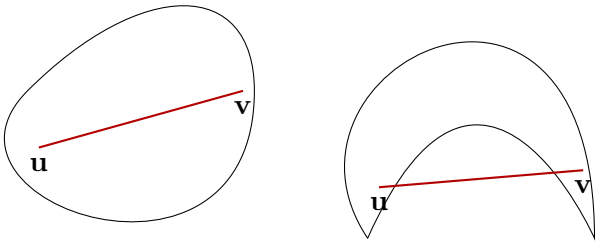

Abbildung 1: Links eine konvexe Menge, rechts eine nicht-konvexe Menge.

In unserem Fall enthält S alle $(d+1)$-dimensionalen zulässigen Vektoren $(\mathbf{w}, u)$. Das heißt, wir fügen unter die d Komponenten von $\mathbf{w}$ mit u einer weitere Komponente an. Im Fall von Soft-SVM gibt es keine weiteren Einschränkungen, also ist $S = \mathbb{R}^n$ mit $n = d+1$. Im Fall von Hard-SVM müssen mittels S die Nebenbedingungen berücksichtigen.

Glücklicherweise sind sowohl die Menge S als auch die Funktion f konvex. Deshalb werden wir die Probleme mithilfe von Algorithmen aus der Konvexen Optimierung lösen können.

Die Menge S ist jeweils *konvex*. Das heißt, dass für zwei Punkte $\mathbf{u}, \mathbf{v} \in S$ alle Punkte auf der Verbindungslinie wieder in S enthalten ist (siehe Abbildung 1). Formal also $\lambda\mathbf{u} + (1-\lambda)\mathbf{v} \in S$ für alle $\lambda \in [0,1]$.

Zusätzlich ist auch die Funktion f konvex. Das bedeutet, dass der Funktionsgraph zwischen zwei Punkten jeweils unterhalb der Verbindungslinie dieser beiden Punkte liegt. Das heißt, für $\mathbf{u}, \mathbf{v} \in S$ gilt $f(\lambda\mathbf{u} + (1-\lambda)\mathbf{v}) \leq \lambda f(\mathbf{u}) + (1-\lambda) f(\mathbf{v})$ für alle $\lambda \in [0,1]$. Ein typisches Beispiel einer konvexen Funktion, das man immer im Kopf haben sollte, ist eine quadratische Funktion in einer Dimension (siehe Abbildung 2 links).

Wenn die Funktionen differenzierbar sind, gibt es viele äquivalente Definitionen von Konvexität. Betrachten wir zunächst den eindimensionalen Fall. Hier muss beispielsweise die zweite Ableitung nicht-negativ sein. Im Kontext von Konvexer Optimierung werden wir jedoch folgende äquivalente Definition nutzen: Die Funktion fällt niemals unterhalb ihre Tangenten. Ausgedrückt in der ersten Ableitung bedeutet dies, dass eine differenzierbare Funktion $f\colon S \to \mathbb{R}$ konvex ist, wenn für alle $u, v \in S$ gilt

$$f(u) \geq f(v) + f'(v)(u - v) \ .$$

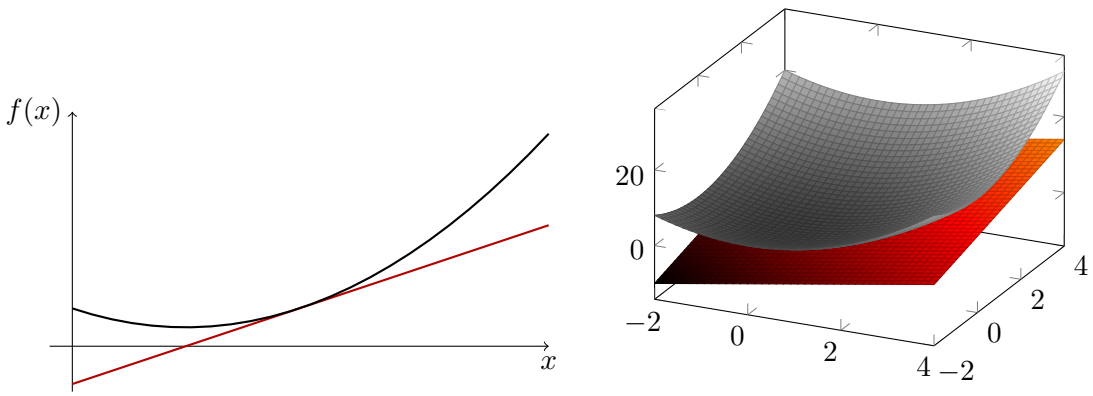

Abbildung 2: Typische konvexe Funktionen in einer bzw. zwei Dimensionen, jeweils mit einer Tangente bzw. Tangentialhyperebene in rot.

All diese Definitionen lassen sich auch ins Mehrdimensionale übertragen. Die Funktion f hat nun einen Gradienten ∇f, der der Vektor aller partiellen Ableitungen ist; $(\nabla f(\mathbf{u}))_i = \frac{\partial f}{\partial u_i}(\mathbf{u})$. Eine differenzierbare Funktion $f\colon S \to \mathbb{R}^n$ ist konvex, wenn sie niemals unter ihre Tangentialhyperebene fällt (siehe Abbildung 2 rechts). Das heißt, dass für alle $\mathbf{u}, \mathbf{v}$

$$f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), (\mathbf{u} - \mathbf{v}) \rangle \ . \tag{1}$$

Die Soft-SVM-Zielfunktion ist nicht differenzierbar. Trotzdem erfüllt sie eine ähnliche Bedingung, wie wir sehen werden.

Wir werden nicht nachweisen, dass die Hard- und Soft-SVM-Zielfunktionen konvex sind. Dies folgt aus relativ einfachen Rechnungen. Folgende Abschlusseigenschaften sind dabei hilfreich.

Lemma 8.2. *1. Sind f und g konvex, dann sind auch $f + g$ und $\max\{f, g\}$ konvex.*

2. Ist f konvex und $\alpha \geq 0$, dann ist auch αf konvex.

3. Sind f und g konvex und g zusätzlich monoton steigend, dann ist auch $g \circ f$ konvex.