Der Kernel-Trick

M. Gruber

KW 45, Rev.2

Bei der Herleitung des SVM-Lernmodells kommen die Elemente $x^{(1)}, \ldots, x^{(N)} \subset \mathcal{X}$ der Lernmenge ausschließlich in Form von Skalarprodukten $\langle x^{(i)}, x^{(j)} \rangle$ vor. Der *kernel*-Trick besteht darin, dass man diese Skalarprodukt durch Ausdrücke der Form $k(x^{(i)}, x^{(j)})$ mit einem *kernel* $k(.,.)$ ersetzt und so ein anderes SVM-Lernmodell erhält. Die formale Rechtfertigung für dieses Vorgehen liegt darin, dass $k(.,.)$ als inneres Produkt $\langle .,. \rangle_{\mathcal{H}_k}$ eines Hilbertraums $\mathcal{H}_k$ identifiziert werden kann, in den $\mathcal{X}$ mittels einer geeigneten *feature map* Φ abgebildet wird. Man liest $k(x^{(i)}, x^{(j)})$ als $\langle \Phi(x^{(i)}), \Phi(x^{(j)}) \rangle_{\mathcal{H}_k}$.

Definition 1 (*kernel*) *Eine Abbildung $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ ist ein kernel, wenn k symmetrisch ist und die Gram-Matrizen $[k(x^{(j)}, x^{(j)})]_{1 \leq i,j \leq N}$ für alle Teilmengen $\{x^{(1)}, \ldots, x^{(N)}\} \subset \mathcal{X}$ positiv semidefinit sind.*

Bemerkung 1 *Eine reelle $(N \times N)$-Matrix A ist positiv semidefinit, wenn $x^T A x \geq 0$ für jedes $x \in \mathbf{R}^N$ ist. Ist A zudem symmetrisch, gibt es eine Faktorisierung $A = U \Lambda U^T$ mit einer Orthonormalmatrix U und einer Diagonalmatrix Λ, deren Diagonalelemente nichtnegativ sind.*

Bemerkung 2 *Für einen kernel k gilt stets $k(x,x) \geq 0$, denn $[k(x,x)]$ ist eine 1×1-Gram-Matrix und als solche positiv semidefinit.*

Bemerkung 3 (Cauchy-Schwarz für *kernels*) *Für einen kernel k gilt stets*

$$k(x,x') \leq \sqrt{k(x,x)}\sqrt{k(x',x')}.$$

Um dies zu sehen, betrachte man die Gram-Matrix $K = \begin{bmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{bmatrix}$ und den Vektor $c = \begin{bmatrix} k(x',x') & -k(x,x') \end{bmatrix}^T$. Aus $0 \leq c^T K c = (k(x',x')k(x,x) - k(x,x')^2)k(x',x')$ folgt die Behauptung, sofern $k(x',x') > 0$ ist. Im Fall $k(x',x') = 0$ betrachte man zunächst $k(.,.) + \varepsilon$ statt k und gehe dann mit $\varepsilon \downarrow 0$ zu k über.

Beim *kernel*-Trick in seiner einfachsten Form konstruiert man die *feature map* Φ und den Hilbertraum $\mathcal{H}_k$ mit seinem inneren Produkt explizit.

Beispiel 1 *Sei* $(\mathcal{X}, \langle .,. \rangle_{\mathcal{X}\times\mathcal{X}}) = (\mathbf{R}^2, \langle .,. \rangle_{\mathbf{R}^2\times\mathbf{R}^2})$ *und* $(\tilde{\mathcal{X}}, \langle .,. \rangle_{\tilde{\mathcal{X}}\times\tilde{\mathcal{X}}}) = (\mathbf{R}^3, \langle .,. \rangle_{\mathbf{R}^3\times\mathbf{R}^3})$. *Die Abbildung* Φ *vom feature-Raum* $\mathcal{X}$ *in den feature-Raum* $\tilde{\mathcal{X}}$ *sei* $\begin{bmatrix} x_1 & x_2 \end{bmatrix} \mapsto \begin{bmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{bmatrix}$. *Der kernel, der* $\langle .,. \rangle_{\mathcal{X}\times\mathcal{X}}$ *ersetzt, ist*

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\tilde{\mathcal{X}}\times\tilde{\mathcal{X}}} .$$

Man kann nachrechnen, dass $k(.,.) = (\langle .,. \rangle_{\mathcal{X}\times\mathcal{X}})^2$ *ist. Ist es Zufall, dass das Quadrat eines kernels wieder ein kernel ist? Hätten wir zum kernel* k *übergehen können, ohne uns um* Φ *und* $\mathcal{H}_k$ *zu kümmern?*

Satz 1 (Moore-Aronszajn) *Zu einem gegebenen feature-Raum* $\mathcal{X}$ *mit kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ *gibt es eine feature map* Φ *in einen Hilbertraum* $\mathcal{H}_k$, *sodass* $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}_k}$ *ist.*

Beweis

1. Als *feature map* $\Phi : \mathcal{X} \to \mathbf{R}^{\mathcal{X}}$ wählen wir die Abbildung $\Phi(x) = k(., x)$. (Mit $\mathbf{R}^{\mathcal{X}}$ bezeichnet man die Menge aller Abbildungen von $\mathcal{X}$ nach $\mathbf{R}$.)
2. Sei $\mathcal{H}_k$ der von $\{\Phi(x), x \in \mathcal{X}\}$ aufgespannte Vektorraum, d.h. die Menge aller endlichen Linearkombinationen der Form $\sum_i \alpha_i \Phi(x^{(i)})$.
3. Auf $\mathcal{H}_k$ definieren wir das innere Produkt $\left\langle \sum_i \alpha_i \Phi(x^{(i)}), \sum_j \beta_j \Phi(x^{\prime(j)}) \right\rangle_{\mathcal{H}_k} = \sum_i \sum_j \alpha_i \beta_j k(x^{(i)}, x^{\prime(j)})$.
4. Das innere Produkt $\langle .,. \rangle_{\mathcal{H}_k}$ ist wohldefiniert, symmetrisch und bilinear.
5. Das innere Produkt $\langle .,. \rangle_{\mathcal{H}_k}$ ist positiv semidefinit: $\left\langle \sum_i \alpha_i \Phi(x^{(i)}), \sum_i \alpha_i \Phi(x^{(i)}) \right\rangle_{\mathcal{H}_k} = \alpha^T K \alpha \geq 0$ mit der Gram-Matrix $K = [k(x^{(i)}, x^{(j)})]_{1 \leq i,j \leq N}$.
6. Das innere Produkt von $\Phi(x)$ und $f = \sum_i \alpha_i \Phi(x^{(i)})$ ist $\langle \Phi(x), f \rangle_{\mathcal{H}_k} = \sum_i \alpha_i k(x, x^{(i)}) = f(x)$.
7. Für einen Hilbertraum genügt es nicht, dass das innere Produkt positiv semidefinit ist. Es muss positiv definit sein. Unser inneres Produkt ist positiv definit. Man kann das folgendermaßen sehen: Für $f \neq 0$ gibt es $x \in \mathcal{X}$ mit $f(x) \neq 0$ und es ist $|f(x)|^2 = \langle \Phi(x), f \rangle^2_{\mathcal{H}_k} \leq ||\Phi(x)||^2_{\mathcal{H}_k} ||f||^2_{\mathcal{H}_k}$. Es muss also $||f||_{\mathcal{H}_k} > 0$ sein und $\langle .,. \rangle_{\mathcal{H}_k}$ positiv definit (d.h. $\langle f, f \rangle_{H_k} = 0 \Leftrightarrow f = 0$).
8. Der Raum $\mathcal{H}_k$ ist mit den bis hierher nachgewiesenen Eigenschaften ein Prähilbertraum. In einem Prähilbertraum ist nicht garantiert, dass alle Cauchyfolgen innerhalb des Raumes einen Grenzwert haben. Es ist möglicherweise "unvollständig". Die "Vervollständigung" eines Prähilbertraumes zu einem Hilbertraum ist eine Standardkonstruktion der Funktionalanalysis. Es werden dabei fehlende Grenzwerte als abstrakte Größen dem Raun hinzugefügt. Wir verzichten hier auf die Darstellung dieses Details.

□

Bemerkung 4 *Im Beweis von Satz 1 benutzen wir eine Cauchy-Schwarz-Ungleichung für Seminormen. In [1], p.119, wird die Gültigkeit solcher Ungleichungen etwa so nachgewiesen:*

1. *Für* $f, g \in \mathcal{H}_k$ *und mit* $a = \langle f, f \rangle_{\mathcal{H}_k}$, $b = \langle f, g \rangle_{\mathcal{H}_k}$, $c = \langle g, g \rangle_{\mathcal{H}_k}$ *ist* $a \geq 0$ *und* $c \geq 0$.

2. *Für beliebiges* λ *ist* $\langle f+\lambda g, f+\lambda g\rangle_{\mathcal{H}_k} = a + 2\lambda b + \lambda^2 c \geq 0$.

3. *Ist* $c \neq 0$, *so setze man* $\lambda = -b/c$. *Dann ist* $b^2 \leq a \cdot c$.

4. *Ist* $c = 0$, *aber* $a \neq 0$, *argumentiert man analog.*

5. *Sind* $c = 0$ *und* $a = 0$, *folgt mit* $\lambda = -b$, *dass* $-b^2 \geq 0$ *ist. Dann ist auch* $b = 0$.

Die positiv semidefinite Bilinearform $\langle .,. \rangle_{\mathcal{H}_k}$ *induziert durch* $\|f\|_k = \sqrt{\langle f, f\rangle}_{\mathcal{H}_k}$ *eine Seminorm auf* $\mathcal{H}_k$ *(mit den Eigenschaften* $\|\lambda f\|_{\mathcal{H}_k} = |\lambda| \|f\|_{\mathcal{H}_k}$, $\|f\|_{\mathcal{H}_k} \geq 0$ *und* $\|f+g\|_{\mathcal{H}_k} \leq \|f\|_{\mathcal{H}_k} + \|g\|_{\mathcal{H}_k}$*).*

Da wir nun wissen, dass der *kernel*-Trick mit jedem *kernel* funktioniert, machen wir uns auf die Suche nach *kernels*.

Bemerkung 5 $k(x, x') = c$ *mit* $c > 0$ *ist ein kernel.*

Bemerkung 6 $\tilde{k}(x, x') = k(h(x), h(x'))$ *mit* $h : X \to X$ *ist ein kernel, falls* $k(x, x')$ *ein kernel ist. Denn wenn* $k(x, x')$ *ein kernel ist, ist auch jede Gram-Matrix* K *mit* $K_{ij} = k(h(x^{(i)}), h(x^{(j)})$ *symmetrisch und positiv semidefinit (denn die* $h(x^{(i)})$ *liegen im feature-Raum)*

Bemerkung 7 $k(x, x') = g(x)g(x')$ *mit* $g : X \to \mathbf{R}$ *ist ein kernel, denn die Gram-Matrizen sind hier von der Form*

$$K = \begin{bmatrix} g(x^{(1)}) \\ \vdots \\ g(x^{(N)}) \end{bmatrix} \begin{bmatrix} g(x^{(1)}) & \dots & g(x^{(N)}) \end{bmatrix}$$

und man hat Symmetrie und positive Semidefinitheit.

Satz 2 1. *Das Produkt eines kernels mit einem positiven Skalar ergibt einen kernel.*

2. *Die Summe zweier kernel ist ein kernel.*

3. *Das Produkt zweier kernel ist ein kernel.*

Beweis Seien $x^{(1)}, \dots, x^{(N)}$ Elemente des *feature*-Raums. Seien k_1, k_2 *kernels* und seien

$$K_1 = \left[k_1(x^{(i)}, x^{(j)})\right]_{1 \leq i,j \leq N}, \quad K_2 = \left[k_2(x^{(i)}, x^{(j)})\right]_{1 \leq i,j \leq N}, \quad K = \left[k_1(x^{(i)}, x^{(j)}) k_2(x^{(i)}, x^{(j)})\right]_{1 \leq i,j \leq N}$$

die Gram-Matrizen zu $k_1, k_2, k_1 \cdot k_2$.

1. Ist K_1 symmetrisch und positiv semidefinit, dann auch αK_1, sofern $\alpha > 0$ ist. Denn auch αK_1 ist dann symmetrisch und $c^T(\alpha K_1)c = \alpha(c^T K_1 c) \geq 0$.
2. Sind K_1 und K_2 symmetrisch und positiv semidefinit, dann auch $K_1 + K_2$. Denn auch $K_1 + K_2$ ist dann symmetrisch und $c^T(K_1 + K_2)c = c^T K_1 c + c^T K_2 c \geq 0$.

3. Die Symmetrie von K sieht man leicht. Zu zeigen bleibt die positive Semidefinitheit.

   Aus der linearen Algebra weiß man, dass K_1 aufgrund von Symmetrie und positiver Definitheit eine Darstellungen der Form $K_1 = U_1 \Lambda_1 U_1^T$ hat, wobei $U_1^T U_1 = I$ ist (mit Einheitsmatrix I) und Λ_1 eine Diagonalmatrix mit nichtnegativen Einträgen auf der Hauptdiagonalen (nämlich die Eigenwerte von K_1). Analoges gilt für K_2: $K_2 = U_2 \Lambda_2 U_2^T$, $U_2^T U_2 = I$ und Λ_2 ist eine Diagonalmatrix mit nichtnegativen Einträgen auf der Hauptdiagonalen.
   $\Lambda_1^{1/2}$ und $\Lambda_2^{1/2}$ seien die Diagonalmatrizen, die man durch Wurzelziehen aus Λ_1 und Λ_2 bekommt. Mit $\Phi_n(x^{(j)}) = \Lambda^{1/2}(U_n^T)_{*j}$ ist $K_n = \left[\Phi_n(x^{(i)})^T \Phi_n(x^{(j)})\right]_{1 \le i,j \le N}$ für $n = 1, 2$. (Mit $(U_n^T)_{*j}$ ist die j-te Spalte von U_n^T gemeint.) Für das Element in der i-ten Zeile und j-ten Spalte von K gilt

$$K_{ij} = \Phi_1(x^{(i)})^T \Phi_1(x^{(j)}) \Phi_2(x^{(i)})^T \Phi_2(x^{(j)}) = \Big(\sum_{1 \le k \le N} \Phi_1(x^{(i)})_k \Phi_1(x^{(j)})_k \Big) \cdot \Big(\sum_{1 \le l \le N} \Phi_2(x^{(i)})_l \Phi_2(x^{(j)})_l \Big).$$

   Multipliziert man das Produkt der Summen aus und ordnet die einzelnen Produkte um, erhält man

$$K_{ij} = \sum_{1 \le k \le N} \sum_{1 \le l \le N} \Phi_1(x^{(i)})_k \Phi_2(x^{(i)})_l \Phi_1(x^{(j)})_k \Phi_2(x^{(j)})_l.$$

   Die Abbildung $f : \{1, \ldots, N^2\} \to \{(i,j) \mid 1 \le i, j \le N\}$, $(f_1(m), f_2(m)) = \left(\lceil \frac{m}{N} \rceil, m - (\lceil \frac{m}{N} \rceil - 1)N\right)$ ist bijektiv. Sei $\Phi(x^{(i)})$ der Vektor in $\mathbf{R}^{N^2}$ mit

$$\Phi(x^{(i)})_m = \Phi_1(x^{(i)})_{f_1(m)} \Phi_2(x^{(i)})_{f_2(m)}.$$

   Damit ist

$$K_{ij} = \Phi(x^{(i)})^T \Phi(x^{(j)}),$$

   und es folgt, dass K positiv semidefinit ist, denn

$$c^T K c = c^T \left[\Phi(x^{(1)}) \quad \ldots \quad \Phi(x^{(N)})\right]^T \left[\Phi(x^{(1)}) \quad \ldots \quad \Phi(x^{(N)})\right] c = \| \left[\Phi(x^{(1)}) \quad \ldots \quad \Phi(x^{(N)})\right] c \|^2 \ge 0.$$

□

Beispiel 2 *1. $\alpha k_1(x, x') + \beta k_2(x, x')$ mit $\alpha, \beta \ge 0$ ist ein kernel, falls k_1, k_2 kernel sind.*

2. *$p(k(x, x'))$ ist ein kernel, falls p ein Polynom mit positiven Koeffizienten und $k(x, x')$ ein kernel ist.*

3. *$e^{k(x,x')}$ ist ein kernel, falls $k(x, x')$ ein kernel ist.*

Beispiel 3 (*Gaussian kernel*) *(nach [2])* *$\tilde{k}(x, x') = e^{-\|x - x'\|^2 / \sigma^2}$ ist ein kernel, denn man hat einen kernel wie in Beispiel 7, der mit einem exponentiellen kernel wie in Beispiel 2 multipliziert wird: $e^{-\|x-x'\|^2/\sigma^2} = e^{-\|x\|^2/\sigma^2} e^{-\|x'\|^2/\sigma^2} e^{2x^T x'/\sigma^2}$.*

Das Ergebnis eines *kernel*-Tricks kann man so zusammenfassen: Wenn das QP-Problem mit der Systemmatrix $Q = [y^{(i)} y^{(j)} k(x^{(i)}, x^{(j)})]_{1 \leq i,j \leq N}$ gelöst ist, hat man den Klassifizierer

$$\operatorname{sgn}(w^T u + b) = \operatorname{sgn}\Big(\sum_{x^{(n)} \text{ support vector}} \alpha_n y^{(n)} (k(x^{(n)}, u) - k(x^{(n)}, x^{(n^*)})) + (1/y^{(n^*)}) \Big).$$

Literatur

[1] J. Dieudonné. *eléments d'analyse*, volume 1. Gauthier-Villars, Paris, 1968.

[2] C. Rudin. Prediction: Machine Learning and Statistics. MIT OCW 15.097.