Kernel-Methode

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Eine Kernel-Methode wendet eine lineare Methode, wie ein lineares Modell oder einen linearer Klassifikator, auf transformierte Merkmalsvektoren an. Die Transformation wird aus einer Kernel konstruiert: Jeder Merkmalsvektor wird auf eine Funktion abgebildet, deren Definitionsbereich gleich dem ursprünglichen Merkmalsraum ist, und Skalarprodukte zwischen transformierten Merkmalsvektoren reduzieren sich auf Auswertungen einer Kernel. Kernel-Methoden sind auf jeden Merkmalsraum anwendbar, auf dem eine Kernel definiert ist, einschließlich Mengen von Textdokumenten und Graphen.

Definition

Zu entscheiden, ob ein Bild ein bestimmtes Objekt zeigt (Lampert, 2009) oder zu welchem Thema ein Textdokument gehört (Cristianini and Shawe-Taylor, 2000), sind Aufgaben für Kernel-Methoden. Eine Kernel-Methode ist eine Methode des maschinelles Lernen (ML), die eine lineare Methode, wie ein lineares Modell für die Regression oder einen linearer Klassifikator für die Klassifikation, auf transformierte Merkmalsvektoren anwendet, die aus einer Kernel konstruiert werden (Lampert, 2009; Schölkopf and Smola, 2002).

Lineare Methoden lernen eine Hypothese, welche die Merkmale eines Datenpunkts linear kombiniert, z. B. $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ für den Merkmalsraum $\mathcal{X} = \mathbb{R}^d$. Eine solche Hypothese liefert nur dann gute Vorhersagen, wenn der Zusammenhang zwischen den Merkmalsvektoren und den Labels der Datenpunkte näherungsweise linear ist. Zudem erfordert eine lineare Methode numerische Merkmalsvektoren und ist nicht direkt auf Datenpunkte aus einem beliebigen Merkmalsraum $\mathcal{X}$, wie etwa Textdokumente oder Graphen, anwendbar.

Beide Einschränkungen lassen sich durch eine Merkmalstransformation $\phi : \mathcal{X} \to \mathcal{H}$ vom ursprünglichen Merkmalsraum $\mathcal{X}$ in einen transformierten Merkmalsraum $\mathcal{H}$, der als Hilbertraum gewählt wird, abmildern. Abb. 1 zeigt ein Problem der binäre Klassifikation mit der ersten Einschränkung: Keine Gerade trennt die beiden Klassen im ursprünglichen Merkmalsraum, sodass kein linearer Klassifikator eine Entscheidungsgrenze besitzt, die sie korrekt klassifiziert. Eine geeignete Merkmalstransformation macht einen Datensatz „linearer“: Der Zusammenhang zwischen den transformierten Merkmalsvektoren $\phi(\mathbf{x})$ und den

Labels ist in $\mathcal{H}$ näher an einem linearen Zusammenhang als der Zusammenhang zwischen den ursprünglichen Merkmalsvektoren $\mathbf{x}$ und den Labels in $\mathcal{X}$.

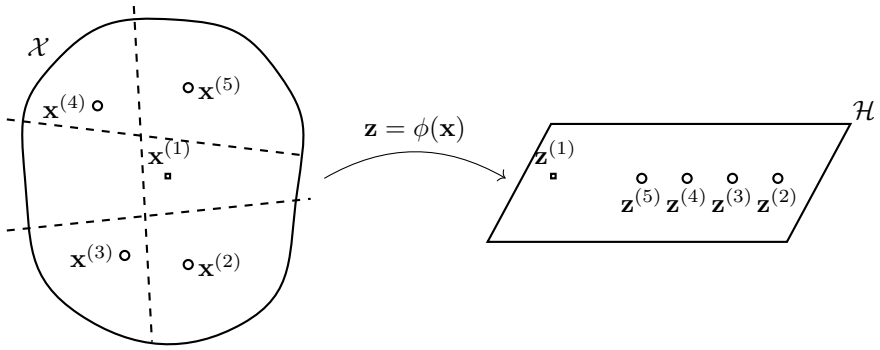

Abbildung 1: Fünf Datenpunkte, charakterisiert durch Merkmalsvektoren $\mathbf{x}^{(r)}$ und Labels $y^{(r)} \in \{\circ, \square\}$ für $r = 1, \ldots, 5$. Die gekrümmte Begrenzung deutet den ursprünglichen Merkmalsraum $\mathcal{X}$ an (links); das Parallelogramm deutet einen Ausschnitt des transformierten Merkmalsraums $\mathcal{H}$ an (rechts), bei dem es sich um einen linearen Raum handelt (siehe Hilbertraum). Mit diesen Merkmalsvektoren ist es nicht möglich, die beiden Klassen durch eine Gerade (welche die Entscheidungsgrenze eines linearer Klassifikators darstellt) zu trennen. Die gestrichelten Linien veranschaulichen drei fehlgeschlagene Versuche: Jeder platziert den Merkmalsvektor $\mathbf{x}^{(1)}$ (Klasse $\square$) auf derselben Seite wie mindestens einen Merkmalsvektor der Klasse $\circ$, sodass ein linearer Klassifikator mit jeder dieser Entscheidungsgrenzen mindestens einen Datenpunkt falsch klassifiziert. Im Gegensatz dazu erlauben die transformierten Merkmalsvektoren $\mathbf{z}^{(r)} = \phi\big(\mathbf{x}^{(r)}\big)$ eine Trennung der Datenpunkte mithilfe eines linearer Klassifikators

Eine nützliche Merkmalstransformation ϕ liefert oft transformierte Merkmalsvektoren in einem Hilbertraum $\mathcal{H}$ hoher (möglicherweise unendlicher) Dimension (Schölkopf and Smola, 2002). Das explizite Berechnen und Speichern der transformierten Merkmalsvektoren $\phi(\mathbf{x})$ kann dann undurchführbar sein. Lineare Methoden greifen jedoch auf die Merkmalsvektoren von Datenpunkteen nur über Skalarprodukte zwischen Paaren von Merkmalsvektoren zu (siehe lineares Modell). Beispielsweise ist die von einer trainierten Stützvektormaschine (SVM) gelieferte Vorhersage eine gewichtete Summe von Skalarprodukte zwischen dem Merkmalsvektor eines neuen Datenpunkts und den Merkmalsvektoren im Trainingsdatensatz (Hastie et al., 2009, Gl. (12.20)). Um eine lineare Methode auf die transformierten Merkmalsvektoren anzuwenden, genügt es daher, die Skalarprodukte $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ für jedes mögliche Paar von Merkmalsvektoren $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ zu kennen (Boser et al., 1992).

Diese Skalarprodukte sind die Werte einer einzigen Funktion zweier Argu-

mente, der Kernel, definiert durch

$$k\big(\mathbf{x}, \mathbf{x}'\big) := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle, \text{ für } \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Der Wert $k\big(\mathbf{x}, \mathbf{x}'\big)$ quantifiziert die Ähnlichkeit der Merkmalsvektoren $\mathbf{x}$ und $\mathbf{x}'$. Als Skalarprodukt transformierter Merkmalsvektoren ist die Kernel symmetrisch, und jede Matrix paarweiser Werte der Kernel ist positiv semidefinit; diese beiden Eigenschaften charakterisieren Kernel (Shalev-Shwartz and Ben-David, 2014, Lemma 16.2). Eine lineare Methode, die auf die transformierten Merkmalsvektoren angewendet wird, benötigt somit nur die Kernel k; die Merkmalstransformation ϕ selbst wird nie ausgewertet.

Kernel-Methoden kehren diese Konstruktion um: Sie gehen von einer Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ aus und konstruieren daraus die Merkmalstransformation. Die Merkmalstransformation bildet einen Merkmalsvektor $\mathbf{x} \in \mathcal{X}$ auf die Funktion $k\big(\mathbf{x}, \cdot\big)$ ab, d. h., der transformierte Merkmalsvektor $\mathbf{z} := \phi(\mathbf{x}) = k\big(\mathbf{x}, \cdot\big)$ ist selbst eine Funktion mit der Definitionsbereich $\mathcal{X}$, für jedes $\mathbf{x} \in \mathcal{X}$. Diese Funktionen gehören zu dem mit der Kernel k assoziierten RKHS (RKHS) $\mathcal{H}_k$ (Aronszajn, 1950, Teil I, Abschn. 2).

Betrachtet seien zwei ursprüngliche Merkmalsvektoren $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Das Skalarprodukt zwischen ihren transformierten Merkmalsvektoren ist eine einzige Auswertung der Kernel,

$$\langle k\big(\mathbf{x}, \cdot\big), k\big(\mathbf{x}', \cdot\big) \rangle = k\big(\mathbf{x}, \mathbf{x}'\big).$$

Allgemeiner ist das Skalarprodukt zwischen dem transformierten Merkmalsvektor $k\big(\mathbf{x}, \cdot\big)$ und einer beliebigen Funktion $h \in \mathcal{H}_k$ eine Punktauswertung,

$$\langle h, k\big(\mathbf{x}, \cdot\big) \rangle = h(\mathbf{x}),$$

was als reproduzierende Eigenschaft des RKHS $\mathcal{H}_k$ bezeichnet wird (Aronszajn, 1950, Teil I, Abschn. 1). Diese reproduzierende Eigenschaft enthält $\langle k\big(\mathbf{x}, \cdot\big), k\big(\mathbf{x}', \cdot\big) \rangle = k\big(\mathbf{x}, \mathbf{x}'\big)$ als Spezialfall, der sich für die Wahl $h = k\big(\mathbf{x}', \cdot\big)$ ergibt. Somit gilt die Schlussfolgerung der Hinrichtung auch hier: Die (möglicherweise unendlichdimensionalen) transformierten Merkmalsvektoren müssen nie explizit berechnet werden.

Abb. 2 veranschaulicht den RKHS $\mathcal{H}_k$ für ein Trainingsdatensatz mit zwei Datenpunkteen: Eine Hypothese h ist ein einzelner Vektor in $\mathcal{H}_k$, die beiden transformierten Merkmalsvektoren spannen einen Unterraum auf, und h besitzt eine orthogonale Projektion auf diesen. Der Wert $\langle h, k\big(\mathbf{x}, \cdot\big) \rangle = h(\mathbf{x})$ ist eine lineare Funktion des transformierten Merkmalsvektors $k\big(\mathbf{x}, \cdot\big) \in \mathcal{H}_k$, aber die auf dem ursprünglichen Merkmalsraum $\mathcal{X}$ induzierte Funktion $\mathbf{x} \mapsto h(\mathbf{x})$ ist im Allgemeinen nichtlinear. Linearität im ursprünglichen Merkmalsvektor $\mathbf{x}$ ist möglicherweise nicht einmal definiert, da der Merkmalsraum $\mathcal{X}$ kein Vektorraum sein muss, z. B. wenn er aus Textdokumenten oder Graphen besteht.

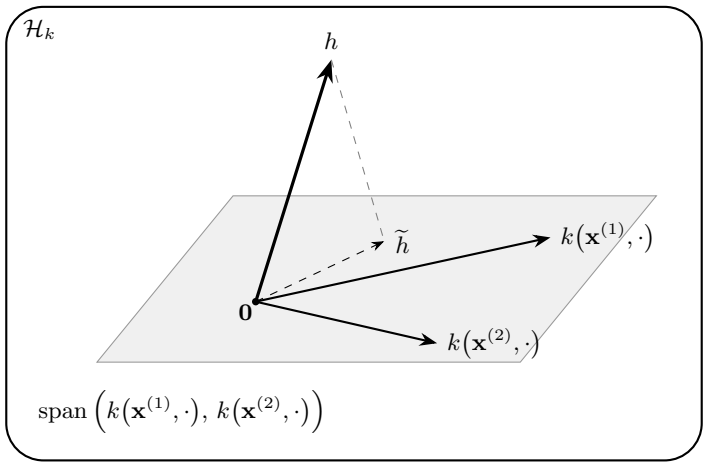

Abbildung 2: Der RKHS $\mathcal{H}_k$ für Trainingsdatensatz mit zwei Datenpunkte, die die Merkmalsvektoren $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ besitzen. Ihre transformierten Merkmalsvektoren $k\big(\mathbf{x}^{(1)},\cdot\big)$ und $k\big(\mathbf{x}^{(2)},\cdot\big)$ spannen einen zweidimensionalen Unterraum auf (schattiertes Parallelogramm). Eine Hypothese h ist ein Vektor in $\mathcal{H}_k$; ihre Vorhersage für einen Datenpunkt mit Merkmalsvektor $\mathbf{x}$ ist das Skalarprodukt $\langle h, k\big(\mathbf{x},\cdot\big)\rangle = h(\mathbf{x})$. Diese Vorhersage ist eine lineare Funktion des transformierten Merkmalsvektors $k\big(\mathbf{x},\cdot\big)$, die auf $\mathcal{X}$ induzierte Funktion $\mathbf{x} \mapsto h(\mathbf{x})$ ist jedoch im Allgemeinen nichtlinear. Der gestrichelte Pfeil stellt die orthogonale Projektion $\widetilde{h}$ von h auf den schattierten Unterraum dar

Kernel-Methoden lassen sich als regularisierte empirische Risikominimierung (RERM) über dem RKHS $\mathcal{H}_k$ formulieren,

$$\min_{h\in\mathcal{H}_k} \frac{1}{m}\sum_{r=1}^{m} L\left(\big(\mathbf{x}^{(r)}, y^{(r)}\big), h\right) + \alpha\|h\|^2_{\mathcal{H}_k},$$

über einem Trainingsdatensatz mit Merkmalsvektoren $\mathbf{x}^{(r)}$ und Labels $y^{(r)}$, für $r = 1,\ldots,m$. Der Parameter der Regularisierung $\alpha > 0$ gewichtet den Strafterm $\|h\|^2_{\mathcal{H}_k}$. Die Verlustfunktion $L\left(\big(\mathbf{x}^{(r)}, y^{(r)}\big), h\right)$ hängt von h nur über die Vorhersage $h\big(\mathbf{x}^{(r)}\big)$ ab, d. h., sie ist eine Funktion ausschließlich der Vorhersage $h\big(\mathbf{x}^{(r)}\big)$ und des Labels $y^{(r)}$. Da $h \in \mathcal{H}_k$ gilt, liefert die reproduzierende Eigenschaft diese Vorhersage als Skalarprodukt, $h\big(\mathbf{x}^{(r)}\big) = \langle h, k\big(\mathbf{x}^{(r)},\cdot\big)\rangle$. Nach dem Representer-Theorem besitzt dieses Optimierungsproblem einen Minimierer der Form $\widehat{h} = \sum_{r=1}^{m} \beta_r k\big(\mathbf{x}^{(r)},\cdot\big)$ mit Entwicklungskoeffizienten $\beta_1,\ldots,\beta_m \in \mathbb{R}$ (Schölkopf and Smola, 2002, Satz 4.2; Shalev-Shwartz and Ben-David, 2014, Satz 16.1).

Abb. 2 veranschaulicht das zugrundeliegende Argument über die Projektion. Wird eine Hypothese h durch ihre orthogonale Projektion $\widetilde{h}$ auf den von $k\big(\mathbf{x}^{(1)},\cdot\big),\ldots,k\big(\mathbf{x}^{(m)},\cdot\big)$ aufgespannten Unterraum ersetzt, bleiben die Vorhersagen auf den Trainingsdatensatz unverändert, und der Strafterm nimmt

niemals zu. Ein Minimierer $\widehat{h}$ liegt daher in diesem Unterraum und besitzt die obige Form.

Das Training reduziert sich somit auf ein Optimierungsproblem in diesen m Koeffizienten. Für eine Verlustfunktion, die konvex in der Vorhersage ist, handelt es sich um ein konvexes Optimierungsproblem: Jede Vorhersage $\widehat{h}\big(\mathbf{x}^{(r')}\big) = \sum_{r=1}^{m} \beta_r k\big(\mathbf{x}^{(r)}, \mathbf{x}^{(r')}\big)$ auf den Trainingsdatensatz ist linear in den Koeffizienten, und der aus der reproduzierenden Eigenschaft erhaltene Strafterm $\left\|\widehat{h}\right\|_{\mathcal{H}_k}^2 = \sum_{r=1}^{m}\sum_{r'=1}^{m} \beta_r \beta_{r'} k\big(\mathbf{x}^{(r)}, \mathbf{x}^{(r')}\big)$ ist eine quadratische Form mit der positiv semidefiniten Matrix der paarweisen Kernel-Werte und folglich konvex. Die Vorhersage $\widehat{h}(\mathbf{x}) = \sum_{r=1}^{m} \beta_r k\big(\mathbf{x}^{(r)}, \mathbf{x}\big)$ erfordert lediglich Auswertungen des Kernels. Beispiele für solche Kernel-Methoden (Cristianini and Shawe-Taylor, 2000) sind die Kernel-SVM, die sich für den Hinge-Verlust ergibt (Hastie et al., 2009, Abschn. 12.3.2), und die Kernel-Ridge-Regression, die sich für den quadratischer Verlust ergibt (Hastie et al., 2009, Abschn. 5.8.1). Für $\mathcal{X} = \mathbb{R}^d$ ist eine weit verbreitete Wahl für den Kernel k der Gauß-Kernel $k\big(\mathbf{x}, \mathbf{x}'\big) = \exp\big(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\sigma^2)\big)$ mit Bandbreite $\sigma > 0$. Kernel stehen auch für Datenpunkte ohne numerische Merkmale zur Verfügung, wie etwa Zeichenketten und Graphen (Schölkopf and Smola, 2002). Abb. 3 zeigt die nichtlineare Entscheidungsgrenze $\widehat{h}(\mathbf{x}) = 0$ einer Hypothese $\widehat{h}$, die mit einer Gauß-Kernel-Methode mit quadratischer Verlust aus Trainingsdatensatz bestehend aus zwei konzentrischen Ringen gelernt wurde, welche kein linearer Klassifikator trennt. Hier passt die Kernel-Ridge-Regression die Labels $y^{(r)} \in \{-1, +1\}$ an, und die Vorhersage ist das Vorzeichen des angepassten Werts; diese Konstruktion wird als regularisierte Kleinste-Quadrate-Klassifikation bezeichnet (Rifkin et al., 2003).

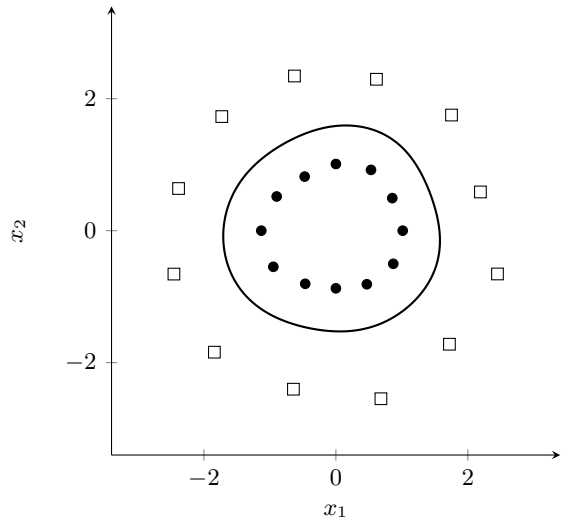

Abbildung 3: Eine Gauß-Kernel-Methode (Kernel-Ridge-Regression mit $\sigma = 0.8$, $\alpha = 10^{-3}$), trainiert auf Trainingsdatensatz aus zwei konzentrischen Ringen. Datenpunkte mit Label $y = +1$ sind als ausgefüllte Kreise dargestellt, solche mit $y = -1$ als offene Quadrate. Die Entscheidungsgrenze $\widehat{h}(\mathbf{x}) = 0$ der gelernten Hypothese (durchgezogene geschlossene Kurve) umschließt den inneren Ring. Der Datensatz ist im ursprünglichen Merkmalsraum nicht linear trennbar. Daten erzeugt mit `pythondemos/kernelmethod.py`

Kernel-Methoden setzen die Glattheitsannahme über die RKHS-Norm durch. Diese Norm $\|h\|_{\mathcal{H}_k}$ quantifiziert die Glattheit einer Hypothese h: Sie ist groß für eine stark schwankende h und klein für eine glatte h. Darüber hinaus erzwingen unterschiedliche Kernel unterschiedliche Arten von Glattheit. Der Gauß-Kernel ist unendlich oft differenzierbar (Rasmussen and Williams, 2006, Abschn. 4.2.1), und damit auch jedes $h(\mathbf{x}) = \langle h, k\big(\mathbf{x}, \cdot\big)\rangle$ in seinem RKHS; die RKHS-Norm bestraft Ableitungen aller Ordnungen und folglich hochfrequente Komponenten (Schölkopf and Smola, 2002, Abschn. 4.4.2). Die Matérn-Kernel bilden eine durch einen Glattheits-Parameter $\nu > 0$ indizierte Familie von Kernel: Ihr RKHS stimmt, mit äquivalenten Normen, mit dem Sobolev-Raum einer durch ν bestimmten Ordnung überein, dessen Funktionen Ableitungen bis zu dieser Ordnung im schwachen Sinne besitzen (Rasmussen and Williams, 2006, Gl. 4.15; Kanagawa et al., 2018, Beispiel 2.6). Der Strafterm $\alpha\|h\|^2_{\mathcal{H}_k}$ in der obigen RERM ist die quadrierte RKHS-Norm. Er setzt daher Glattheit wie durch $\|h\|_{\mathcal{H}_k}$ quantifiziert um: Unter Hypothesen mit demselben Trainingsfehler ist die RERM-Zielfunktion für diejenige Hypothese am kleinsten, die die kleinste RKHS-Norm aufweist.

Die Wahl des Kernels bestimmt nicht nur das Training, sondern auch die Vorhersagen der gelernten Hypothese: Für lokalisierte Kernel, wie etwa den Gauß-Kernel, wird die Vorhersage $\widehat{h}(\mathbf{x})$ von denjenigen Datenpunkte in den

Trainingsdatensatz dominiert, deren Merkmalsvektoren nahe bei $\mathbf{x}$ liegen. In diesem Sinne setzen Kernel-Methoden die Glattheitsannahme um: Datenpunkte mit nahe beieinander liegenden Merkmalsvektoren erhalten ähnliche Vorhersagen. Die Lokalität der Vorhersagen macht Kernel-Methoden zudem zu einem weich gewichteten Analogon von k-nächste Nachbarn (k-NN): Anstatt die Labels einer festen Anzahl nächstgelegener Datenpunkte zu mitteln, gewichtet die Vorhersage alle Datenpunkte in den Trainingsdatensatz mit dem Kernel-Wert $k\big(\mathbf{x}^{(r)}, \mathbf{x}\big)$ (Hastie et al., 2009, Abschn. 6.1).

Synonyme: Kernel-Maschine.

Siehe auch: Kernel, Merkmalstransformation, Hilbertraum, RKHS, SVM, Kernel-Ridge-Regression, Glattheitsannahme.

Querverweise

Literatur

1. Boser et al. (1992). A Training Algorithm for Optimal Margin Classifiers. In: Proc. 5th Annual Workshop on Computational Learning Theory (COLT), pp. 144–152. ACM.
2. Cristianini and Shawe-Taylor (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge Univ. Press, New York, NY, USA.
3. Lampert (2009). Kernel Methods in Computer Vision. Found. Trends Comput. Graph. Vis., 4(3), pp. 193–285.
4. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.
5. Rasmussen and Williams (2006). Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA.
6. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.
7. Aronszajn (1950). Theory of Reproducing Kernels. Trans. Am. Math. Soc., 68(3), pp. 337–404.
8. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.
9. Kanagawa et al. (2018). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. URL: https://arxiv.org/abs/1807.02582.
10. Rifkin et al. (2003). Regularized Least-Squares Classification. In: Advances in Learning Theory: Methods, Model and Applications, pp. 131–154. IOS Press.