Kernel-Ridge-Regression

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

Die Kernel-Ridge-Regression wendet die Ridge-Regression auf transformierte Merkmalsvektoren an, die aus einem Kernel konstruiert wurden, und lernt dadurch eine nichtlineare Hypothese. Sie führt regularized empirical risk minimization (RERM) über Hypothese-Abbildungen durch, die Elemente eines reproducing kernel Hilbert space (RKHS) sind. Nach dem Representer-Theorem ist die gelernte Hypothese eine gewichtete Summe von Kernel-Auswertungen mit Koeffizienten, die in geschlossener Form durch das Lösen eines einzelnen Systems von lineare Gleichungen erhalten werden. Der Kernel ist ein Hyperparameter, der an die Struktur des Merkmalsraum angepasst werden muss: Für die Stunden eines Tages macht ein Kernel, der aus der kürzesten Pfadlänge auf dem Zyklus-Graph der Stunden aufgebaut ist, die Stunden 23 und 0 so ähnlich wie die Stunden 0 und 1. Wie die Ridge-Regression entspricht auch die Kernel-Ridge-Regression der empirische Risikominimierung (ERM) auf einem erweiterten Trainingsdatensatz: Die transformierten Merkmalsvektoren werden durch Realisierungen eines Gauß-Prozesses mit Mittelwert Null gestört, dessen covariance function der Kernel ist, skaliert mit dem Regularisierung Parameter.

Definition

Die Vorhersage der Temperatur in Krems an der Donau für eine gegebene Tagesstunde ist eine machine learning (ML)-Anwendung. Ein Datenpunkt ist eine Stunde eines Tages. Ihr Merkmal $x \in \{0, \ldots, 23\}$ ist die Tagesstunde, und ihr Label y ist die in dieser Stunde gemessene Temperatur. Der Merkmalsraum ist $\mathcal{X} = \{0, \ldots, 23\}$ und der Labelraum ist $\mathcal{Y} = \mathbb{R}$, daher ist die Anwendung ein regression-Problem. Sie erfordert einen Hypothesenraum nichtlinearer Hypothese-Abbildungen $h(x)$, da sich die Temperatur selten linear mit der Tagesstunde ändert. Abb. 2 zeigt 24 der Datenpunkte eines Trainingsdatensatz der Größe $m = 744$, die die stündlichen Temperaturmessungen vom Juli 2024 an der GeoSphere Wetterstation Krems an der Donau (GeoSphere Austria, 2026) darstellen. Die Beziehung zwischen Merkmal und Label ist nichtlinear: Die Temperatur fällt bis zum frühen Morgen, steigt bis zum Nachmittag und fällt dann wieder.

Die Kernel-Ridge-Regression erweitert die Ridge-Regression, die ein lineares Modell verwendet, indem sie das ursprüngliche Merkmal x durch einen transfor-

mierten Merkmalsvektor $\phi(x)$ ersetzt. Die Merkmalstransformation ϕ wird aus einem Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Schölkopf and Smola, 2002) konstruiert. Der Wert $k\big(x, x'\big)$ misst die Ähnlichkeit zwischen zwei Datenpunkte mit Merkmal-Werten x und x'. Der Kernel ist ein Hyperparameter der Kernel-Ridge-Regression, der an die Struktur des Merkmalsraum angepasst werden muss. Für das Temperaturbeispiel sind die Stunden eines Tages die Knoten eines Zyklus-Graph (siehe Abb. 1), und die Ähnlichkeit zweier Merkmal-Werte wird durch die Länge $\min\{|x - x'|, 24 - |x - x'|\}$ des kürzesten Pfades zwischen den entsprechenden Knoten bestimmt,

$$k\big(x, x'\big) := \exp\big(- \min\{|x - x'|, 24 - |x - x'|\}/\sigma\big), \tag{1}$$

mit einer Bandbreite $\sigma > 0$, gemessen in Stunden. Die Stunden 23 und 0 sind dann so ähnlich wie die Stunden 0 und 1, was der absolute Unterschied $|x - x'|$ allein nicht liefern würde.

Da $\mathcal{X}$ 24 Elemente hat, ist die Funktion (1) genau dann ein Kernel, wenn die 24×24 Matrix ihrer Werte positive semi-definite (psd) ist; für $\sigma = 3$ beträgt ihr kleinster Eigenwert 0.16. Der Kernel (1) verwendet die Exponentialfunktion der kürzesten Pfadlänge und nicht deren Gaußsche Funktion $\exp\big(- \min\{|x - x'|, 24 - |x - x'|\}^2/(2\sigma^2)\big)$. Die Gaußsche Funktion einer Abstand ist nur dann ein Kernel für jede Bandbreite, wenn die Datenpunkte mit dieser Abstand in einem Euclidean space platziert werden können (Feragen et al., 2015). Die Knoten eines Zyklus-Graph mit der kürzesten Pfadlänge können dies nicht: Für $\sigma = 3$ hat die 24×24 Matrix der Gaußschen Werte bereits einen negativen Eigenwert.

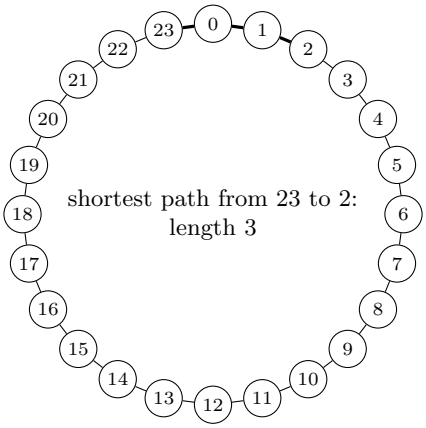

Abbildung 1: Die Stunden eines Tages als Knoten eines Zyklen-Graph: Jede Stunde ist mit der nächsten verbunden, und 23 ist mit 0 verbunden. Das Kernel (1) misst die Ähnlichkeit zweier Stunden durch die Länge des kürzesten Pfades zwischen ihren Knoten. Die dicken Kanten bilden den kürzesten Pfad von 23 nach 2, mit der Länge 3, wobei die absolute Differenz $|23 - 2| = 21$ die beiden Stunden unähnlich machen würde

Mit dem Kernel (1) für $\sigma = 3$ und $\alpha = 10^{-2}$ folgt die aus den 744 Daten-

punkte gelernte Hypothese dem Tageszyklus (siehe Abb. 2), mit einem Training mean squared error (MSE) von 9.0, und ihre Werte zu den Stunden 23 und 0 unterscheiden sich um 0.5 Grad. Die Ridge-Regression auf der rohen Stunde kann nur eine lineare Hypothese lernen, mit einem Training MSE von 18.6, deren Werte zu den Stunden 23 und 0 sich um 5.2 Grad unterscheiden.

Mathematisch lernt die Kernel-Ridge-Regression eine Hypothese aus dem reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ eines Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, indem sie den bestraften durchschnittlichen squared error loss auf einem Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}$ mit numerischen Labels $y^{(r)} \in \mathbb{R}$ minimiert,

$$\widehat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_k} \frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - h\left(\mathbf{x}^{(r)}\right)\right)^2 + \alpha \|h\|_{\mathcal{H}_k}^2. \tag{2}$$

Aufgrund der reproduzierenden Eigenschaft von $\mathcal{H}_k$ (siehe RKHS) erfüllt jede Hypothese $h \in \mathcal{H}_k$ die Bedingung $h(\mathbf{x}) = \langle h, \phi(\mathbf{x}) \rangle_{\mathcal{H}_k}$, wobei $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ das Skalarprodukt von $\mathcal{H}_k$ und $\phi(\mathbf{x}) = k\left(\mathbf{x}, \cdot\right)$ der transformierte Merkmalsvektor ist. Die gelernte Hypothese $\widehat{h}$ ist daher linear im transformierten Merkmalsvektor und typischerweise nichtlinear als Abbildung auf dem ursprünglichen Merkmalsraum. Sie ist nur dann linear auf $\mathcal{X}$, wenn die Merkmalstransformation selbst linear ist, was bei einem Kernel der Form $k\left(\mathbf{x}, \mathbf{x}'\right) = \mathbf{x}^\top \mathbf{Q} \mathbf{x}'$ mit einer psd Matrix $\mathbf{Q}$ der Fall ist. Algorithmus 1 beschreibt die Methode vollständig; die folgenden Absätze leiten ihre Schritte ab.

Algorithm 1 Kernel-Ridge-Regression (Murphy, 2012, Sect. 14.4.3)

Input: Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}$, Kernel k, Regularisierung Parameter $\alpha > 0$

1: compute the Gram Matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ with entries $K_{r,r'} = k\left(\mathbf{x}^{(r)}, \mathbf{x}^{(r')}\right)$

2: solve the system of lineare Gleichungen $\left(\mathbf{K} + \alpha m \mathbf{I}\right)\boldsymbol{\beta} = \mathbf{y}$ with the label vector $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(m)}\right)^\top$ for $\widehat{\boldsymbol{\beta}}$

Output: the Hypothese $\widehat{h}(\mathbf{x}) = \sum_{r=1}^{m} \widehat{\beta}_r k\left(\mathbf{x}^{(r)}, \mathbf{x}\right)$

Der Strafterm im Kernel-Ridge-Regression-Problem (2) ist die skalierte quadrierte RKHS Norm $\alpha \|h\|_{\mathcal{H}_k}^2$ mit einem Regularisierung Parameter $\alpha > 0$. Die resultierende Methode ist eine Instanz von regularized empirical risk minimization (RERM), ebenso wie die Ridge-Regression, deren Strafterm die skalierte quadrierte Euclidean norm $\alpha \|\mathbf{w}\|_2^2$ der Modellparameter ist (Hastie et al., 2009, Sect. 5.8.2). Der durchschnittliche squared error loss ist eine konvex Funktion von h, und die quadrierte Norm ist strictly convex. Das Kernel-Ridge-Regression-Problem (2) hat daher für jedes Trainingsdatensatz einen eindeutigen Minimierer $\widehat{h}$, ebenso wie die Ridge-Regression.

Nach dem Representer-Theorem (siehe Kernel-Methode) (Schölkopf and Smola, 2002, Thm. 4.2) hat der Minimierer des Kernel-Ridge-Regression-Problems (2) die Form $\widehat{h} = \sum_{r=1}^{m} \beta_r k\left(\mathbf{x}^{(r)}, \cdot\right)$ mit einem Koeffizienten-Vektor

$\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_m\big)^\top$. Das Einsetzen dieser Form in das Kernel-Ridge-Regression-Problem (2) ergibt das endlich-dimensionale Optimierungsproblem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \frac{1}{m} \|\mathbf{y} - \mathbf{K}\boldsymbol{\beta}\|_2^2 + \alpha\, \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta},$$

mit dem label vector $\mathbf{y} = \big(y^{(1)}, \ldots, y^{(m)}\big)^\top$ und der Gram-Matrix $\mathbf{K}$ mit Einträgen $K_{r,r'} = k\big(\mathbf{x}^{(r)}, \mathbf{x}^{(r')}\big)$ (siehe Kernel).
Gemäß der zero-gradient condition erfüllt ein Koeffizienten-Vektor, der dieses Optimierungsproblem löst, das System von lineare Gleichungen

$$\mathbf{K}\big(\big(\mathbf{K} + \alpha m \mathbf{I}\big)\boldsymbol{\beta} - \mathbf{y}\big) = \mathbf{0}.$$

Da $\mathbf{K}$ psd ist und $\alpha > 0$, ist die Matrix $\mathbf{K} + \alpha m \mathbf{I}$ invertible, und ein Minimierer wird in geschlossener Form erhalten als

$$\widehat{\boldsymbol{\beta}} = \big(\mathbf{K} + \alpha m \mathbf{I}\big)^{-1} \mathbf{y}. \tag{3}$$

Die Vorhersage für einen Datenpunkt mit Merkmalsvektor $\mathbf{x}$ ist

$$\widehat{h}(\mathbf{x}) = \sum_{r=1}^{m} \widehat{\beta}_r k\big(\mathbf{x}^{(r)}, \mathbf{x}\big). \tag{4}$$

Das Training ist eine einmalige Berechnung: Die Kernel-Ridge-Regression erfordert keine iterative Optimierungsverfahren (und somit keine Fixpunktinterpretation), sondern nur die eindeutige Lösung des $m \times m$-Systems von lineare Gleichungen mit der invertible Matrix $\mathbf{K} + \alpha m \mathbf{I}$, welche die geschlossene Form (3) ist, gefolgt von der Vorhersage (4).

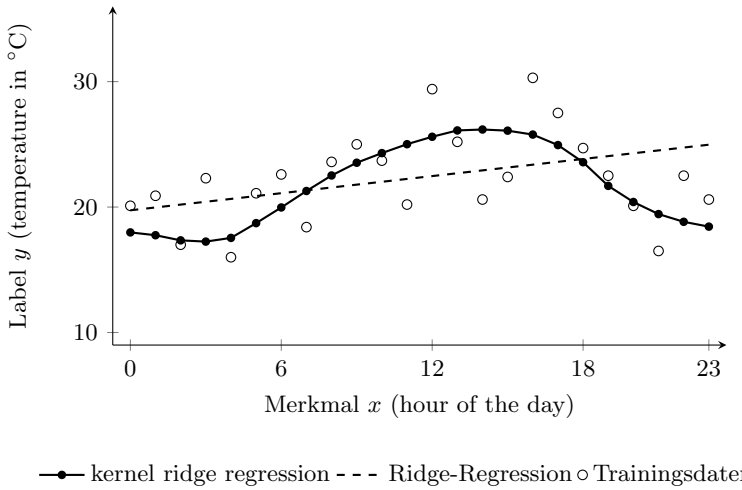

Abbildung 2: Die durch Kernel-Ridge-Regression mit dem Kernel (1) ($\sigma = 3$, $\alpha = 10^{-2}$) gelernte Hypothese unter Verwendung von $m = 744$ Datenpunkte, die die stündlichen Temperaturen vom Juli 2024 in Krems an der Donau darstellen; 24 der Datenpunkte des Trainingsdatensatz sind gezeichnet (Kreise), jeder 31. in zeitlicher Reihenfolge, einer für jede Stunde des Tages. Die gelernte Hypothese $\widehat{h}$ (durchgezogene Linie durch ihre Werte zu den 24 Stunden, Punkte) folgt dem Tageszyklus, mit einem Training-MSE von 9.0; die Ridge-Regression auf der rohen Stunde (gestrichelte Linie) lernt eine lineare Hypothese, mit einem Training-MSE von 18.6. Daten generiert von `pythondemos/kernelridgeregression.py`

Für den linearen Kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$ auf $\mathcal{X} = \mathbb{R}^d$ erzeugt die Kernel-Ridge-Regression dieselben Vorhersagen wie die Ridge-Regression, angewendet auf die rohen Merkmalsvektoren (Rasmussen and Williams, 2006, Sect. 2.1, Eqs. (2.11)–(2.12)). Die Gram-Matrix ist dann $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$, mit der Merkmalsmatrix $\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)^\top$. Das Ausmultiplizieren beider Seiten zeigt, dass $\left(\mathbf{X}^\top\mathbf{X} + \alpha m\mathbf{I}\right)\mathbf{X}^\top = \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \alpha m\mathbf{I}\right)$, und somit $\mathbf{X}^\top\left(\mathbf{K} + \alpha m\mathbf{I}\right)^{-1} = \left(\mathbf{X}^\top\mathbf{X} + \alpha m\mathbf{I}\right)^{-1}\mathbf{X}^\top$. Die Vorhersage $\widehat{h}(\mathbf{x}) = \mathbf{x}^\top\mathbf{X}^\top\widehat{\boldsymbol{\beta}}$ der Kernel-Ridge-Regression entspricht daher $\mathbf{x}^\top\left(\mathbf{X}^\top\mathbf{X} + \alpha m\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, der Vorhersage der Ridge-Regression.

Wie die Ridge-Regression hat die Kernel-Ridge-Regression eine Interpretation als eine Form der Datenerweiterung. Das Training mit zu den Merkmalsvektoren hinzugefügtem Rauschen ist äquivalent zum Hinzufügen eines Strafterm (Bishop, 1995, Gleichungen (10)–(11)). Das Ersetzen jedes Datenpunkt durch eine darauf zentrierte Wahrscheinlichkeitsverteilung ist die vicinale Form der empirische Risikominimierung (ERM) (Chapelle et al., 2001): Der durchschnittliche Verlust über die Datenpunkte wird durch den erwarteten Verlust unter diesen Wahrscheinlichkeitsverteilungen ersetzt, wobei jede die Umgebung eines Datenpunkt beschreibt. Ihr Effekt auf Kernel-Methoden ist eine Regulari-

sierung der gelernten Hypothese (Dao et al., 2019).

Das folgende Beispiel verdeutlicht die Äquivalenz für einen linearen Kernel in der Ebene, sodass die gestörten Kopien gezeichnet werden können. Betrachten Sie Datenpunkte mit zwei Merkmale, den Merkmalsraum $\mathcal{X} = \mathbb{R}^2$, und den Kernel

$$k\big(\mathbf{x}, \mathbf{x}'\big) := \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}', \tag{5}$$

mit einer positive definite (pd) Matrix $\mathbf{C} \in \mathbb{R}^{2\times 2}$. Dieser Kernel ist ein Skalarprodukt von $\mathbb{R}^2$ mit einer modifizierten Metrik: Die quadrierte Norm $k\big(\mathbf{x}, \mathbf{x}\big) = \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}$ skaliert die beiden Merkmale unterschiedlich und koppelt sie, und ihre Einheitskugel ist eine Ellipse anstelle des Kreises der Euclidean norm. Der RKHS $\mathcal{H}_k$ dieses Kernel ist $\mathbb{R}^2$ mit diesem Skalarprodukt. Der transformierte Merkmalsvektor $k\big(\mathbf{x}, \cdot\big)$ ist die lineare Funktion $\mathbf{x}' \mapsto \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}'$, jede Hypothese in $\mathcal{H}_k$ ist ein lineares Modell $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ mit Modellparameter $\mathbf{w} \in \mathbb{R}^2$, und ihre RKHS-Norm ist $\big\|h^{(\mathbf{w})}\big\|_{\mathcal{H}_k}^2 = \mathbf{w}^\top \mathbf{C} \mathbf{w}$.

Das Kernel-Ridge-Regression-Problem (2) lautet daher

$$\widehat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^2} \frac{1}{m} \sum_{r=1}^{m} \big(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\big)^2 + \alpha\, \mathbf{w}^\top \mathbf{C} \mathbf{w}, \tag{6}$$

was Ridge-Regression ist, wobei der Strafterm $\alpha \|\mathbf{w}\|_2^2$ durch $\alpha\, \mathbf{w}^\top \mathbf{C} \mathbf{w}$ ersetzt wurde. Ihr eindeutiger Minimierer ist $\widehat{\mathbf{w}} = \big(\mathbf{X}^\top \mathbf{X} + \alpha m \mathbf{C}\big)^{-1} \mathbf{X}^\top \mathbf{y}$ mit der Merkmalsmatrix $\mathbf{X} = \big(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\big)^\top$, und er stimmt mit $\mathbf{C}^{-1} \mathbf{X}^\top \widehat{\boldsymbol{\beta}}$ für die Koeffizienten (3) mit der Gram-Matrix $\mathbf{K} = \mathbf{X} \mathbf{C}^{-1} \mathbf{X}^\top$ überein. Für $\mathbf{C} = \mathbf{I}$ ist der Kernel (5) der lineare Kernel und das penalisierte Problem (6) ist die Ridge-Regression selbst.

Das Hinzufügen des Strafterm im penalisierten Problem (6) ist äquivalent zur ERM auf einem augmentierten Trainingsdatensatz, d.h. zur Berechnung des durchschnittlichen squared error loss von $h^{(\mathbf{w})}$ darauf. Dieses augmentierte Trainingsdatensatz wird aus dem ursprünglichen Trainingsdatensatz gewonnen. Für jeden Datenpunkt $\big(\mathbf{x}^{(r)}, y^{(r)}\big)$, $r = 1, \ldots, m$, wird der Merkmalsvektor $\mathbf{x}^{(r)}$ durch Realisierungen unendlich vieler unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) $\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}$ mit Gaußsche Zufallsvariablen $\boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}\,(\mathbf{0}, \alpha \mathbf{C})$ ersetzt. Das Label $y^{(r)}$ bleibt unverändert.

Der Strafterm ist genau der Betrag, um den der durchschnittliche squared error loss von $h^{(\mathbf{w})}$ über die gestörten Kopien seinen squared error loss auf dem ursprünglichen Datenpunkt übersteigt. Da $\mathbb{E}\{\boldsymbol{\varepsilon}^{(r)}\} = \mathbf{0}$, ist die Zunahme

$$\mathbb{E}\big\{\big(\mathbf{w}^\top \boldsymbol{\varepsilon}^{(r)}\big)^2\big\} = \mathbf{w}^\top \mathbb{E}\big\{\boldsymbol{\varepsilon}^{(r)} \big(\boldsymbol{\varepsilon}^{(r)}\big)^\top\big\} \mathbf{w} = \alpha\, \mathbf{w}^\top \mathbf{C} \mathbf{w}.$$

Die Äquivalenz ist exakt für das lineares Modell $h^{(\mathbf{w})}$, während sie für eine nichtlineare Hypothese bis auf Terme höherer Ordnung in der Rauschamplitude gilt (Bishop, 1995, Gleichungen (10)–(11)). Wie bei der Ridge-Regression ist diese Zunahme des durchschnittlichen Verlust unter Datenerweiterung eine Schätzung für den generalization gap (Hastie et al., 2009, Sect. 7.4). Dementsprechend stellt

die lineare Regression auf einem großen augmentierten Trainingsdatensatz die Vorhersagen der Kernel-Ridge-Regression wieder her.
Das Kernel bestimmt die Form dieser Störungen (siehe Abb. 3). Die Kovarianzmatrix $\alpha\mathbf{C}$ von $\boldsymbol{\varepsilon}^{(r)}$ ist die skalierte Inverse der Matrix im Kernel (5). Die Ein-Standardabweichungs-Kontur der gestörten Kopien ist daher

$$\left(\mathbf{x} - \mathbf{x}^{(r)}\right)^\top \mathbf{C}^{-1} \left(\mathbf{x} - \mathbf{x}^{(r)}\right) = \alpha.$$

Diese Kontur ist die Menge der Merkmalsvektoren $\mathbf{x}$ mit $k\left(\mathbf{x}-\mathbf{x}^{(r)}, \mathbf{x}-\mathbf{x}^{(r)}\right) = \alpha$, d.h. die Kugel mit Radius $\sqrt{\alpha}$ um $\mathbf{x}^{(r)}$ in der durch das Kernel induzierten Norm. Für $\mathbf{C} = \mathbf{I}$ ist diese Kugel ein Kreis mit Radius $\sqrt{\alpha}$: Die Störungen in der Datenerweiterung-Interpretation der Ridge-Regression haben die Kovarianzmatrix $\alpha\mathbf{I}$ und keine bevorzugte Richtung. Für eine allgemeine $\mathbf{C}$ ist es eine Ellipse, sodass die um die Datenpunkte gezeichneten Ellipsen die Kugeln mit Radius $\sqrt{\alpha}$ der durch das Kernel induzierten Norm sind. Die Größe einer Störung wird durch ihre Kernel-Norm gemessen, und der Strafterm $\alpha\, \mathbf{w}^\top \mathbf{C} \mathbf{w}$ ist der durchschnittliche squared error loss, den Störungen der Kernel-Norm $\sqrt{\alpha}$ verursachen.

Die Ellipsen zeigen auch die Eigenwertzerlegung der Matrix $\mathbf{C}$. Das Einsetzen der Eigenwertzerlegung $\mathbf{C} = \sum_{j=1}^{2} \lambda_j \mathbf{u}^{(j)} \left(\mathbf{u}^{(j)}\right)^\top$, mit Eigenwerte $\lambda_1 \geq \lambda_2 > 0$ und orthonormalen Eigenvektoren $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}$, in die Kontur ergibt

$$\sum_{j=1}^{2} \frac{\left(\left(\mathbf{u}^{(j)}\right)^\top \left(\mathbf{x} - \mathbf{x}^{(r)}\right)\right)^2}{\alpha\lambda_j} = 1.$$

Die Hauptachsen jeder Ellipse zeigen daher entlang der Eigenvektoren von $\mathbf{C}$, und die Halbachsen haben die Längen $\sqrt{\alpha\lambda_1}$ und $\sqrt{\alpha\lambda_2}$. Die Störungen breiten sich am stärksten entlang des Eigenvektor mit dem größten Eigenwert aus, und der Strafterm wächst mit dem Quadrat einer Verschiebung von $\mathbf{w}$ entlang $\mathbf{u}^{(j)}$, proportional zu λ_j.

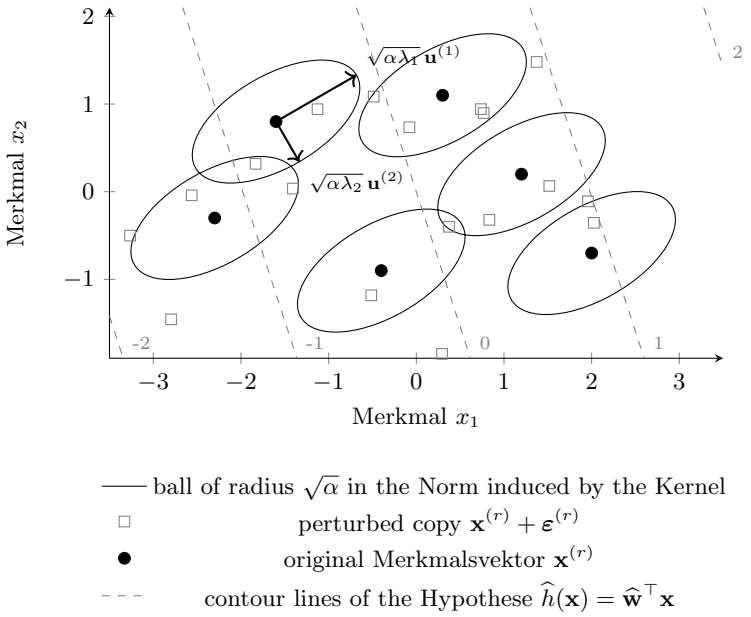

Abbildung 3: Die Datenerweiterung-Interpretation der Kernel-Ridge-Regression mit dem Kernel (5), für einen Trainingsdatensatz von sechs Datenpunkte mit zwei Merkmale, $\alpha = 1/2$, und einer Matrix $\mathbf{C}$ mit Eigenwerte $\lambda_1 = 1.5^2$, $\lambda_2 = 0.75^2$, und Eigenvektoren, die um 30° gegen die Merkmal-Achsen gedreht sind. Jeder ursprüngliche Datenpunkt mit Merkmalsvektor $\mathbf{x}^{(r)}$ (gefüllter Kreis) wird durch drei Kopien $\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}$ (offene Quadrate) ersetzt, deren Merkmalsvektoren durch Gaußsche Zufallsvariablen mit Kovarianzmatrix $\alpha\mathbf{C}$ gestört werden, wobei das Label beibehalten wird. Die Ellipse um jeden Datenpunkt ist die Ein-Standardabweichungs-Kontur ihrer Kopien, die Kugel mit Radius $\sqrt{\alpha}$ in der durch das Kernel induzierten Norm; für $\mathbf{C} = \mathbf{I}$ ist es der Kreis mit Radius $\sqrt{\alpha}$ der isotropen Störungen der Ridge-Regression. Die Pfeile auf einer Ellipse sind ihre Hauptachsen $\sqrt{\alpha\lambda_j}\,\mathbf{u}^{(j)}$, die Eigenvektoren von $\mathbf{C}$, skaliert mit den Quadratwurzeln von α mal ihren Eigenwerte. Die gestrichelten Linien sind Konturlinien der Hypothese $\widehat{h}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x}$, die aus den sechs Datenpunkte durch das regularisierte Problem (6) gelernt wurde, beschriftet mit ihrem Wert: Eine entlang einer Konturlinie verschobene Kopie behält ihre Vorhersage, und eine über die Konturlinien verschobene Kopie ändert sie proportional zur Verschiebung. Daten generiert von `pythondemos/kernelridgeregression.py`

Für das Kernel (5) wirkt die Störung linear auf die transformierten Merkmalsvektoren, $k\big(\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}, \cdot\big) = k\big(\mathbf{x}^{(r)}, \cdot\big) + k\big(\boldsymbol{\varepsilon}^{(r)}, \cdot\big)$. Die zufällige Funktion $k\big(\boldsymbol{\varepsilon}^{(r)}, \cdot\big)$ ist daher ein Nullmittelwert-Gaussian process (GP), dessen covariance function das durch den Regularisierungs-Parameter skalierte Kernel ist,

$$\mathbb{E}\big\{k\big(\boldsymbol{\varepsilon}^{(r)}, \mathbf{x}\big)\, k\big(\boldsymbol{\varepsilon}^{(r)}, \mathbf{x}'\big)\big\} = \mathbf{x}^\top \mathbf{C}^{-1} \mathbb{E}\big\{\boldsymbol{\varepsilon}^{(r)} \big(\boldsymbol{\varepsilon}^{(r)}\big)^\top\big\} \mathbf{C}^{-1} \mathbf{x}' = \alpha\, k\big(\mathbf{x}, \mathbf{x}'\big). \quad (7)$$

Für ein allgemeines Kernel können die rohen Merkmalsvektoren auf diese Weise nicht gestört werden. Die Merkmalstransformation ϕ ist nichtlinear, sodass Rauschen, das zu $\mathbf{x}^{(r)}$ hinzugefügt wird, keine zufällige Funktion mit Nullmittelwert zu $\phi(\mathbf{x}^{(r)})$ hinzufügt. In einem endlichen Merkmalsraum wie den Stunden eines Tages gibt es überhaupt kein Rauschen hinzuzufügen. Die Störung des transformierten Merkmalsvektor selbst bleibt definiert: Sie ersetzt $\phi(\mathbf{x}^{(r)}) = k\big(\mathbf{x}^{(r)}, \cdot\big)$ durch $\phi(\mathbf{x}^{(r)}) + g^{(r)}$ mit einem Nullmittelwert-GP $g^{(r)}$, dessen covariance function αk ist. Da $h(\mathbf{x}) = \langle h, \phi(\mathbf{x})\rangle_{\mathcal{H}_k}$, ändert die Störung die Vorhersage um $\langle h, g^{(r)}\rangle_{\mathcal{H}_k}$. Die Entwicklung von h und $g^{(r)}$ in den Eigenfunktionen des Kernel (Kanagawa et al., 2018, Thms. 4.2 and 4.3) ergibt

$$\mathbb{E}\big\{\langle h, g^{(r)}\rangle^2_{\mathcal{H}_k}\big\} = \alpha\|h\|^2_{\mathcal{H}_k}, \tag{8}$$

somit ist der Strafterm die erwartete Zunahme des squared error loss unter dieser Störung für jedes Kernel. Für ein Kernel mit einem unendlichdimensionalen RKHS sind die Realisierungen von $g^{(r)}$ keine Elemente von $\mathcal{H}_k$ (Kanagawa et al., 2018, Cor. 4.10). Das Skalarprodukt in der erwarteten Zunahme (8) ist dann der Grenzwert der Partialsummen der Entwicklung, der konvergiert, da $\|h\|_{\mathcal{H}_k}$ endlich ist. Für die Stunden eines Tages ist $\phi(x)$ die Spalte für Stunde x der 24×24 Matrix der Kernel-Werte, und $g^{(r)}$ ist ein Gaußscher Vektor, dessen Kovarianzmatrix α mal diese Matrix ist. Bis auf die Skalierung mit α in der covariance function (7) ist dies der GP mit covariance function k und Rausch-Varianz αm, dessen posterior mean die Kernel-Ridge-Regression berechnet: Die Vorhersage $\widehat{h}(\mathbf{x})$ stimmt mit diesem Mittelwert überein (Kanagawa et al., 2018, Prop. 3.6).

Siehe auch: Kernel-Methode, Ridge-Regression, Kernel, RKHS, RERM, squared error loss, regression, GP, Datenerweiterung, Merkmalstransformation, Strafterm.

Querverweise

Literatur

1. Bishop (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), pp. 108–116.

2. Chapelle et al. (2001). Vicinal Risk Minimization. In: Advances in Neural Information Processing Systems 13 (NIPS 2000), pp. 416–422. MIT Press.

3. Dao et al. (2019). A Kernel Theory of Modern Data Augmentation. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 1528–1537. PMLR.

4. Feragen et al. (2015). Geodesic exponential kernels: When curvature and linearity conflict. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3032–3042.

5. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/hist v2-1d.

6. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

7. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

8. Rasmussen and Williams (2006). Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA.

9. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

10. Kanagawa et al. (2018). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. URL: https://arxiv.org/abs/1807.02582.