probabilistische Hauptkomponentenanalyse (probabilistische PCA)

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 probabilistische Hauptkomponentenanalyse (probabilistische PCA) ist ein Wahrscheinlichkeitsmodell für die Merkmalsvektoren eines Datensatzes: Ein niedrigdimensionaler gaußscher Zufallsvektor, der Code, wird durch eine Matrix in den Merkmalsraum abgebildet, und isotropes gaußsches Rauschen wird addiert. Die Wahrscheinlichkeitsverteilung des Merkmalsvektors ist dann eine multivariate Normalverteilung, deren Kovarianzmatrix eine Matrix vom Rang d' plus ein Vielfaches der Einheits-Matrix ist, sodass das Modell jedem Punkt des Merkmalsraums einen Wert der Wahrscheinlichkeitsdichte zuweist. Dieser Wert ist das, was die Hauptkomponentenanalyse (PCA) nicht liefert: Ein Datenpunkt außerhalb des Trainingsdatensatzs kann danach gereiht werden, wie wahrscheinlich er ist, und fehlende Merkmale in einem Merkmalsvektor werden durch einen Erwartungswert ergänzt. Die maximale Likelihood-Schätzung der Modellparameter besitzt eine geschlossene Form in den Eigenwerten und Eigenvektorenen der Stichproben-Kovarianzmatrix: Die Spalten der Matrix spannen den Unterraum auf, den die PCA beibehält, und die Rausch-Varianz ist der Durchschnitt der verworfenen Eigenwerte. Der EM-Algorithmus (EM)-Algorithmus berechnet dieselbe Schätzung als Fixpunktiteration, ohne eine Eigenwertzerlegung. Der Posterior-Mittelwert des Codes zieht die PCA-Koordinaten zum Ursprung hin und geht in die orthogonale Projektion auf den Unterraum über, wenn die Rausch-Varianz gegen null strebt, wodurch sich wieder die PCA ergibt. Das isotrope Rauschen ist die Haupteinschränkung des Modells: Es führt dazu, dass der gelernte Unterraum von den Einheiten der Merkmale abhängt; die Faktorenanalyse weist stattdessen jeder Messgröße ihre eigene Rausch-Varianz zu.

Definition

Eine Wetterstation in Krems an der Donau (Österreich) zeichnet für jeden Tag des Jahres 2024 acht Zahlen auf: die tiefste, höchste und mittlere Lufttemperatur, den Niederschlag, die Sonnenscheindauer, die Luftfeuchtigkeit, den Luftdruck und die Windgeschwindigkeit. Die Hauptkomponentenanalyse (PCA) ersetzt die acht Zahlen durch zwei, wenig genug für ein Streudiagramm, und sagt nichts darüber aus, wie wahrscheinlich ein Tag ist: Der 14. September 2024, der

81.7 mm Regen brachte, wird wie jeder andere Tag in der Ebene platziert. Die probabilistische PCA behält den Unterraum bei, den die PCA liefert, und fügt eine Wahrscheinlichkeitsverteilung über die acht Zahlen hinzu, gegeben durch eine Wahrscheinlichkeitsdichte auf dem Merkmalsraum. Der Regentag trägt dann einen Dichtewert, der weit unter dem jedes anderen Tages des Jahres liegt.

Die probabilistische PCA ist ein Wahrscheinlichkeitsmodell mit latente Variable. Jeder Datenpunkt eines Datensatzes trägt einen Merkmalsvektor $\mathbf{x}^{(r)} \in \mathbb{R}^d$, für $r = 1, \ldots, m$, mit $d = 8$ und $m = 366$ für die Tage des Jahres 2024. Das Modell bildet ihn aus einem niedrigdimensionalen Vektor $\mathbf{z}^{(r)} \in \mathbb{R}^{d'}$, dem Code, dessen Länge $d' < d$ ein Hyperparameter ist:

$$\mathbf{x}^{(r)} = \mathbf{W}\mathbf{z}^{(r)} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}^{(r)}, \text{ für } r = 1, \ldots, m. \tag{1}$$

Hierbei ist $\mathbf{W} \in \mathbb{R}^{d \times d'}$ eine Matrix und $\boldsymbol{\mu} \in \mathbb{R}^d$ ein Mittelwert-Vektor. Die Codes und die Rausch-Vektoren des Modells (1) sind Realisierungen unabhängiger Zufallsvektoren, $\mathbf{z}^{(r)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ und $\boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, mit der Einheits-Matrix $\mathbf{I}$ und einer Rausch-Varianz $\sigma^2 > 0$. Das Rauschen ist isotrop: Es hat in jeder Richtung dieselbe Varianz und keine Kovarianz zwischen Merkmalen. Die Matrix $\mathbf{W}$ bildet einen Code auf einen Merkmalsvektor ab, jene Rolle, die der Decoder in der PCA spielt, und die Modellparameter sind $\mathbf{W}$, $\boldsymbol{\mu}$ und σ^2.

Das Herausintegrieren des Codes aus dem Modell (1) hinterlässt eine multivariate Normalverteilung für den Merkmalsvektor selbst,

$$\mathbf{x}^{(r)} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}), \text{ wobei } \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}. \tag{2}$$

Die Kovarianzmatrix $\mathbf{C}$ der multivariate Normalverteilung (2) ist eine Matrix vom Rang höchstens d' plus ein Vielfaches der Einheits-Matrix. Das ist die einzige Einschränkung, welche die probabilistische PCA von einer unbeschränkten multivariate Normalverteilung trennt: Die Varianz entlang jeder der d' beibehaltenen Richtungen wird einzeln modelliert, während jede zu ihnen allen orthogonale Richtung dieselbe Varianz σ^2 erhält. Da das Modell (1) angibt, wie ein Datenpunkt zustande kommt, ist es ein generatives Modell: Das Ziehen eines Codes und das Addieren von Rauschen erzeugt einen neuen Tag. Abb. 1 zeigt die aus den Tagen in Krems mit zwei Merkmalen, der tiefsten und der höchsten Tagestemperatur, gelernte Wahrscheinlichkeitsverteilung (2) als Höhenlinie im Abstand einer Standardabweichung vom Mittelwert-Vektor. Für einen Code aus einer einzigen Zahl ist der Unterraum, den die PCA beibehält, eine Gerade, und die Wahrscheinlichkeitsverteilung füllt eine Ellipse um sie herum aus.

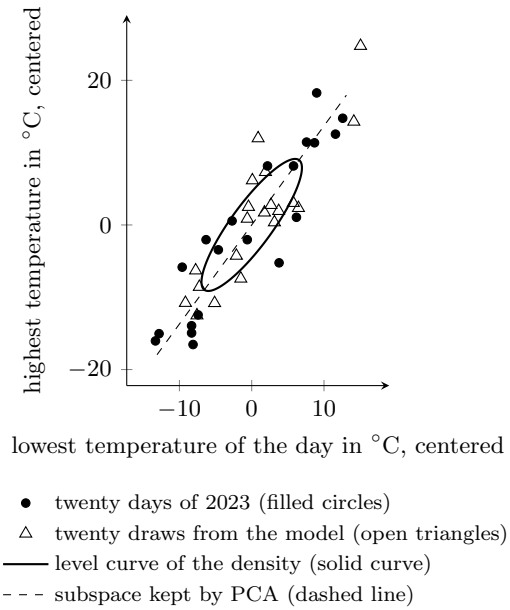

Abbildung 1: Zwanzig der 365 Tage des Jahres 2023 in Krems (jeweils jeder 19-te Tag), jeweils ein Datenpunkt mit der niedrigsten und der höchsten Temperatur des Tages als seinen beiden Merkmale, zentriert (gefüllte Kreise). Die durchgezogene Kurve ist die Ein-Standardabweichung-Niveaukurve der aus allen 365 Tagen gelernten Wahrscheinlichkeitsdichte (2) mit $d' = 1$, und die gestrichelte Linie markiert den von der PCA beibehaltenen Unterraum. Die offenen Dreiecke sind zwanzig aus dem gelernten Modell gezogene Merkmalsvektoren. Daten erzeugt mit `pythondemos/ppca.py`

Was die Wahrscheinlichkeitsverteilung (2) im Gegensatz zur PCA liefert, ist eine Zahl für jeden Punkt des Merkmalsraums. Ein Tag, der nicht zum Trainingsdatensatz gehörte, erhält einen Dichtewert, sodass Tage danach gereiht werden können, für wie wahrscheinlich das Modell sie hält. Mit den aus den 365 Tagen des Jahres 2023 gelernten Modellparametern und $d' = 3$ ist der unwahrscheinlichste Tag des Jahres 2024 in Krems der 14. September, an dem der Logarithmus der Wahrscheinlichkeitsdichte (2) bei -440.8 liegt, gegenüber -8.5 für den Median-Tag und -105.6 für den nächstniedrigeren (Abb. 2). Seine 81.7 mm Regen liegen 20.4 Stichproben-Standardabweichungen über dem Durchschnitt von 2023 von 1.5 mm. Das Reihen von Datenpunkte nach ihrem Dichtewert ist ein Zugangsweg zur Anomalieerkennung, und er bleibt der PCA verschlossen, da diese Koordinaten und keine Wahrscheinlichkeit zurückgibt.

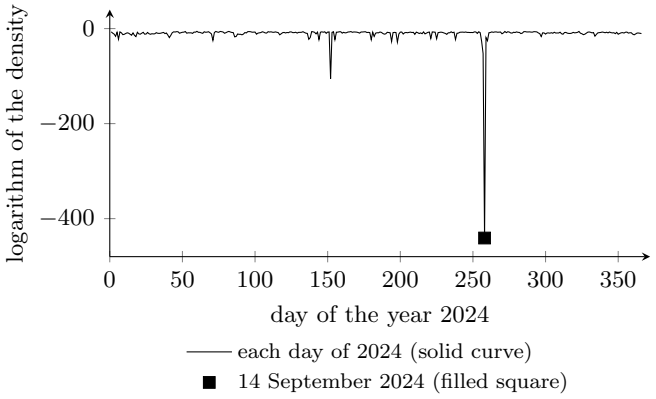

Abbildung 2: Der Logarithmus der Wahrscheinlichkeitsdichte (2) an jedem der 366 Tage des Jahres 2024 in Krems, unter den aus den 365 Tagen von 2023 mit $d' = 3$ und allen acht Messungen gelernten Modellparameter (durchgezogene Kurve). Das ausgefüllte Quadrat markiert den 14. September 2024, den Tag mit 81.7 mm Regen, bei -440.8, während der Median Tag des Jahres bei -8.5 liegt, wo die durchgezogene Kurve für den Großteil des Jahres verläuft. Daten erzeugt mit `pythondemos/ppca.py`

Die Modellparameter werden mittels maximale Likelihood gelernt, und die Schätzung besitzt eine geschlossene Form. Die Eigenwerte der Stichproben-Kovarianzmatrix $\widehat{\mathbf{C}}$ des Trainingsdatensatzs, geordnet als $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, gehen mit orthonormalen Eigenvektorenen $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(d)}$ einher. Die Schätzung des Mittelwert-Vektors ist der Stichprobenmittel $\widehat{\boldsymbol{\mu}}$, die Schätzung der Rausch-Varianz ist

$$\widehat{\sigma}^2 = \frac{1}{d - d'} \sum_{j=d'+1}^{d} \lambda_j, \tag{3}$$

und die Schätzung $\widehat{\mathbf{W}}$ hat die Spalten $\sqrt{\lambda_j - \widehat{\sigma}^2}\, \mathbf{u}^{(j)}$, für $j = 1, \ldots, d'$ (Bishop, 2006, Abschn. 12.2.1, Gl. (12.45) und (12.46); Tipping and Bishop, 1999, Abschn. 3.2, Gl. (7) und (8)). Der Durchschnitt (3) der verworfenen Eigenwerte ist die Varianz, welche die beibehaltenen Richtungen unerklärt lassen, gleichmäßig verteilt über die $d - d'$ Richtungen, die sie tragen. Die Spalten von $\widehat{\mathbf{W}}$ spannen den Unterraum auf, den die PCA beibehält, nämlich jenen, der von den ersten d' Eigenvektorenen von $\widehat{\mathbf{C}}$ aufgespannt wird. Ein Drehen dieser Spalten untereinander lässt die Kovarianzmatrix $\mathbf{C}$ unverändert, sodass maximale Likelihood den Unterraum und nicht die einzelnen Spalten bestimmt. Algorithmus 1 fasst die Berechnung zusammen.

Algorithm 1 Probabilistische Hauptkomponentenanalyse durch maximale Likelihood (Tipping and Bishop, 1999, Abschn. 3.2, Gl. (7) und (8); Bishop, 2006, Abschn. 12.2.1, Gl. (12.45) und (12.46))

Input: Merkmalsvektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$, code length $d' < d$

1: $\widehat{\boldsymbol{\mu}} :=$ the Stichprobenmittel of the Merkmalsvektoren
2: form the Stichproben-Kovarianzmatrix $\widehat{\mathbf{C}}$ of the Merkmalsvektoren centered by $\widehat{\boldsymbol{\mu}}$
3: compute the Eigenwertzerlegung of $\widehat{\mathbf{C}}$, giving Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_d$ and Eigenvektoren $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(d)}$
4: $\widehat{\sigma}^2 := \big(1/(d-d')\big) \sum_{j=d'+1}^{d} \lambda_j$
5: $\widehat{\mathbf{W}} :=$ the Matrix whose j-th column is $\sqrt{\lambda_j - \widehat{\sigma}^2}\, \mathbf{u}^{(j)}$, for $j = 1, \ldots, d'$

Output: the Modellparameter $\widehat{\boldsymbol{\mu}}$, $\widehat{\mathbf{W}}$, and $\widehat{\sigma}^2$ of the multivariate Normalverteilung (2); the code of a Merkmalsvektor is the Posterior-Mittelwert (4)

Der EM-Algorithmus (EM)-Algorithmus gelangt ohne eine Eigenwertzerlegung von $\widehat{\mathbf{C}}$ zu derselben Schätzung, was der günstigere Weg ist, wenn d groß ist und nur wenige Richtungen gesucht sind (Bishop, 2006, Abschn. 12.2.2). Fasst man die Modellparameter $\mathbf{W}$ und σ^2 in $\mathbf{w}$ zusammen, so ist eine Iteration eine Fixpunktiteration

$$\mathbf{w}^{(t+1)} = T\big(\mathbf{w}^{(t)}\big),$$

deren Operator T die Matrix $\mathbf{W}$ aus $\widehat{\mathbf{C}}$ und der A-posteriori-Verteilung der Codes neu schätzt und anschließend σ^2 auf die Varianz setzt, die das neue $\mathbf{W}$ unerklärt lässt (Tipping and Bishop, 1999, Anhang B, Gl. (29) und (30)). Keine Iteration verringert die Likelihood-Funktion, und die Fixpunkte von T sind die stationären Punkte dieser Likelihood-Funktion. Die Schätzung in geschlossener Form ist unter ihnen jene, an der die Likelihood-Funktion ihr Maximum annimmt; die stationären Punkte, die aus anderen Eigenvektorenen als den führenden d' gebildet werden, sind Sattelpunkte (Tipping and Bishop, 1999, Anhang A.3).

Im Vokabular des empirische Risikominimierung (ERM) ändert die probabilistische Hauptkomponentenanalyse zwei der drei Teile der PCA und belässt den dritten unverändert. Der Hypothesenraum erhält eine Zahl hinzu. PCA sucht über die Produkte eines Encoder und eines Decoder, das sind die Matrizen von Rang höchstens d'. Die probabilistische Hauptkomponentenanalyse sucht über die multivariate Normalverteilungen (2), deren Kovarianzmatrizen eine solche Matrix plus $\sigma^2\mathbf{I}$ sind. Die Spalten von $\mathbf{W}$ müssen nicht orthonormal sein: Die von $\widehat{\mathbf{W}}$ haben die quadrierten Längen $\lambda_j - \widehat{\sigma}^2$ für $j = 1, \ldots, d'$, sie legen also die Varianz entlang jeder beibehaltenen Richtung fest und nicht nur den Unterraum. Die Verlustfunktion ist der negative Logarithmus der Wahrscheinlichkeitsdichte der Wahrscheinlichkeitsverteilung $\mathbb{P}$ (2), anstelle der quadratischer Verlust der PCA,

$$L\left(\mathbf{x}, \mathbb{P}\right) = -\ln \mathcal{N}\left(\mathbf{x}\,;\, \boldsymbol{\mu}, \mathbf{C}\right),$$

sodass das Minimieren des empirisches Risiko maximale Likelihood ist. Der dritte Teil bleibt unverändert: Die Eigenwertzerlegung von $\widehat{\mathbf{C}}$, die die PCA löst, liefert auch hier dieses Minimum in geschlossener Form, und der EM Algorithmus ist der Weg, der diese umgeht. Die beiden Verlustfunktionen treffen sich, wenn die Rausch-Varianz gegen null geht: $2\sigma^2$ mal der empirisches Risiko nähert sich dem Minimum des Rekonstruktionsfehlers $\sum_{j=d'+1}^{d} \lambda_j$ der PCA (Tipping and Bishop, 1999, Anhang A.2, Gl. (17)). Diese Summe beträgt 1.636 für die acht standardisierten Messungen der 365 Tage des Jahres 2023 mit $d' = 3$. Multipliziert mit $2\sigma^2$ erreicht der empirisches Risiko 0.979 davon bei $\sigma^2 = 10^{-2}$, 0.998 bei 10^{-4} und 1.000 bei 10^{-6}.

Ein Datenpunkt wird durch die A-posteriori-Verteilung seines Codes im Unterraum platziert. Der Posterior-Mittelwert des Codes ist

$$\mathbb{E}\{\mathbf{z} \mid \mathbf{x}\} = \big(\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}} + \widehat{\sigma}^2 \mathbf{I}\big)^{-1} \widehat{\mathbf{W}}^\top \big(\mathbf{x} - \widehat{\boldsymbol{\mu}}\big), \tag{4}$$

dessen j-te Komponente für die geschlossene Lösung die j-te PCA-Koordinate $\langle \mathbf{u}^{(j)}, \mathbf{x} - \widehat{\boldsymbol{\mu}} \rangle$ multipliziert mit $\sqrt{\lambda_j - \widehat{\sigma}^2}/\lambda_j$ ist, für $j = 1, \ldots, d'$. Dieser Faktor liegt für die drei beibehaltenen Richtungen in Krems unter eins, nämlich 0.50, 0.71 und 0.78. Wenn $\widehat{\sigma}^2$ gegen null geht, nähert sich der Faktor $1/\sqrt{\lambda_j}$, hier 0.53, 0.80 und 0.91. Der Grenzwert des Posterior-Mittelwert (4) ist daher die PCA-Koordinate geteilt durch $\sqrt{\lambda_j}$, was beide Quellen als orthogonale Projektion in den Raum des Codes bezeichnen (Bishop, 2006, Abschn. 12.2.1, Gl. (12.50); Tipping and Bishop, 1999, Abschn. 3.4, Gl. (9)). Gegenüber diesem Grenzwert wird der Code bei $\widehat{\sigma}^2 > 0$ durch den Faktor $\sqrt{1 - \widehat{\sigma}^2/\lambda_j}$ zum Ursprung gezogen, hier 0.95, 0.89 und 0.85.

Im Merkmalsraum entspricht der PCA die Rekonstruktion $\widehat{\mathbf{W}}\mathbb{E}\{\mathbf{z} \mid \mathbf{x}\} + \widehat{\boldsymbol{\mu}}$: Bei $\widehat{\sigma}^2 = 10^{-12}$ stimmt sie bis auf $3 \cdot 10^{-12}$ mit der orthogonale Projektion der Merkmalsvektor auf den durch die ersten d' Eigenvektoren aufgespannten Unterraum überein. Bei $\widehat{\sigma}^2 = 0.33$ ist sie keine solche orthogonale Projektion mehr, weicht im quadratischen Mittel um 0.153 pro Eintrag ab, und die Rekonstruktion mit dem kleinsten quadratischen Fehler wendet stattdessen $\widehat{\mathbf{W}}\big(\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}}\big)^{-1}\big(\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}} + \widehat{\sigma}^2 \mathbf{I}\big)$ auf den Posterior-Mittelwert an (Tipping and Bishop, 1999, Abschn. 3.4). In diesem Grenzfall ist die Kovarianzmatrix $\mathbf{C}$ singulär und die Wahrscheinlichkeitsdichte existiert nicht mehr, sodass der Grenzwert, der die PCA zurückgewinnt, derjenige ist, der auf die Wahrscheinlichkeitsverteilung verzichtet.

Eine Wahrscheinlichkeitsverteilung über den Merkmalsraum behandelt auch fehlende Daten. Konditioniert auf die Einträge einer Merkmalsvektor, die aufgezeichnet wurden, ist die Wahrscheinlichkeitsverteilung (2) wieder eine Gauß-Verteilung in den nicht aufgezeichneten Einträgen, und ihr Erwartungswert ist der ausgefüllte Wert. Löscht man die höchste Temperatur jedes der 366 Tage des Jahres 2024 und füllt sie auf diese Weise auf, aus den auf 2023 gelernten Modellparameter, ergibt sich eine mittlerer quadratischer Fehler (MSE), deren Quadratwurzel 2.28 Grad beträgt, gegenüber 9.21 Grad, wenn stattdessen der Mittelwert von 2023 eingesetzt wird. Die von der PCA angewandte Projektion benötigt jeden Eintrag der Merkmalsvektor und bietet keine solche Regel,

während das gleiche bedingte Lesen auch das Lernen der Modellparameter aus Merkmalsvektoren mit fehlenden Einträgen ermöglicht (Tipping and Bishop, 1999, Abschn. 4.1).

Das Modell macht zwei Annahmen, und die Tage in Krems zeigen, was jede kostet. Das Rauschen ist isotrop, daher erhält jede Richtung, die zum beibehaltenen Unterraum orthogonal ist, die gleiche Varianz $\widehat{\sigma}^2 = 0.33$ für die acht standardisierten Messungen mit $d' = 3$. Die mittlere Lufttemperatur eines Tages ist der Mittelwert aus seiner niedrigsten und höchsten Temperatur bis auf 0.05 Grad an jedem Tag des Jahres 2023. Die Stichproben-Standardabweichung der Tage entlang des letzten Eigenvektor von $\widehat{\mathbf{C}}$ beträgt daher 0.0027, während das Modell $\widehat{\sigma} = 0.57$ erwartet. Das isotrope Rauschen bewirkt auch, dass der gelernte Unterraum von den Einheiten der Messungen abhängt: Multipliziert man den Niederschlagswert mit zehn, dreht sich der Unterraum um 51 Grad, der größte Winkel zwischen dem Unterraum vor und nach der Umskalierung. Deshalb werden die acht Messungen standardisiert, bevor die Modellparameter gelernt werden. Die Faktorenanalyse unterscheidet sich von der probabilistischen Hauptkomponentenanalyse nur dadurch, dass das Rauschen eine diagonale Kovarianzmatrix erhält, also eine Varianz pro Messung, und eine Umskalierung einer Messung lässt sie unverändert (Bishop, 2006, Abschn. 12.2.4; Tipping and Bishop, 1999, Abschn. 2.1 und 3.3). Mit aus denselben 365 Tagen von 2023 gelernten Modellparameter erhöht sie den Median des Logarithmus der Wahrscheinlichkeitsdichte über die 366 Tage von 2024 von -8.5 auf -3.2, und was sie dafür aufgibt, ist die geschlossene Form (Tipping and Bishop, 1999, Abschn. 2.1).

Die zweite Annahme ist die gaußsche Form der Wahrscheinlichkeitsdichte (2), von der die Rangfolge der Tage nach Dichtewert abhängt. Wären die Tage von 2024 Realisierungen von i.i.d. Zufallsvektoren mit der gelernten multivariate Normalverteilung, wäre die quadratische Form im Exponenten dieser Wahrscheinlichkeitsdichte eine Realisierung einer Chi-Quadrat-Zufallsvariable (ZV) mit acht Freiheitsgraden. Ihr 99-Perzentil liegt bei 20.1, was im Mittel 3.7 der 366 Tage überschreiten würden. Tatsächlich tun dies sechsundzwanzig, ebenso wie 16 der 365 Tage von 2023, aus denen die Modellparameter gelernt wurden. Am 14. September 2024 beträgt die quadratische Form 870.6, was dieselbe Chi-Quadrat-ZV mit Wahrscheinlichkeit $1.2 \cdot 10^{-182}$ überschreitet, und dieser Tag trat ein. Der Dichtewert ordnet also die Tage, sagt aber nicht, wie oft ein Tag dieser Art auftritt. Wo eine multivariate Normalverteilung zu einfach ist, ist die Standarderweiterung ein Gemisch probabilistischer Hauptkomponentenanalyse-Modelle, ein Gaußsches Mischmodell (GMM), dessen Komponenten-Kovarianzmatrizen jeweils die Form $\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$ haben (Tipping and Bishop, 1999, Abschn. 4.2). Was die beiden Annahmen im Gegenzug liefern, ist eine Wahrscheinlichkeitsverteilung über den gesamten Merkmalsraum in geschlossener Form: Sie gibt einem Tag außerhalb des Trainingsdatensatz einen Dichtewert und einem nicht aufgezeichneten Merkmal einen ausgefüllten Wert.

Siehe auch: PCA, robuste Hauptkomponentenanalyse (robuste PCA), Sparse PCA, Dimensionsreduktion, ERM, Wahrscheinlichkeitsmodell, multivariate

Normalverteilung, maximale Likelihood, EM, latente Variable, generatives Modell, Stichproben-Kovarianzmatrix, fehlende Daten, Anomalieerkennung, GMM.

Querverweise

Literatur

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Tipping and Bishop (1999). Probabilistic Principal Component Analysis. J. Roy. Statist. Soc.: Ser. B (Statist. Methodology), 61(3), pp. 611–622.