multivariate Normalverteilung

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 multivariate Normalverteilung ist die Wahrscheinlichkeitsverteilung eines Gaußschen Zufallsvektors in $\mathbb{R}^d$, bestimmt durch einen Mittelwert Vektor und eine Kovarianzmatrix. Ihre Wahrscheinlichkeitsdichte ist proportional zum Exponential einer negativen quadratischen Form im Inversen der Kovarianzmatrix, sodass ihre Niveaumengen Ellipsoide sind, die am Mittelwert Vektor zentriert sind, mit Achsen entlang der Eigenvektoren der Kovarianzmatrix und Halbachsen proportional zu den Quadratwurzeln ihrer Eigenwerte. Die Familie ist abgeschlossen unter affinen Abbildungen und unter Konditionierung, und unter allen Zufallsvektoren mit einer gegebenen Kovarianzmatrix hat der Gaußsche die größte differential entropy. Die maximum likelihood-Schätzungen der beiden Parameter aus einem Datensatz sind der Stichprobenmittel und die sample covariance matrix, in geschlossener Form. Wenn die Kovarianzmatrix singulär ist, existiert keine Wahrscheinlichkeitsdichte und der Zufallsvektor ist auf einen affinen Unterraum beschränkt. Im machine learning (ML) ist die multivariate Normalverteilung die Komponente eines Gaußsches Mischmodell (GMM), das Wahrscheinlichkeitsmodell für das Rauschen hinter dem squared error loss und die klassenbedingte Wahrscheinlichkeitsverteilung hinter linear discriminant analysis (LDA).

Definition

Man betrachte die 365 Tage des Jahres 2024 in Krems an der Donau (Österreich), deren Aufzeichnung die Lufttemperatur sowohl um 06:00 (UTC) als auch um 15:00 (UTC) enthält, wobei jeder Tag ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$ ist, das die beiden Messwerte enthält. Ein warmer Morgen geht gewöhnlich mit einem warmen Nachmittag einher, sodass die Tage eine längliche Wolke um ihren Mittelwert Vektor bilden (Abb. 1(a)). Die multivariate Normalverteilung ist ein Wahrscheinlichkeitsmodell für solche Merkmalsvektoren, das dieses Bild mit zwei Parameter beschreibt (Anderson, 2003, Sect. 2.3; Bishop, 2006, Sect. 2.3). Ein Mittelwert Vektor $\boldsymbol{\mu} \in \mathbb{R}^d$ lokalisiert das Zentrum der Wolke, und eine Kovarianzmatrix $\mathbf{C} \in \mathbb{R}^{d \times d}$ legt ihre Ausbreitung in jede Richtung fest. Die multivariate Normalverteilung wird mit $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ bezeichnet, und ein Zufallsvektor mit dieser Wahrscheinlichkeitsverteilung ist ein Gaußscher Zufallsvektor.

Für eine invertible Kovarianzmatrix $\mathbf{C}$, die dann positive definite (pd) ist, hat die multivariate Normalverteilung die Wahrscheinlichkeitsdichte (Bishop, 2006, Sect. 2.3, Eq. (2.43))

$$\mathcal{N}(\mathbf{x}\,;\,\boldsymbol{\mu},\mathbf{C}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{C})}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) \quad \text{für } \mathbf{x}\in\mathbb{R}^d. \tag{1}$$

Der Exponent ist eine quadratische Form im Inversen $\mathbf{C}^{-1}$, der precision matrix. Die Wahrscheinlichkeitsdichte (1) ist am größten bei $\boldsymbol{\mu}$ und nimmt mit dem Wert dieser quadratischen Form ab; der Faktor vor dem Exponential, der die determinant von $\mathbf{C}$ enthält, sorgt dafür, dass sie zu eins integriert. Ein Zufallsvektor mit der Wahrscheinlichkeitsdichte (1) hat den Mittelwert Vektor $\boldsymbol{\mu}$ und die Kovarianzmatrix $\mathbf{C}$, wofür die beiden Symbole benannt sind (Bishop, 2006, Sect. 2.3, Eqs. (2.59)–(2.64)). Für $d = 1$ ist die Wahrscheinlichkeitsdichte (1) die Wahrscheinlichkeitsdichte eines Gaußsche Zufallsvariable mit Mittelwert μ und Varianz σ^2.

Eine Niveaumenge der Wahrscheinlichkeitsdichte (1) ist die Menge der Merkmalsvektoren, bei denen die quadratische Form einen gegebenen Wert c^2 annimmt, für eine positive Zahl c,

$$\big\{\mathbf{x}\in\mathbb{R}^d : (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\big\}. \tag{2}$$

Die Kovarianzmatrix ist eine symmetrische Matrix und erlaubt eine spectral decomposition $\mathbf{C} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$, mit orthonormalen Eigenvektoren als Spalten von $\mathbf{V}$ und den Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_d > 0$ auf der Diagonalen von $\boldsymbol{\Lambda}$. Die Niveaumenge (2) ist dann ein Ellipsoid, das bei $\boldsymbol{\mu}$ zentriert ist, dessen Achsen entlang der Eigenvektoren zeigen und dessen Halbachse entlang des j-ten Eigenvektor die Länge $c\sqrt{\lambda_j}$ hat (Bishop, 2006, Sect. 2.3, Eqs. (2.44)–(2.52), Fig. 2.7). Auf ihr ist die Wahrscheinlichkeitsdichte gleich $\exp(-c^2/2)$ mal ihrem Wert bei $\boldsymbol{\mu}$. Abb. 1(a) zeigt die Niveaumengen bei $c = 1$ und $c = 2$ für die Tage in Krems. Die Eigenwerte sind 146.5 und 6.9 Quadratgrad Celsius, sodass die Niveaumenge bei $c = 2$ Halbachsen von 24.2 und 5.2 Grad Celsius hat. Ihre Hauptachse zeigt entlang des Eigenvektors $(0.65, 0.76)^\top$, der Richtung, in der die beiden Messwerte gemeinsam ansteigen.

Die Wahrscheinlichkeit, dass ein Gaußscher Zufallsvektor innerhalb der Niveaumenge (2) liegt, hängt nur von c und d ab: Die quadratische Form eines Gaußschen Zufallsvektors hat eine Chi-Quadrat-Verteilung mit d Freiheitsgraden (Anderson, 2003, Thm. 3.3.3). Für $d = 2$ beträgt diese Wahrscheinlichkeit $1 - \exp(-c^2/2)$, das heißt 39.3 Prozent bei $c = 1$, 86.5 Prozent bei $c = 2$ und 98.9 Prozent bei $c = 3$. Abb. 1(b) vergleicht diese Werte mit dem Anteil der Tage in Krems innerhalb jeder Niveaumenge: 31.0, 87.7 und 99.7 Prozent, wobei der letzte Wert alle Tage bis auf einen umfasst. Die Tage sind entlang der Hauptachse gleichmäßiger verteilt, als es ein Gaußscher Zufallsvektor wäre, da die Jahreszeiten einen Tag entlang dieser Achse verschieben. Nur 60.8 Prozent von ihnen liegen innerhalb einer Standardabweichung des Mittelwerts entlang dieser Achse, gegenüber 68.3 Prozent für einen Gaußschen Zufallsvektor. Ein Gaußsches Mischmodell (GMM) mit zwei Komponenten, eine für die kalten und

eine für die warmen Tage, beschreibt eine solche Wolke besser als eine einzelne multivariate Normalverteilung.

Jeder Gaußsche Zufallsvektor ist eine affine Funktion eines standard normal random vector $\mathbf{z} \in \mathbb{R}^\ell$, dessen Einträge unabhängig und identisch verteilt (u.i.v.) Standard-Gaußsche Zufallsvariablen sind: $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ mit einer Matrix $\mathbf{A} \in \mathbb{R}^{d \times \ell}$, und dann $\mathbf{C} = \mathbf{A}\mathbf{A}^\top$ (Anderson, 2003, Sect. 2.4). Diese Konstruktion deckt auch eine singuläre Kovarianzmatrix ab, für die die Wahrscheinlichkeitsdichte (1) undefiniert ist. Der Zufallsvektor ist dann auf einen affinen Unterraum beschränkt, den column space von $\mathbf{A}$ verschoben um $\boldsymbol{\mu}$, der in $\mathbb{R}^d$ Volumen Null hat, sodass keine Wahrscheinlichkeitsdichte auf $\mathbb{R}^d$ existiert (Anderson, 2003, Sect. 2.4; Lapidoth, 2017, Ch. 23). Eine singuläre Kovarianzmatrix entsteht immer dann, wenn ein Merkmal eine affine Funktion der anderen ist. Das Anhängen des Anstiegs vom 06:00- zum 15:00-Messwert als drittes Merkmal eines Tages in Krems ergibt eine Kovarianzmatrix vom rank zwei, und die dreidimensionalen Merkmalsvektoren liegen auf einer Ebene.

Zwei Eigenschaften machen die multivariate Normalverteilung zum Standard-Wahrscheinlichkeitsmodell für numerische Merkmalsvektoren. Erstens, ist die Familie unter affinen Abbildungen abgeschlossen (Anderson, 2003, Thm. 2.4.5): für jede Matrix $\mathbf{B} \in \mathbb{R}^{q \times d}$ und jeden Vektor $\mathbf{c} \in \mathbb{R}^q$,

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \quad \text{impliziert} \quad \mathbf{B}\mathbf{x} + \mathbf{c} \sim \mathcal{N}\left(\mathbf{B}\boldsymbol{\mu} + \mathbf{c}, \mathbf{B}\mathbf{C}\mathbf{B}^\top\right). \tag{3}$$

Wählt man $\mathbf{B}$ so, dass es einige Einträge von $\mathbf{x}$ auswählt, zeigt sich, dass jede Randverteilung einer multivariaten Normalverteilung wieder eine multivariate Normalverteilung ist (Bishop, 2006, Sect. 2.3.2). Für die Tage in Krems ist der Anstieg vom ersten zum zweiten Messwert die affine Abbildung (3) mit $\mathbf{B} = (-1, 1)$ und $\mathbf{c} = 0$; unter dem Wahrscheinlichkeitsmodell hat er einen Mittelwert von 6.2 und eine Varianz von $\mathbf{B}\mathbf{C}\mathbf{B}^\top = 15.4$ Quadratgrad Celsius. Die Familie ist auch unter Konditionierung abgeschlossen. Gegeben die Werte einiger Einträge von $\mathbf{x}$, ist die conditional probability distribution der verbleibenden Einträge eine multivariate Normalverteilung. Ihr Mittelwert ist eine affine Funktion der gegebenen Werte, und ihre Kovarianzmatrix hängt nicht von diesen ab (Anderson, 2003, Thm. 2.5.1; Bishop, 2006, Sect. 2.3.1). Betrachten wir $d = 2$, und schreiben wir $C_{1,1}$, $C_{1,2}$ und $C_{2,2}$ für die Einträge der Kovarianzmatrix und μ_1 und μ_2 für die Einträge des Mittelwert Vektor. Die conditional probability distribution von x_2, gegeben x_1, ist dann

$$\mathcal{N}\left(\mu_2 + \frac{C_{1,2}}{C_{1,1}}(x_1 - \mu_1), C_{2,2} - \frac{C_{1,2}^2}{C_{1,1}}\right). \tag{4}$$

Der bedingte Mittelwert in der conditional probability distribution (4) ist die Vorhersage des 15:00-Messwerts aus dem 06:00-Messwert mit dem kleinsten Risiko unter dem squared error loss, und ist linear in x_1 (siehe lineare Regression). Für die Tage in Krems ist dies die durch least squares gelernte Linie, mit einer Steigung von 1.04, und die bedingte Standardabweichung beträgt 3.9 Grad Celsius. Zweitens, unter allen Zufallsvektoren mit der Kovarianzmatrix $\mathbf{C}$ hat

der Gaußsche Zufallsvektor die größte differential entropy (Cover and Thomas, 2006, Thm. 8.6.5). Wenn nur der Mittelwert Vektor und die Kovarianzmatrix bekannt sind, ist die multivariate Normalverteilung daher das am wenigsten restriktive Wahrscheinlichkeitsmodell.

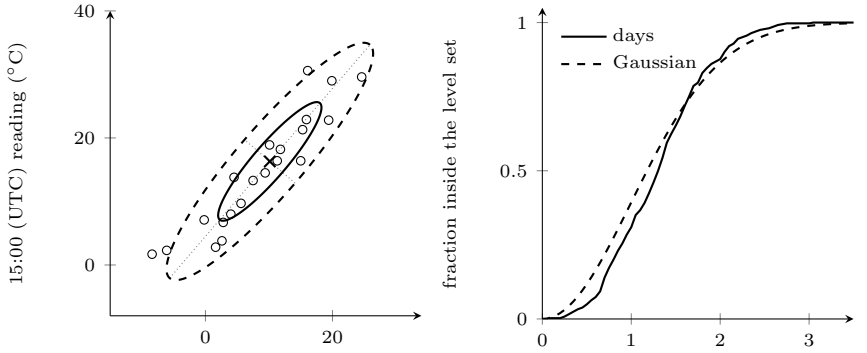
Die Parameter $\boldsymbol{\mu}$ und $\mathbf{C}$ werden aus einem Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ mittels maximum likelihood gelernt. Unter der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) sind die Datenpunkte Realisierungen von u.i.v. Zufallsvektoren $\mathbf{x}^{(r)} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$. Der Verlust, der durch einen Datenpunkt entsteht, ist der negative Logarithmus der Wahrscheinlichkeitsdichte (1) an seinem Merkmalsvektor, und der Durchschnitt dieses Verlust über $\mathcal{D}$ ist das empirical risk des Paares $(\boldsymbol{\mu}, \mathbf{C})$. Ihre Minimierung ist empirische Risikominimierung (ERM) über dem Hypothesenraum aller Wahrscheinlichkeitsdichten (1). Der Minimierer ist in geschlossener Form verfügbar, sodass kein iterativer Algorithmus erforderlich ist (Anderson, 2003, Thm. 3.2.1; Bishop, 2006, Sect. 2.3.4, Eqs. (2.121)–(2.122)):

$$\widehat{\boldsymbol{\mu}} = \frac{1}{m} \sum_{r=1}^{m} \mathbf{x}^{(r)} \quad \text{und} \quad \widehat{\mathbf{C}} = \frac{1}{m} \sum_{r=1}^{m} \left(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}\right) \left(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}\right)^{\top}. \tag{5}$$

Die Schätzungen (5) sind der Stichprobenmittel und die sample covariance matrix von $\mathcal{D}$ mit dem Teiler m; unter der u.i.v.-Annahme ist die sample covariance matrix im Durchschnitt um den Faktor $(m-1)/m$ kleiner als $\mathbf{C}$ (Bishop, 2006, Sect. 2.3.4, Eq. (2.124)). Für die Tage in Krems sind die Schätzungen (5) $\widehat{\boldsymbol{\mu}} = (10.1, 16.3)^{\top}$ und eine sample covariance matrix mit Varianzen 66.1 und 87.2 und Kovarianz 69.0, alle in Quadratgrad Celsius; der Korrelation-Koeffizient der beiden Messwerte beträgt 0.91. Die sample covariance matrix hat einen rank von höchstens $m-1$, sodass sie bei weniger als $d+1$ Datenpunkte singulär ist und die Wahrscheinlichkeitsdichte (1) mit diesen Schätzungen undefiniert ist. Ein übliches Heilmittel ersetzt $\widehat{\mathbf{C}}$ durch einen gewichteten Durchschnitt aus ihr und einem positiven Vielfachen der Einheits-Matrix (Hastie et al., 2009, Sect. 4.3.1, Eq. (4.14)). Dies schränkt den Hypothesenraum auf Kovarianzmatrizen ein, deren Eigenwerte von Null weg begrenzt sind, eine der Wege, die die Regularisierung auszeichnet.

Die multivariate Normalverteilung wird in der gesamten machine learning (ML) verwendet. Ein GMM zieht den Merkmalsvektor eines Datenpunkt aus einer von k multivariaten Normalverteilungen, sodass jeder Cluster eines soft clustering ein Mittelwert Vektor mit einer Kovarianzmatrix ist, dargestellt als Ellipse (Bishop, 2006, Sect. 9.2). Bei der regression wird das Label eines Datenpunkt als der Wert einer Hypothese plus einer Gaußschen Zufallsvariable (ZV) mit Mittelwert Null modelliert. Die maximum likelihood-Schätzung der Hypothese ist dann ERM mit dem squared error loss, sodass die lineare Regression mittels least squares dieses Wahrscheinlichkeitsmodell für das Label annimmt (Bishop, 2006, Sects. 1.2.5 and 3.1.1). Bei der Klassifikation führt die Modellierung des Merkmalsvektor jeder Klasse durch eine multivariate Normalverteilung und der Vergleich der Logarithmen der Wahrscheinlichkeitsdichten (1) zu einem Klassifikator mit dem kleinsten Risiko unter diesem Wahrscheinlichkeitsmodell.

Mit einer von allen Klassen geteilten Kovarianzmatrix heben sich die quadratischen Terme auf und die Entscheidungsgrenze ist eine Hyperebene, was linear discriminant analysis (LDA) ist. Mit einer Kovarianzmatrix pro Klasse ist die Entscheidungsgrenze eine quadratische Oberfläche (Bishop, 2006, Sect. 4.2.1; Hastie et al., 2009, Sect. 4.3, Eqs. (4.9) and (4.12)).



(a) 21 of the 365 days and two level sets

(b) fraction inside the level set at c

Abbildung 1: (a) Jeder 18 der 365 Tage des Jahres 2024 in Krems, 21 Tage (Kreise), der Mittelwert Vektor aller 365 Tage (Kreuz), und die Niveaumengen (2) der multivariaten Normalverteilung, die mittels maximum likelihood aus ihnen gelernt wurde: die innere bei $c = 1$ (durchgezogene Kurve) und die äußere bei $c = 2$ (gestrichelte Kurve). Die gepunkteten Segmente sind die beiden Achsen der äußeren Niveaumenge, entlang der Eigenvektoren der Kovarianzmatrix, mit Halbachsenlängen von 24.2 und 5.2 Grad Celsius. (b) Der Anteil der 365 Tage innerhalb der Niveaumenge bei c (durchgezogene Kurve) im Vergleich zur Wahrscheinlichkeit $1 - \exp(-c^2/2)$, dass ein Gaußscher Zufallsvektor mit denselben Parameter darin liegt (gestrichelte Kurve). Daten generiert von `pythondemos/mvndist.py`

Synonyme: multivariate Gauß-Verteilung, gemeinsame Normalverteilung, gemeinsam Gaußsche Verteilung.

Siehe auch: Gaußsche Zufallsvariable, Normalverteilung, Zufallsvektor, Mittelwert, Kovarianzmatrix, Wahrscheinlichkeitsdichte, precision matrix, spectral decomposition, Eigenwert, Eigenvektor, standard normal random vector, maximum likelihood, Stichprobenmittel, sample covariance matrix, u.i.v.-Annahme, ERM, GMM, lineare Regression, squared error loss, LDA, differential entropy, Wahrscheinlichkeitsmodell.

Querverweise

Literatur

1. Anderson (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., Wiley, Hoboken, NJ.

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

3. Lapidoth (2017). A Foundation in Digital Communication. 2nd ed., Cambridge Univ. Press, Cambridge, U.K.

4. Cover and Thomas (2006). Elements of Information Theory. 2nd ed., Wiley, Hoboken, NJ, USA.

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