lineare Diskriminanzanalyse (LDA)

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 lineare Diskriminanzanalyse (LDA) erlernt eine lineare Merkmalstransformation, die den Merkmalsvektor eines Datenpunkts durch ein einziges Merkmal ersetzt, das so gewählt wird, dass die beiden Label-Werte eines Problems der binäre Klassifikation so gut wie möglich unterschieden werden. Das Kriterium ist das Fisher-Kriterium, das Verhältnis der Streuung zwischen den Klassen zur Streuung innerhalb der Klassen des neuen Merkmals. Die Richtung, die sein Maximum annimmt, wird durch das Lösen eines einzigen linearen Gleichungssystems erhalten, sodass die LDA eine geschlossene Form und kein iteratives Verfahren ist. Die beiden StreuungsMatrizen summieren sich zur Stichproben-Kovarianzmatrix, jener Matrix, deren quadratische Form die Hauptkomponentenanalyse (PCA) maximiert, sodass die PCA die Richtung der größten gesamten Stichprobenvarianz beibehält, während die LDA die beiden Anteile gegeneinander abwägt. Eine Richtung kann fast die gesamte Stichprobenvarianz und fast nichts vom Anteil zwischen den Klassen tragen. Bei mehr als zwei Label-Werten liefert die LDA höchstens eine Richtung weniger, als es Werte gibt. Der Name bezeichnet auch den Klassifikator, der aus einer multivariate Normalverteilung pro Label-Wert mit einer gemeinsamen Kovarianzmatrix dieselbe Richtung erreicht.

Definition

Eine Wetterstation in Krems an der Donau (Österreich) zeichnet für jeden Tag das morgendliche Minimum und das Tagesmaximum der Temperatur auf, und in der Wettervorhersage wird eine einzige Zahl pro Tag gesucht, die die Tage mit Niederschlag von den trockenen unterscheidet. Die beiden Temperaturen bewegen sich gemeinsam: Ein kalter Tag ist morgens und nachmittags kalt, ein warmer Tag ist zu beiden Zeiten warm. Diese gemeinsame Schwankung ist die Richtung, entlang derer die Tage am stärksten variieren, und es ist die Richtung, die die Hauptkomponentenanalyse (PCA) beibehält. Für die gestellte Frage ist sie zudem nahezu nutzlos, weil sie besagt, wie warm ein Tag war, und nicht, ob es geregnet hat. Die LDA erzeugt ebenfalls eine Zahl pro Tag, und sie wählt die Richtung anhand der Labels.

Die LDA ist ein Verfahren des Repräsentationslernens für die binäre Klassifikation. Betrachtet sei ein Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \big\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\big\}$

des Stichprobengrößes m, mit Merkmalsvektoren $\mathbf{x}^{(r)} \in \mathbb{R}^d$ und Labelsn $y^{(r)} \in \{-1, +1\}$. Die LDA erlernt eine lineare Merkmalstransformation

$$\phi^{(\mathbf{w})} : \mathbb{R}^d \to \mathbb{R} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x},$$

die die d rohen Merkmale eines Datenpunkts durch das einzige neue Merkmal $\phi^{(\mathbf{w})}(\mathbf{x})$ ersetzt. Die Labels gehen über zwei Matrizen ein. Sei $\widehat{\boldsymbol{\mu}}^{(c)}$ der Stichprobenmittel der Merkmalsvektoren, die den Label-Wert c tragen, und p_c der Anteil des Trainingsdatensatzs, der ihn trägt. Der Stichprobenmittel aller Merkmalsvektoren ist dann $\widehat{\boldsymbol{\mu}} = \sum_{c \in \{-1,+1\}} p_c \widehat{\boldsymbol{\mu}}^{(c)}$. Die beiden Matrizen sind (Hastie et al., 2009, Abschn. 4.3.3):

$$\mathbf{S} := \frac{1}{m} \sum_{c \in \{-1,+1\}} \sum_{r \,:\, y^{(r)} = c} \left(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}^{(c)}\right)\left(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}^{(c)}\right)^\top, \tag{1}$$

$$\mathbf{B} := \sum_{c \in \{-1,+1\}} p_c \left(\widehat{\boldsymbol{\mu}}^{(c)} - \widehat{\boldsymbol{\mu}}\right)\left(\widehat{\boldsymbol{\mu}}^{(c)} - \widehat{\boldsymbol{\mu}}\right)^\top. \tag{2}$$

Die StreuungsMatrix innerhalb der Klassen $\mathbf{S}$ mittelt die Ausbreitung der Merkmalsvektoren um den Stichprobenmittel ihres eigenen Label-Werts, und die StreuungsMatrix zwischen den Klassen $\mathbf{B}$ misst, wie weit die beiden Stichprobenmittel auseinanderliegen. Entlang einer Richtung $\mathbf{w}$ hat das neue Merkmal die Stichprobenvarianz innerhalb der Klassen $\mathbf{w}^\top \mathbf{S} \mathbf{w}$ und die Stichprobenvarianz zwischen den Klassen $\mathbf{w}^\top \mathbf{B} \mathbf{w}$.

Das Kriterium, das die LDA maximiert, ist das Verhältnis der beiden, das Fisher-Kriterium (Bishop, 2006, Abschn. 4.1.4, Gl. (4.26); Fisher, 1936; Hastie et al., 2009, Abschn. 4.3.3, Gl. (4.15)),

$$\max_{\mathbf{w} \in \mathbb{R}^d \,:\, \|\mathbf{w}\| = 1} \frac{\mathbf{w}^\top \mathbf{B} \mathbf{w}}{\mathbf{w}^\top \mathbf{S} \mathbf{w}}. \tag{3}$$

Eine Richtung erzielt einen hohen Wert, wenn die beiden Stichprobenmittel auf ihr weit auseinanderliegen und die Tage eines Label-Werts auf ihr dicht beieinanderbleiben. Das Verhältnis der beiden Stichprobenvarianzen im Fisher-Kriterium (3) ändert sich nicht, wenn $\mathbf{w}$ umskaliert wird, sodass die Nebenbedingung nichts kostet und eine Richtung erlernt wird. Für ein invertierbares $\mathbf{S}$ ist die maximierende Richtung diejenige von

$$\widehat{\mathbf{w}} = \mathbf{S}^{-1}\left(\widehat{\boldsymbol{\mu}}^{(+1)} - \widehat{\boldsymbol{\mu}}^{(-1)}\right) \tag{4}$$

(Bishop, 2006, Abschn. 4.1.4, Gl. (4.30)). Die LDA-Richtung (4) wird in geschlossener Form durch das Lösen eines einzigen linearen Gleichungssystems berechnet: Die LDA ist kein iteratives Verfahren, sodass kein Operator ihre Ausgabe als Fixpunkt hat.

Der Name bezeichnet auch einen Klassifikator, nicht nur die Merkmalstransformation. Modelliert man jeden Label-Wert durch eine multivariate Normalverteilung mit eigenem Mittelwert und einer gemeinsamen Kovarianzmatrix und

schätzt diese Modellparameter mittels maximale Likelihood, so ergibt sich ein linearer Klassifikator, dessen Entscheidungsgrenze eine Hyperebene mit dem NormalenVektor (4) ist (Hastie et al., 2009, Abschn. 4.3). Die Richtung ist dieselbe, und das Wahrscheinlichkeitsmodell liefert den Schwellenwert, den das Fisher-Kriterium offenlässt.

Der Unterschied zur PCA besteht in einer einzigen Matrix. Die Streuungs-Matrix innerhalb der Klassen (1) und die StreuungsMatrix zwischen den Klassen (2) summieren sich zur Stichproben-Kovarianzmatrix (Hastie et al., 2009, Abschn. 4.3.3),

$$\mathbf{S} + \mathbf{B} = \widehat{\mathbf{C}}. \tag{5}$$

Die Labels sind aus der Summe (5) verschwunden. Die PCA maximiert $\mathbf{w}^\top \widehat{\mathbf{C}} \mathbf{w}$ über die Richtungen mit EinheitsNorm, was die gesamte Stichprobenvarianz des neuen Merkmals ist und nichts darüber aussagt, aus welchem Teil der Summe (5) sie stammt. Das Fisher-Kriterium (3) wägt stattdessen die beiden Teile gegeneinander ab. Eine Richtung kann daher fast die gesamte Stichprobenvarianz und fast nichts vom Anteil zwischen den Klassen tragen, was das Eingangsbeispiel beschreibt.

Dieses Beispiel ist messbar. Das Trainingsdatensatz umfasst die 1461 Tage von 2020 bis 2023 in Krems, jeweils mit dem morgendlichen Minimum und dem Tagesmaximum der Temperatur als ihren beiden Merkmalen und dem Label $+1$, wenn Niederschlag aufgezeichnet wurde. Der zum größten Eigenwert gehörende Eigenvektor von $\widehat{\mathbf{C}}$ trägt 93 Prozent der gesamten Stichprobenvarianz und erreicht ein Fisher-Kriterium von 0.00034, während die LDA-Richtung (4) einen Wert von 0.103 erreicht, was um den Faktor 303 größer ist. Die beiden Richtungen liegen 89 Grad auseinander, sodass das, wovon die Labels abhängen, fast genau entlang jener Richtung liegt, die die PCA an die letzte Stelle setzt (Abb. 1). Die LDA-Richtung gewichtet die beiden Merkmale mit entgegengesetzten Vorzeichen, $(0.817, -0.577)$, sodass ihr neues Merkmal an einem Tag groß ist, dessen Temperatur sich zwischen Morgen und Nachmittag wenig geändert hat, was durch eine Wolkendecke bewirkt wird.

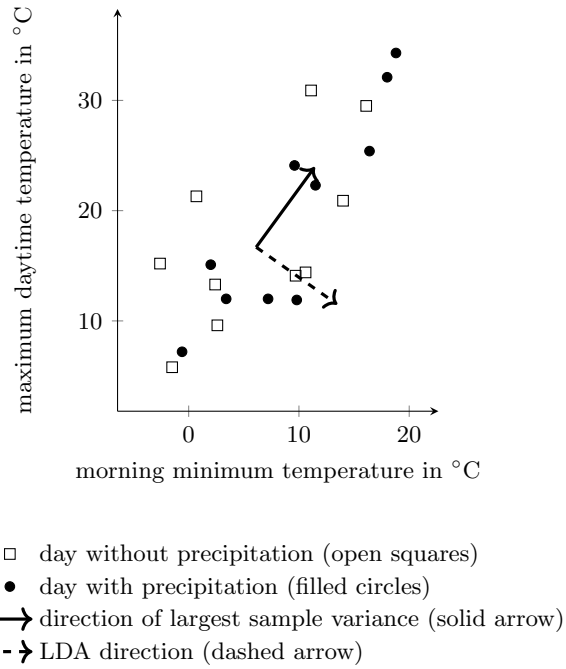


Abbildung 1: Zwanzig der 366 Tage des Jahres 2024 in Krems an der Donau (Österreich), zehn mit und zehn ohne aufgezeichneten Niederschlag, ausgewählt durch Entnahme jedes 14. der 144 nassen Tage und jedes 22. der 222 trockenen Tage. Jeder Tag ist ein Datenpunkt mit seinem morgendlichen Minimum und seinem Tagesmaximum der Temperatur als den beiden Merkmalen, gezeichnet als ausgefüllter Kreis, wenn Niederschlag aufgezeichnet wurde, und andernfalls als offenes Quadrat. Die beiden Pfeile beginnen beim Stichprobenmittel des Trainingsdatensatzs und sind gleich lang gezeichnet, da nur ihre Richtungen erlernt werden. Der durchgezogene Pfeil ist die Richtung der größten Stichprobenvarianz und der gestrichelte Pfeil die Richtung, entlang derer die beiden Arten von Tagen am besten unterschieden werden. Daten erzeugt mit `pythondemos/lda.py`

Das Kriterium (3) hat über die Richtungen der Ebene ein Maximum und ein Minimum, und Abb. 2 zeigt, wo die beiden Richtungen darauf liegen. Eine Schwellenwertbildung für das einzige LDA-Merkmal bei dem auf den 1461 Trainingstagen erlernten Wert klassifiziert 63.7 Prozent der 366 Tage des Jahres 2024, die als Testdatensatz zurückgehalten wurden, korrekt. Das einzige PCA-Merkmal erreicht 60.9 Prozent und die ständige Vorhersage des Mehrheits-Label-Werts 60.7 Prozent. Zwei Temperaturen eines Tages sagen wenig über Regen aus, und das Wenige, das sie aussagen, liegt in der Richtung, die die PCA verwirft.

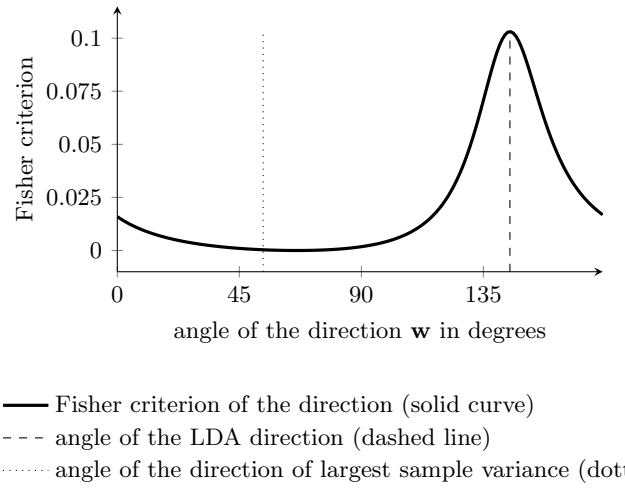


Abbildung 2: Das Fisher-Kriterium (3) jeder Richtung der Ebene für die beiden Temperatur-Merkmale der 1461 Tage von 2020 bis 2023 in Krems. Eine Richtung und ihre Gegenrichtung ergeben denselben Wert, sodass der Winkel über eine halbe Umdrehung läuft. Die gestrichelte Linie markiert den Winkel der LDA-Richtung, bei dem die durchgezogene Kurve ihr Maximum annimmt, und die gepunktete Linie den Winkel der Richtung der größten Stichprobenvarianz, der nahe dem Minimum liegt. Daten erzeugt mit `pythondemos/lda.py`

Mit der geschlossenen Form (4) gehen drei Einschränkungen einher. Sie benötigt ein invertierbares $\mathbf{S}$, was fehlschlägt, wenn das Trainingsdatensatz weniger Datenpunkte als Merkmale enthält. Das Schrumpfen von $\mathbf{S}$ in Richtung eines Vielfachen der EinheitsMatrix, was einer Regularisierung durch einen Strafterm entspricht, ist die übliche Abhilfe (Hastie et al., 2009, Abschn. 4.3.1, Gl. (4.14)). Bei mehr als zwei Label-Werten wird die StreuungsMatrix zwischen den Klassen aus allen Stichprobenmittel gebildet und hat einen Rang, der höchstens um eins kleiner ist als deren Anzahl, sodass die LDA höchstens so viele Richtungen liefert, wie groß d auch sein mag (Hastie et al., 2009, Abschn. 4.3.3). Und das einzige neue Merkmal ist nur so gut, wie eine Linearkombination sein kann: Zwei Labels, die durch eine gekrümmte Entscheidungsgrenze getrennt sind, überlappen sich auf jeder Richtung (Hastie et al., 2009, Abschn. 12.4).
Wo eine Linearkombination nicht ausreicht, besteht die Lösung darin, die LDA auf einem erweiterten Merkmalsvektor auszuführen, der auch die Quadrate und die Produkte der rohen Merkmale enthält. Eine Hyperebene im Merkmalsraum des erweiterten Merkmalsvektors ist eine gekrümmte Entscheidungsgrenze im Merkmalsraum der rohen Merkmale (Hastie et al., 2009, Abschn. 12.5), und eine Kernel-Methode ist die allgemeine Form dieses Schritts. Was die LDA für den Aufwand des Lösens eines einzigen linearen Gleichungssystems erkauft, ist der Faktor 303 im Fisher-Kriterium (3). Von den beiden einzelnen Merkmalen in Krems gewinnt das anhand der Labels gewählte 3.0 Prozentpunkte gegenüber

der Vorhersage des Mehrheits-Label-Werts, und das PCA-Merkmal gewinnt 0.3.

Siehe auch: PCA, Repräsentationslernen, Merkmalstransformation, Dimensionsreduktion, binäre Klassifikation, linearer Klassifikator, Stichproben-Kovarianzmatrix, Entscheidungsgrenze, Kernel-Methode, logistische Regression.

Querverweise

Literatur

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

2. Fisher (1936). The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugenics, 7(2), pp. 179–188.

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