Streudiagramm

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

Ein Streudiagramm stellt einen Datensatz mit zwei numerischen Merkmale durch eine Markierung pro Datenpunkt dar, die in der Ebene an der durch die beiden Merkmal-Werte gegebenen horizontalen und vertikalen Position platziert wird. Auf diese Weise lässt sich ein Datensatz mit zwei Merkmale vor und nach einer Merkmalstransformation visuell untersuchen, da Cluster, Ausreißer und ein Trend entlang einer Geraden im Bild und nicht in der Zahlentabelle sichtbar sind. Trägt die vertikale Achse das Label eines Datenpunkts, so ist eine Hypothese eine Kurve in derselben Ebene, sodass die Markierungen und die Abbildung, die ihre Höhe vorhersagt, gleichzeitig zu sehen sind. Ein Streudiagramm eines aufgeteilten Datensatzes stellt das Trainingsdatensatz durch ausgefüllte Kreise, das Validierungsdatensatz durch offene Dreiecke und das Testdatensatz durch offene Quadrate dar, sodass sich die drei Rollen ohne Farbe unterscheiden lassen. Nahezu zusammenfallende Markierungen verdecken einander, weshalb ein gut lesbares Streudiagramm höchstens etwa zwanzig nach einer angegebenen Regel ausgewählte Markierungen enthält, während jede daneben angegebene Zahl aus dem gesamten Datensatz berechnet wird. Ein Datensatz mit mehr als zwei Merkmale wird mithilfe einer Matrix von Streudiagrammen – mit einem Teilbild für jedes Paar von Merkmale – oder nach einer Dimensionsreduktion auf zwei Merkmale untersucht.

Definition

Betrachtet seien die 365 Tage des Jahres 2024 in Krems an der Donau (Österreich), für die die Lufttemperatur sowohl um 06:00 Uhr (UTC) als auch um 15:00 Uhr (UTC) aufgezeichnet vorliegt. Eine Tabelle mit 365 Zeilen aus je zwei Zahlen verbirgt, was eine Markierung pro Tag in der Ebene auf einen Blick zeigt – an der horizontalen Position des Messwerts von 06:00 Uhr und der vertikalen Position des Messwerts von 15:00 Uhr. Dieses Bild ist ein Streudiagramm: die grafische Darstellung eines Datensatzes $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$, dessen Datenpunkte zwei numerische Merkmale tragen, $\mathbf{x}^{(r)} = \big(x_1^{(r)}, x_2^{(r)}\big)^\top \in \mathbb{R}^2$, wobei die r-te Markierung für $r = 1, \ldots, m$ bei $\big(x_1^{(r)}, x_2^{(r)}\big)$ platziert wird. Nichts außer den beiden Zahlen eines Datenpunkts bestimmt die Position seiner Markierung, sodass das Bild dieselbe Information wie der Datensatz enthält. Ein einzelnes

Merkmal wird stattdessen durch ein Histogramm dargestellt, das die Werte beibehält und die Paarung verwirft.

In einer Anwendung des maschinelles Lernen (ML) ist eine der beiden Zahlen oft das Label und die andere das Merkmal. Die vertikale Achse trägt dann das Label y, und eine Hypothese h wird in derselben Ebene als Kurve durch die Punkte $\big(x, h(x)\big)$ eingezeichnet, sodass die Markierungen und die Abbildung, die ihre Höhe vorhersagt, gleichzeitig zu sehen sind (Bishop, 2006, Sect. 1.1). Ein Streudiagramm eines aufgeteilten Datensatzes gibt die Rolle jedes Datenpunkts durch die Form seiner Markierung wieder und kommt ohne Farbe aus: Das Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ wird durch ausgefüllte Kreise, das Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ durch offene Dreiecke und das Testdatensatz $\mathcal{D}^{(\text{test})}$ durch offene Quadrate dargestellt. Abb. 1(a) stellt die Tage in Krems auf diese Weise dar, zusammen mit der mittels der kleinste Quadrate auf den 219 Tagen des Trainingsdatensatzs gelernten Hypothese $\widehat{h}$, und Abb. 1(b) zeigt dieselben Tage nach einer Merkmalstransformation.

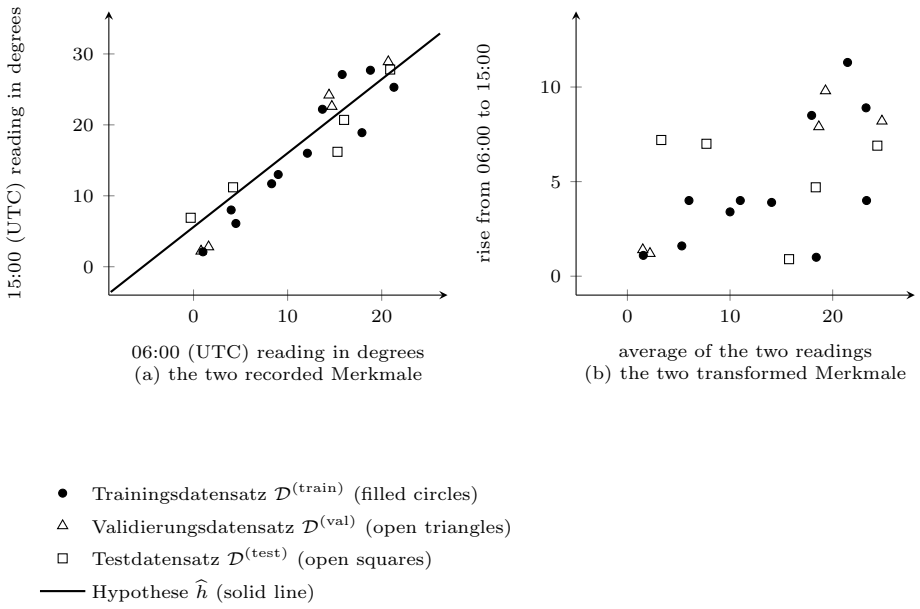

Abbildung 1: Einundzwanzig der 365 Tage des Jahres 2024 in Krems: jeder 20. der 219 Tage des Trainingsdatensatzs, jeder 15. der 73 Tage des Validierungsdatensatzs und jeder 15. der 73 Tage des Testdatensatzs. (a) Die Messwerte von 06:00 Uhr (UTC) und 15:00 Uhr (UTC) jedes Tages in Grad Celsius, mit dem Trainingsdatensatz als ausgefüllte Kreise, dem Validierungsdatensatz als offene Dreiecke, dem Testdatensatz als offene Quadrate und der mittels der kleinste Quadrate auf dem Trainingsdatensatz gelernten Hypothese $\widehat{h}$ als durchgezogene Linie. (b) Dieselben Tage und dieselben Markierungen nach der Merkmalstransformation auf den Mittelwert der beiden Messwerte und den Anstieg zwischen ihnen. Daten erzeugt durch `pythondemos/scatterplot.py`

Vier Eigenschaften eines Datensatzes sind im Bild und nicht in der Zahlentabelle sichtbar. Die Markierungen in Abb. 1(a) liegen entlang einer ansteigenden Geraden, was einem Stichprobenkoeffizienten der Korrelation von 0.91 über alle 365 Tage entspricht. Eine weit von den anderen entfernte Markierung ist ein Ausreißer: Über alle 365 Tage beträgt die größte Abweichung von der gelernten Geraden 11.0 Grad, und zwar am 4. März 2024, an dem auf einen Messwert von 0.7 Grad um 06:00 Uhr ein Messwert von 17.3 Grad um 15:00 Uhr folgte. Markierungen, die in getrennte Gruppen zerfallen, zeigen die Cluster eines Datensatzes noch bevor ein Verfahren zum Clustering ausgeführt wird, und ein leer gebliebener Bereich der Ebene zeigt, wo eine Hypothese ohne Datengrundlage eingesetzt wird.

Eine Merkmalstransformation verändert das Bild, und die beiden Bilder nebeneinander zeigen, was sie bewirkt hat. Abb. 1(b) ersetzt die beiden aufgezeichneten Messwerte durch ihren Mittelwert $(x_1+x_2)/2$ und den Anstieg x_2-x_1, eine

invertierbare lineare Abbildung. Aus der langgestreckten Punktwolke wird eine Wolke ohne erkennbare Gerade: Über alle 365 Tage fällt der Stichprobenkoeffizient der Korrelation von 0.91 auf 0.32, und allein auf den Mittelwert entfallen 82.6 Prozent der Summe der beiden Stichprobenvarianzen, gegenüber 43.1 Prozent für den Messwert von 06:00 Uhr. Eines von zwei Merkmale beizubehalten ist die kleinste Dimensionsreduktion, und anhand eines Streudiagramms lassen sich ihre Kosten visuell beurteilen.

Nahezu zusammenfallende Markierungen verdecken einander, weshalb ein Streudiagramm nur solange gut lesbar bleibt, wie es wenige davon enthält. Von den 365 Tagen in Krems liegen 180 innerhalb eines halben Grades zu einem anderen Tag und überlappen sich bei jeder lesbaren Markierungsgröße. Ein Streudiagramm stellt daher höchstens etwa zwanzig nach einer angegebenen Regel ausgewählte Datenpunkte dar, während jede daneben angegebene Zahl aus dem gesamten Datensatz berechnet wird. Die 21 Tage in Abb. 1 sind auf diese Weise ausgewählt, und nur zwei davon liegen innerhalb eines halben Grades zu einem anderen eingezeichneten Tag; ihr Stichprobenkoeffizient der Korrelation beträgt 0.95 und liegt damit nahe bei dem Wert von 0.91 für alle 365 Tage.

Für einen Datensatz mit mehr als zwei Merkmale gibt es kein einzelnes Streudiagramm. Eine Möglichkeit ist eine Matrix aus Streudiagrammen mit einem Teilbild für jedes Paar von Merkmale (Hastie et al., 2009, Sect. 3.2.1), eine andere ist eine Dimensionsreduktion auf zwei Merkmale, etwa mittels einer Hauptkomponentenanalyse (PCA), wonach das Bild zeigt, was nach der Reduktion erhalten geblieben ist. Keine der beiden Möglichkeiten ist ohne Nachteil. Eine Matrix aus Teilbildern wächst mit dem Quadrat der Anzahl der Merkmale und verlangt, die Teilbilder gedanklich miteinander abzugleichen. Eine Reduktion auf zwei Merkmale stellt die neuen Merkmale und nicht die aufgezeichneten dar. Was das Streudiagramm um diesen Preis liefert, ist das, was eine Zahlentabelle vorenthält: einen Datensatz, der sich auf einen Blick erfassen lässt. Das Darstellungsmittel hat jeden Wandel in der Speicherung von Daten überdauert; das erste bekannte Streudiagramm ist Herschels Darstellung des Positionswinkels gegen den Abstand von Doppelsternen aus dem Jahr 1833 (Friendly and Denis, 2005).

Siehe auch: Datenpunkt, Datensatz, Merkmal, Merkmalsvektor, Label, Hypothese, Trainingsdatensatz, Validierungsdatensatz, Testdatensatz, Ausreißer, Cluster, Merkmalstransformation, Dimensionsreduktion, PCA, Korrelation, Histogramm.

Querverweise

Literatur

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

2. Friendly and Denis (2005). The Early Origins and Development of the Scatterplot. Journal of the History of the Behavioral Sciences, 41(2),

pp. 103–130.

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.