zufällige Projektion

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

Eine zufällige Projektion ersetzt den Merkmalsvektor $\mathbf{x} \in \mathbb{R}^d$ eines Datenpunkts durch $\mathbf{A}\mathbf{x} \in \mathbb{R}^{d'}$, wobei $d' < d$ und $\mathbf{A}$ eine breite Matrix ist, deren Einträge Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) sind. In $\mathbf{A}$ fließen keine Daten ein, was eine zufällige Projektion von der Hauptkomponentenanalyse (PCA) und von einem Autoencoder unterscheidet, deren Abbildungen zur Dimensionsreduktion aus einem Trainingsdatensatz gelernt werden. Haben die Einträge den Mittelwert null und die Varianz $1/d'$, so gleicht der Erwartungswert der quadrierten Länge von $\mathbf{A}\mathbf{x}$ der quadrierten Länge von $\mathbf{x}$, sodass quadrierte euklidische Abstände zwischen Datenpunkten im Mittel erhalten bleiben. Das Johnson-Lindenstrauss-Lemma (JL-Lemma) macht daraus eine Garantie für einen gesamten endlichen Datensatz: Eine Zieldimension von mindestens $4 \ln m/(\varepsilon^2/2 - \varepsilon^3/3)$ hält jeden paarweisen quadrierten euklidischer Abstand innerhalb des Faktors $1 \pm \varepsilon$. Diese Bedingung hängt vom Stichprobengröße m und von der Toleranz ε ab, nicht aber von der Ausgangsdimension d, sodass die Einsparung mit der Länge des ursprünglichen Merkmalsvektors wächst. Die Einträge können normalverteilte Zufallsvariablen sein oder mit einer Wahrscheinlichkeit von jeweils ein halb die Werte $\pm 1/\sqrt{d'}$ annehmen. Eine zufällige Projektion kostet eine Multiplikation mit einer Matrix pro Datenpunkt und keinen Durchlauf über die Daten.

Definition

Ein Wetterarchiv in Krems an der Donau (Österreich) speichert für jeden Tag die Lufttemperatur um 15:00 Uhr (UTC) der 3650 aufgezeichneten Tage bis einschließlich dieses Tages, also von etwa zehn Jahren. Ein solcher Tag ist ein Datenpunkt, dessen Merkmalsvektor $\mathbf{x} \in \mathbb{R}^d$ genau $d = 3650$ Zahlen enthält, und eine Suche nach den Tagen, die einem gegebenen Tag am ähnlichsten sind, vergleicht diese Merkmalsvektoren anhand ihres euklidischer Abstands. Eine zufällige Projektion verkürzt jeden Datensatz vor der Suche: Sie ersetzt $\mathbf{x}$ durch $\mathbf{A}\mathbf{x} \in \mathbb{R}^{d'}$, wobei $d' < d$ und $\mathbf{A} \in \mathbb{R}^{d' \times d}$ eine breite Matrix ist, deren Einträge Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) sind. In $\mathbf{A}$ fließen keine Daten ein, und genau das unterscheidet eine zufällige Projektion von der Hauptkomponentenanalyse (PCA) und von einem Autoencoder, deren Abbildungen zur Dimensionsreduktion aus einem Trainings-

datensatz gelernt werden. Das Wort *Projektion* ist hier historisch bedingt: Die Matrix $\mathbf{A}$ bildet in den $\mathbb{R}^{d'}$ ab, ist also keine Projektion des $\mathbb{R}^d$ auf eine seiner Teilmengen.

Die Wahrscheinlichkeitsverteilung dieser Zufallsvariablen hat den Mittelwert null und die Varianz $1/d'$. Betrachtet wird ein festes $\mathbf{u} \in \mathbb{R}^d$. Jeder Eintrag von $\mathbf{Au}$ ist dann eine ZV mit dem Mittelwert null und der Varianz $\|\mathbf{u}\|_2^2/d'$. Die Addition der Varianzen der d' Einträge ergibt den Erwartungswert der quadrierten Länge von $\mathbf{Au}$,

$$\mathbb{E}\{\|\mathbf{Au}\|_2^2\} = \|\mathbf{u}\|_2^2. \tag{1}$$

Gelesen mit $\mathbf{u} = \mathbf{x}^{(r)} - \mathbf{x}^{(r')}$ besagt die Identität (1), dass der quadrierte euklidischer Abstand zwischen zwei Datenpunkten die Projektion im Mittel übersteht. Eine einzelne Ziehung von $\mathbf{A}$ verhält sich anders. Für zwei der Kremser Tage und $d' = 16$ lag der quadrierte euklidischer Abstand nach der Projektion über 4000 Ziehungen hinweg beim 0.19- bis 3.28-Fachen des ursprünglichen Werts.

Diese Streuung verengt sich mit wachsendem d', und das von Johnson and Lindenstrauss (1984) bewiesene Johnson-Lindenstrauss-Lemma (JL-Lemma) macht aus dieser Verengung eine Garantie, die jedes Paar eines endlichen Datensatzes gleichzeitig abdeckt. Betrachtet werden ein Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ vom Stichprobengröße m und eine Toleranz ε mit $0 < \varepsilon < 1$. Sobald die Zieldimension

$$d' \geq \frac{4 \ln m}{\varepsilon^2/2 - \varepsilon^3/3}, \tag{2}$$

erfüllt, gibt es eine Abbildung $\mathbf{\Phi}\colon \mathbb{R}^d \to \mathbb{R}^{d'}$ mit

$$(1-\varepsilon)\left\|\mathbf{x}^{(r)} - \mathbf{x}^{(r')}\right\|_2^2 \leq \left\|\mathbf{\Phi}\big(\mathbf{x}^{(r)}\big) - \mathbf{\Phi}\big(\mathbf{x}^{(r')}\big)\right\|_2^2 \leq (1+\varepsilon)\left\|\mathbf{x}^{(r)} - \mathbf{x}^{(r')}\right\|_2^2, \tag{3}$$

für alle $r, r' \in \{1, \ldots, m\}$, und eine zufällige Projektion liefert eine solche Abbildung mit positiver Wahrscheinlichkeit, die sich durch wiederholte Ziehungen beliebig nahe an eins bringen lässt (Dasgupta and Gupta, 2003, Thm. 2.1). In die Bedingung (2) gehen der Stichprobengröße m und die Toleranz ε ein, nicht aber die Ausgangsdimension d. Ihr Nenner ist bei $\varepsilon = 1$ am größten, wo er $1/6$ beträgt; somit wird keine Toleranz unter eins garantiert, bevor $d' \geq 24 \ln m$ erreicht ist, was für $m = 256$ Tage dem Wert 134 entspricht. Der Stichprobengröße geht über seinen Logarithmus ein: Bei $\varepsilon = 1/2$ erhöht ein hundertmal größerer Datensatz die Anforderung von 266 auf 487. Abb. 1 stellt die zweiseitige Schranke (3) dem gegenüber, was eine einzelne Ziehung von $\mathbf{A}$ mit den 32640 paarweisen quadrierten euklidische Abständen von 256 solchen Tagen macht.

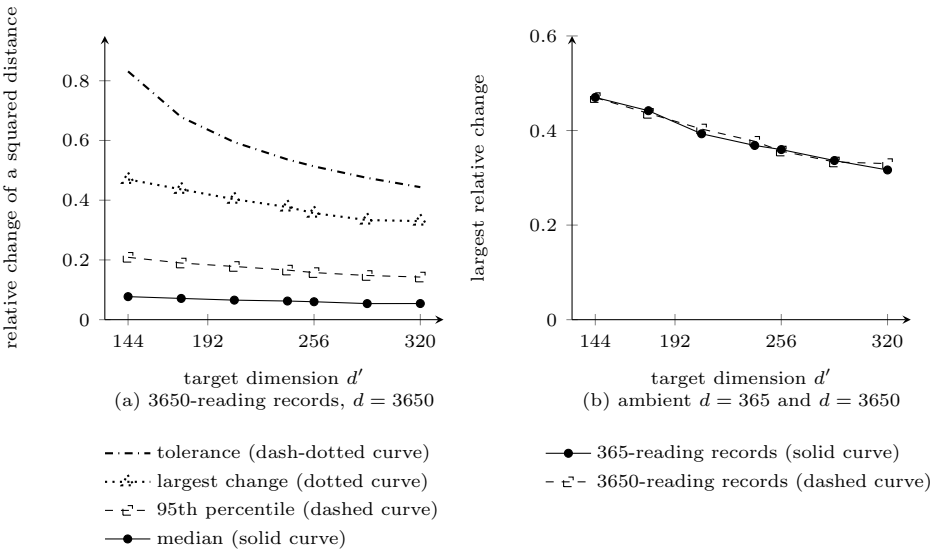


Abbildung 1: Die relative Änderung der 32640 paarweisen quadrierten euklidische Abstände von 256 Tagen in Krems unter einer zufälligen Projektion, aufgetragen gegen die Zieldimension d', wobei jeder Wert der Median über 40 Ziehungen von $\mathbf{A}$ ist. (a) Für die Aufzeichnungen mit 3650 Messwerten, $d = 3650$: der Median über die Paare (durchgezogene Kurve, ausgefüllte Kreise), ihr 95. Perzentil (gestrichelte Kurve, offene Quadrate), die größte Änderung (gepunktete Kurve, offene Dreiecke) und die Toleranz ε, welche die Bedingung (2) garantiert (strichpunktierte Kurve). (b) Die größte Änderung für die Aufzeichnungen mit 365 Messwerten, $d = 365$ (durchgezogene Kurve, ausgefüllte Kreise), und für jene mit 3650 Messwerten, $d = 3650$ (gestrichelte Kurve, offene Quadrate) – ein zehnfacher Unterschied in d, der die Kurve unverändert lässt. Daten erzeugt mit `pythondemos/randomprojection.py`

Zwei Wahrscheinlichkeitsverteilungen für die Einträge von $\mathbf{A}$ sind gebräuchlich. Normalverteilte Zufallsvariablen mit dem Mittelwert null und der Varianz $1/d'$ liefern die Garantie in der oben angegebenen Form (Shalev-Shwartz and Ben-David, 2014, Lemma 23.4). Einträge, die jeweils mit der Wahrscheinlichkeit ein halb gleich $1/\sqrt{d'}$ oder $-1/\sqrt{d'}$ sind, erfüllen dieselbe Bedingung (2) (Achlioptas, 2001; Dasgupta and Gupta, 2003, Sect. 3), und die Produkte in $\mathbf{Ax}$ reduzieren sich dann auf Additionen und Subtraktionen der Merkmale. In beiden Fällen beträgt der Aufwand dd' Multiplikationen pro Datenpunkt, die nur einmal anfallen: Es gibt keine Iteration und kein Training.

Für das Kremser Archiv und $d' = 256$ schrumpft die 3650 Messwerte umfassende Aufzeichnung eines Tages von 3650 Zahlen auf 256, also um den Faktor 14.3. Bei einer typischen Ziehung von $\mathbf{A}$ ändert sich keines der 32640 Paare quadrierter euklidische Abstände um mehr als 36 Prozent und die Hälfte von ihnen um weniger als 6 Prozent, verglichen mit den 51 Prozent, welche die Bedingung

(2) bei dieser Zieldimension garantiert. Ein Verfahren, das die Datenpunkte ausschließlich über ihre paarweisen euklidische Abstände einliest, wie etwa k-nächste Nachbarn (k-NN), arbeitet daher auf nahezu demselben Datensatz wie zuvor. Sind sowohl die Abbildung als auch die Rekonstruktion linear, ist die PCA die Wahl mit dem kleinsten mittleren quadratischen Fehler über dem Trainingsdatensatz, eine zufällige Projektion hingegen nicht. Darf der Rückschritt in den $\mathbb{R}^d$ nichtlinear sein, kann sich die Rangfolge umkehren: Merkmalsvektoren mit wenigen von null verschiedenen Merkmalen lassen sich aus einer zufälligen Projektion besser rekonstruieren als aus der PCA (Shalev-Shwartz and Ben-David, 2014, Sect. 23.5).

Siehe auch: Dimensionsreduktion, Repräsentationslernen, JL-Lemma, PCA, Autoencoder, Zufallsmatrix, breite Matrix, euklidischer Abstand, Merkmalsvektor, k-NN.

Querverweise

Literatur

1. Achlioptas (2001). Database-friendly random projections. In: PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 274–281. ACM.

2. Dasgupta and Gupta (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms, 22(1), pp. 60–65.

3. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

4. Johnson and Lindenstrauss (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary Math., 26, pp. 189–206.