Verteilungsverschiebung

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 Datenpunkt trägt Merkmale und ein Label, die zusammen aus einer gemeinsamen Wahrscheinlichkeitsverteilung gezogen werden. Verteilungsverschiebung ist eine Diskrepanz zwischen dieser gemeinsamen Wahrscheinlichkeitsverteilung während des Training und derjenigen, auf die die gelernte Hypothese nach dem Einsatz angewendet wird. Der durchschnittliche Verlust auf einem Testdatensatz schätzt das Risiko unter der Training-Wahrscheinlichkeitsverteilung ab, sodass ein kleiner Testfehler kein Beweis für den Verlust ist, der nach dem Einsatz entsteht. Eine gemeinsame Wahrscheinlichkeitsverteilung hat zwei Komponenten, eine Randverteilung und eine bedingte Wahrscheinlichkeitsverteilung, und die genannten Spezialfälle unterscheiden sich darin, welche von ihnen sich ändert: covariate shift ändert die Randverteilung der Merkmale allein, Label shift die Randverteilung der Labels allein, und concept drift die bedingte Wahrscheinlichkeitsverteilung. Sample weighting korrigiert covariate shift, wenn beide Wahrscheinlichkeitsverteilungen bekannt sind; Online-Lernen kommt zur Anwendung, wenn sie es nicht sind.

Definition

Die Wetterstation Krems an der Donau (Österreich) zeichnet die Lufttemperatur um 06:00 und um 15:00 (UTC) jedes Tages auf. Jeder Wintertag ist ein Datenpunkt: sein Merkmal ist der 06:00-Wert dieses Tages, sein Label der 15:00-Wert desselben Tages. Die Generierung jedes Datenpunkt kann durch Zufallsvariablen (ZV) mit einer gemeinsamen Wahrscheinlichkeitsverteilung $\mathbb{P}$ von Merkmal x und Label y angenähert werden.

Abb. 1 zeigt einige dieser Datenpunkte zusammen mit einer linearen Hypothese $\widehat{h}$. Diese Hypothese wird mittels lineare Regression aus 59 der 88 Datenpunkte gelernt, die zwischen Dezember 2025 und Februar 2026 gesammelt wurden. Sie verursacht einen Trainingsfehler von 14.4 und einen Testfehler von 4.0 auf einem Testdatensatz, das aus den verbleibenden 29 Datenpunkte besteht.

Die gelernte Hypothese $\widehat{h}$ wird dann auf die Datenpunkte $\mathcal{D}'$ der 92 Tage von Juni bis August 2026 an derselben Station angewendet. Wie beim Trainingsdatensatz kann die Generierung dieser Datenpunkte ebenfalls annähernd als Zufallsvariablen mit einer gemeinsamen Wahrscheinlichkeitsverteilung $\mathbb{P}'$ modelliert werden. Der resultierende durchschnittliche Verlust auf $\mathcal{D}'$ beträgt 147.0,

was 37-mal größer ist als der Testfehler 4.0 von $\widehat{h}$ (Abb. 1). Der Grund für diese Diskrepanz zwischen Testfehler und der Leistung auf $\mathcal{D}'$ ist eine Verteilungsverschiebung.

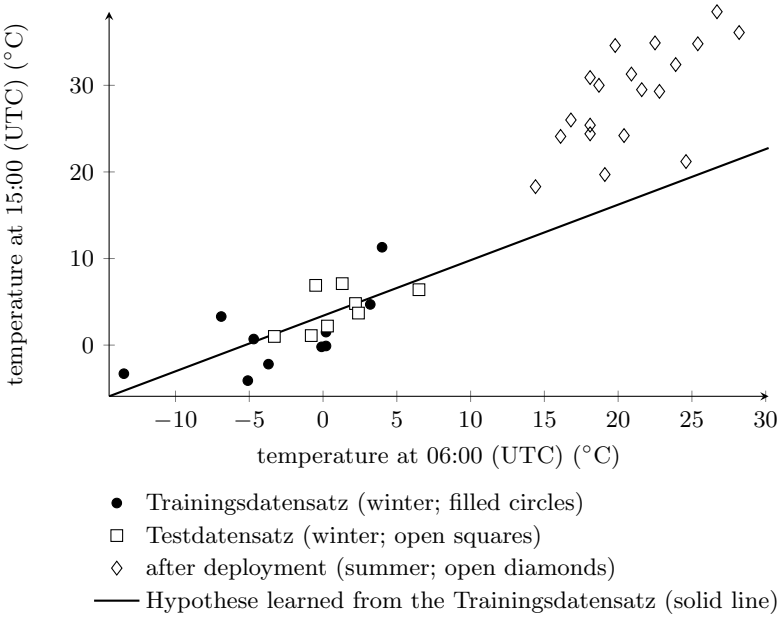


Abbildung 1: Verteilungsverschiebung zwischen den Winter-Datenpunkte, aus denen die Hypothese gelernt wurde, und den Sommer-Datenpunkte, auf die sie angewendet wird, beide in Krems an der Donau (Österreich). Jede Markierung ist ein Datenpunkt, dessen Merkmal die Lufttemperatur um 06:00 und dessen Label die Lufttemperatur um 15:00 (UTC) desselben Tages ist. Die Markierungen zeigen jeden sechsten Datenpunkt des Trainingsdatensatz (gefüllte Kreise), jeden vierten des Testdatensatz (offene Quadrate) und jeden fünften nach dem Einsatz (offene Rauten) an. Die Linie stellt die Hypothese $\widehat{h}$ dar, die mittels lineare Regression aus dem Trainingsdatensatz gelernt wurde. Die Merkmale der nach dem Einsatz angetroffenen Datenpunkte liegen außerhalb des vom Trainingsdatensatz abgedeckten Bereichs. Darüber hinaus unterschätzt die gelernte Hypothese $\widehat{h}$ jedes Label nach dem Einsatz, im Durchschnitt um 11.4 °C. Daten generiert von `pythondemos/distshift.py`

Verteilungsverschiebung bezieht sich auf eine Diskrepanz zwischen der Wahrscheinlichkeitsverteilung $\mathbb{P}$, die der Generierung der für das Training verwendeten Datenpunkte zugrunde liegt, und der Wahrscheinlichkeitsverteilung $\mathbb{P}'$ der Datenpunkte, auf die die gelernte Hypothese nach dem Einsatz angewendet wird (Quiñonero-Candela et al., 2009). Eine Verteilungsverschiebung kann nicht anhand des Testfehlers erkannt werden, wenn das Testdatensatz ebenfalls durch $\mathbb{P}$

generiert wird. Der durchschnittliche Verlust von $\widehat{h}$ auf einem solchen Testdatensatz ist ein unbiased estimator des Risiko $\mathbb{E}_{\mathbb{P}}\{ L\left(\mathbf{z}, \widehat{h}\right) \}$ unter $\mathbb{P}$, der erheblich vom Risiko $\mathbb{E}_{\mathbb{P}'}\{ L\left(\mathbf{z}, \widehat{h}\right) \}$ unter $\mathbb{P}'$ abweichen kann.

Das Risiko unter $\mathbb{P}'$ ist nach dem Einsatz relevant, aber das Testdatensatz wird aus $\mathbb{P}$ generiert. Hier ist ein kleiner Testfehler daher kein Beweis für den Verlust, der unter $\mathbb{P}' \neq \mathbb{P}$ entsteht. Man betrachte Datenpunkte mit einem einzelnen Merkmal x und einer Hypothese, die durch Vergleich von x mit einem Schwellenwert vorhersagt (dies ist ein einfacher Entscheidungsbaum). Die beiden Wahrscheinlichkeitsdichten in Abb. 2, die einer Standard-Normalverteilung $\mathbb{P}$ und die einer Normalverteilung $\mathbb{P}'$ mit einem größeren Mittelwert und einer größeren Standardabweichung, zeigen den Effekt des Ersetzens von $\mathbb{P}$ durch $\mathbb{P}'$ auf einem Schwellenwert, der für $\mathbb{P}$ optimiert ist. Der Anteil der Datenpunkte jenseits des Schwellenwerts steigt von 14% auf 88%, sodass das Risiko unter $\mathbb{P}'$ das Risiko unter $\mathbb{P}$ übersteigt, welches durch den Testfehler geschätzt wird.

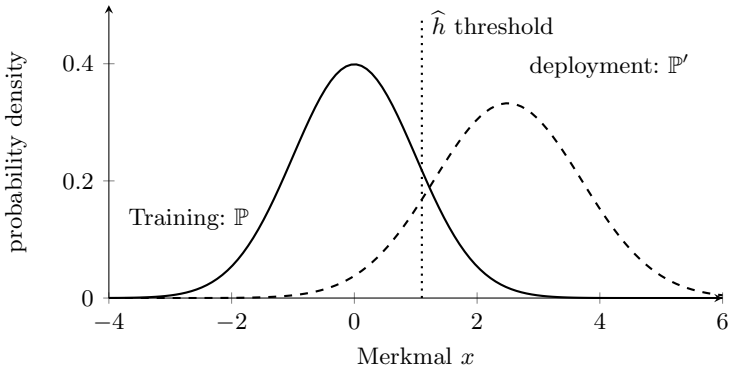


Abbildung 2: Verteilungsverschiebung für ein einzelnes Merkmal. Die durchgezogene Kurve ist die Wahrscheinlichkeitsdichte der Wahrscheinlichkeitsverteilung $\mathbb{P}$, die das Trainingsdatensatz generiert hat, hier eine Standard-Normalverteilung mit Mittelwert 0 und Standardabweichung 1; die gestrichelte Kurve ist die Wahrscheinlichkeitsdichte der Wahrscheinlichkeitsverteilung $\mathbb{P}'$ beim Einsatz, eine Normalverteilung mit Mittelwert 2.5 und Standardabweichung 1.2. Der gepunktete Schwellenwert der gelernten Hypothese $\widehat{h}$ bei $x = 1.1$ teilt jede Wahrscheinlichkeitsdichte in zwei Teile; die Fläche rechts davon ist der Anteil der Datenpunkte, deren Merkmal den Schwellenwert überschreitet, 14% unter $\mathbb{P}$ gegenüber 88% unter $\mathbb{P}'$, beide berechnet aus der Normalverteilung.

Drei benannte Spezialfälle der Verteilungsverschiebung entsprechen unterschiedlichen Ursprüngen der Verschiebung von $\mathbb{P}$ zu $\mathbb{P}'$. Diese unterschiedlichen Ursprünge entsprechen den Faktoren in zwei alternativen Faktorisierungen der gemeinsamen Wahrscheinlichkeitsverteilung. Beide gelten für jede gemeinsame Wahrscheinlichkeitsverteilung eines numerischen Merkmalsvektor und Label

(Billingsley, 1986, Thm. 33.3; Kallenberg, 1997, Thms. 5.3 and 5.4):

$$\mathbb{P}^{(\mathbf{x},y)} = \mathbb{P}^{(\mathbf{x})} \cdot \mathbb{P}^{(y|\mathbf{x})} = \mathbb{P}^{(y)} \cdot \mathbb{P}^{(\mathbf{x}|y)}.$$

Covariate shift ändert die Randverteilung $\mathbb{P}^{(\mathbf{x})}$ des Merkmalsvektor (wobei $\mathbb{P}^{(y|\mathbf{x})}$ unverändert bleibt), und Label shift ändert die Randverteilung $\mathbb{P}^{(y)}$ des Label (wobei $\mathbb{P}^{(\mathbf{x}|y)}$ unverändert bleibt). Concept drift bezieht sich auf eine Änderung der conditional probability distribution $\mathbb{P}^{(y|\mathbf{x})}$ (wobei $\mathbb{P}^{(\mathbf{x})}$ unverändert bleibt) oder $\mathbb{P}^{(\mathbf{x}|y)}$ (wobei $\mathbb{P}^{(y)}$ unverändert bleibt).

Das obige Zwei-Jahreszeiten-Beispiel ist kein reiner Fall von einem der drei, da jeder Faktor beider Faktorisierungen sich ändert. Die Randverteilung $\mathbb{P}^{(\mathbf{x})}$ ändert sich, wie unter covariate shift, da die Sommer-Datenpunkte höhere Merkmal-Werte aufweisen als jeder Wintertag. Die conditional probability distribution $\mathbb{P}^{(y|\mathbf{x})}$ ändert sich ebenfalls, wie unter concept drift: eine lineare Hypothese, die aus den 92 Sommer-Datenpunkte gelernt wurde, hat eine Steigung von 1.07, gegenüber der Steigung von 0.64 der Hypothese, die aus den Winter-Datenpunkte gelernt wurde. Ein Datenpunkt, dessen Merkmal einen Grad höher ist, hat daher im Sommer ein Label, das 1.07 Grad höher ist, gegenüber 0.64 im Winter. Die Randverteilung $\mathbb{P}^{(y)}$ ändert sich ebenfalls, wie unter Label shift, und damit auch die klassenbedingte $\mathbb{P}^{(\mathbf{x}|y)}$, die zweite Lesart von concept drift. Insbesondere haben die beiden Datensätze keinen Label-Wert gemeinsam, da jedes Label im Winter höchstens 13.6 °C und jedes Label im Sommer mindestens 13.7 °C beträgt.

Verteilungsverschiebung kann erkannt werden, indem man zusammenfassende Statistiken der Merkmale zwischen dem Trainingsdatensatz und den nach dem Einsatz eintreffenden Datenpunkte vergleicht, oder indem man den entstandenen Verlust überwacht, sobald Labels verfügbar werden (Quiñonero-Candela et al., 2009).

Eine Minderung der Verteilungsverschiebung ist sample weighting, das verwendet werden kann, wenn die Wahrscheinlichkeitsverteilungen $\mathbb{P}$ und $\mathbb{P}'$ bekannt sind. Hier, wobei $p^{(\mathrm{src})}(\cdot)$ die Wahrscheinlichkeitsdichte der Quelldistribution $\mathbb{P}$ ist, die das Trainingsdatensatz generiert hat, und $p^{(\mathrm{tar})}(\cdot)$ die Wahrscheinlichkeitsdichte der Zieldistribution $\mathbb{P}'$ ist, die die Datenpunkte nach dem Einsatz generiert, wird jeder Datenpunkt des Trainingsdatensatz mit dem Verhältnis $p^{(\mathrm{tar})}(\mathbf{x}) / p^{(\mathrm{src})}(\mathbf{x})$ gewichtet. Diese Gewichtung korrigiert covariate shift unter zwei Bedingungen: $\mathbb{P}^{(y|\mathbf{x})}$ ist unverändert, und $\mathbb{P}'$ ist absolut continuous bezüglich $\mathbb{P}$ auf dem Merkmalsraum, sodass das Gewicht als Radon-Nikodym-Ableitung von $\mathbb{P}'$ bezüglich $\mathbb{P}$ existiert (Billingsley, 1986, Thm. 32.2). Wo beide Wahrscheinlichkeitsdichten existieren, ist diese Ableitung das obige Verhältnis, definiert überall dort, wo $p^{(\mathrm{src})}(\mathbf{x}) > 0$ (Quiñonero-Candela et al., 2009). Unter diesen beiden Bedingungen (Shimodaira, 2000),

$$\begin{aligned} p^{(\mathrm{tar})}(\mathbf{x}, y) &= p^{(\mathrm{src})}(y \mid \mathbf{x})\, p^{(\mathrm{tar})}(\mathbf{x}) \\ &= \frac{p^{(\mathrm{tar})}(\mathbf{x})}{p^{(\mathrm{src})}(\mathbf{x})}\, p^{(\mathrm{src})}(\mathbf{x}, y) \quad \text{wherever } p^{(\mathrm{src})}(\mathbf{x}) > 0. \end{aligned} \tag{1}$$

Somit hat der gewichtete Verlust unter $\mathbb{P}$ denselben Erwartungswert wie der

Verlust unter $\mathbb{P}'$, was die Reweighting-Identität (1) zeigt.

Ein weiterer Ansatz zur Minderung der Verteilungsverschiebung ist die Verwendung von Online-Lernen, das die Hypothese aktualisiert, sobald Datenpunkte aus der sich ändernden Wahrscheinlichkeitsverteilung eintreffen (Quiñonero-Candela et al., 2009). Es erfordert, dass ihre Labels nach dem Einsatz verfügbar werden, was bei sample weighting nicht der Fall ist, und im Gegenzug muss keine der Wahrscheinlichkeitsverteilungen bekannt sein.

Synonyme: Dataset-Shift, Domain-Shift.

Siehe auch: Generalisierung, Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme), Wahrscheinlichkeitsverteilung, Risiko, Trainingsdatensatz, Testdatensatz, covariate shift, Transfer-Learning, domain adaptation, sample weighting, Online-Lernen.

Querverweise

Literatur

1. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.

2. Kallenberg (1997). Foundations of Modern Probability. Springer-Verlag, New York, NY, USA.

3. Quiñonero-Candela et al. (2009). Dataset Shift in Machine Learning. MIT Press, Cambridge, MA.

4. Shimodaira (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Statist. Plann. Inference, 90(2), pp. 227–244.