Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme)

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 Annahme unabhängig und identisch verteilter Daten (u. i. v.-Annahme) ist das Wahrscheinlichkeitsmodell, das die Erzeugung einer Folge von Datenpunkte durch unabhängig und identisch verteilt (u.i.v.)e Zufallsvariablen (ZV) darstellt. Sie besteht aus zwei getrennten Anforderungen, gemeinsamer Unabhängigkeit und einer gemeinsamen Wahrscheinlichkeitsverteilung. Durch diese Annahme ist der durchschnittliche Verlust auf einem Trainingsdatensatz eine erwartungstreue Schätzung des Risikos, was die empirische Risikominimierung sowie die Verwendung eines Validierungsdatensatzes zur Schätzung des Verlusts außerhalb des Trainingsdatensatzes rechtfertigt. Jede Anforderung lässt sich überprüfen. Unabhängigkeit kann über ihre Konsequenz für die Vorhersagbarkeit einzelner Datenpunkte getestet werden. Eine gemeinsame Wahrscheinlichkeitsverteilung lässt sich anhand des größten Abstands zwischen den empirischen Verteilungsfunktionen zweier Blöcke der Folge testen. Die Annahme ist bei einer Verteilungsverschiebung verletzt, wenn die nach dem Training auftretenden Datenpunkte einer anderen Wahrscheinlichkeitsverteilung folgen.

Definition

Eine Wetterstation in Krems an der Donau (Österreich) zeichnete im gesamten Jahr 2024 täglich um 15:00 Uhr (UTC) die Lufttemperatur auf. Abb. 1 zeigt vier dieser Monate, jeweils ein Datensatz mit einem Wert pro Tag. Die Werte eines solchen Monats lassen sich als Realisierungen unabhängig und identisch verteilt (u.i.v.)er Zufallsvariablen (ZV) auffassen. Diese Interpretation ist die u. i. v.-Annahme, und die Abbildung zeigt Belege dagegen. Die vier Monate weisen unterschiedliche Durchschnittswerte auf. Innerhalb jedes Monats folgt auf einen warmen Tag häufig ein weiterer warmer Tag, sodass der abgelesene Tag darüber entscheidet, wie die benachbarten Werte ausfallen.

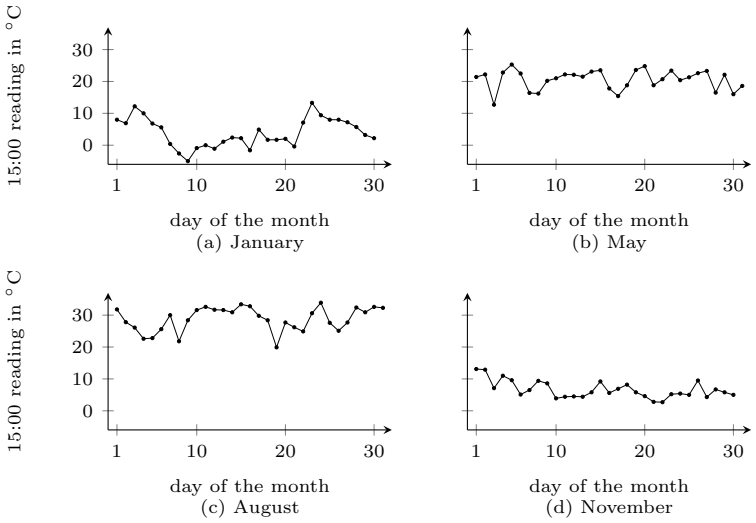

Abbildung 1: Die Lufttemperatur jeweils um 15:00 Uhr (UTC) in Krems an der Donau für vier Monate des Jahres 2024. Jeder Monat wird als ein Datensatz aufgefasst. Die Durchschnittswerte unterscheiden sich von Monat zu Monat (3.95 Grad im Januar gegenüber 28.76 im August), und innerhalb jedes Monats folgt auf einen warmen Tag tendenziell ein weiterer warmer Tag. Daten erzeugt mit `pythondemos/iidasspt.py`

Die u. i. v.-Annahme ist das in der statistischen Lerntheorie übliche Wahrscheinlichkeitsmodell (Vapnik, 2000, Sect. 1.1), das die Erzeugung der Datenpunkte $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)} \in \mathcal{Z}$ eines Datensatzes durch u.i.v.e Zufallsvariablen mit einer unbekannten Wahrscheinlichkeitsverteilung $p(\mathbf{z})$ darstellt (Shalev-Shwartz and Ben-David, 2014, Sect. 2.3.1). Für beliebige Ereignisse $\mathcal{A}_1, \ldots, \mathcal{A}_m \subseteq \mathcal{Z}$ gilt

$$\mathbb{P}\left(\mathbf{z}^{(1)} \in \mathcal{A}_1, \ldots, \mathbf{z}^{(m)} \in \mathcal{A}_m\right) = \prod_{r=1}^{m} \mathbb{P}\left(\mathbf{z}^{(r)} \in \mathcal{A}_r\right), \tag{1}$$

und jeder Faktor wird aus demselben $p(\mathbf{z})$ berechnet. Die Annahme besteht daher aus zwei getrennten Anforderungen: Die Datenpunkte folgen einer gemeinsamen Wahrscheinlichkeitsverteilung, und sie sind gemeinsam unabhängig. Gemeinsame Unabhängigkeit ist stärker als die Unabhängigkeit jedes Paares, da die Produktregel (1) für alle m Ereignisse gleichzeitig gefordert wird: Es gibt Familien, in denen jedes Paar faktorisiert, die gesamte Familie jedoch nicht (Murphy, 2012, Exercise 2.7).

Die Annahme benennt eine Zielgröße jenseits des Trainingsdatensatzes: das Risiko einer Hypothese, also ihren erwarteten Verlust $\mathbb{E}\big\{L\left(\mathbf{z}, h\right)\big\}$ auf einem aus $p(\mathbf{z})$ gezogenen Datenpunkt. Unter der u. i. v.-Annahme ist der durchschnittliche Verlust auf einem Trainingsdatensatz eine erwartungstreue Schätzung dafür (Shalev-Shwartz and Ben-David, 2014, Sect. 4.2). Sei σ^2 die Varianz des Ver-

lusts einer festen Hypothese auf einem einzelnen Datenpunkt. Der durchschnittliche Verlust über m Datenpunkte hat dann die Varianz σ^2/m, und für auf ein Intervall $[a, b]$ beschränkte Verluste begrenzt die Hoeffding-Ungleichung die Wahrscheinlichkeit, dass er um mehr als ϵ vom Risiko abweicht, durch $2\exp\left(-2m\epsilon^2/(b-a)^2\right)$ (Hoeffding, 1963; Shalev-Shwartz and Ben-David, 2014, Lemma B.6). Diese Schranke liefert das Argument für die empirische Risikominimierung und für die Verwendung eines Validierungsdatensatzes zur Schätzung des Verlusts außerhalb des Trainingsdatensatzes (Shalev-Shwartz and Ben-David, 2014, Thm. 11.1).

Die Annahme kann verletzt sein: Eine Verteilungsverschiebung verletzt die Anforderung identischer Verteilung über Datensätze hinweg, da sich die Wahrscheinlichkeitsverteilung, die den Trainingsdatensatz erzeugt hat, von derjenigen unterscheidet, welche die im praktischen Einsatz auftretenden Datenpunkte erzeugt (Quiñonero-Candela et al., 2009). Die Überprüfung, ob sich ein gegebener Datensatz durch Realisierungen u. i. v. Zufallsvariablen annähern lässt, besteht aus zwei Teilen, die jeweils eine Form des Hypothesentests darstellen. Ein erster Test prüft, ob die Datenpunkte Realisierungen unabhängiger Zufallsvariablen sind. Ein zweiter Test untersucht anschließend, ob diese Zufallsvariablen eine gemeinsame Wahrscheinlichkeitsverteilung teilen.

Ein grundlegender Unabhängigkeitstest nutzt die Konsequenz, die statistische Unabhängigkeit für die Vorhersagbarkeit eines Datenpunkts aus einigen vorangegangenen hat. Sind $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(d)}, \mathbf{z}^{(d+1)}$ Realisierungen u.i.v.er Zufallsvariablen, so kann keine Hypothese den Wert $\mathbf{z}^{(d+1)}$ aus $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(d)}$ vorhersagen.

Die Hypothese wird aus Datenpunkte gelernt, die aus der Folge gebildet werden: Jeder besteht aus einem überlappenden Block von $d+1$ aufeinanderfolgenden Werten, dessen erste d Werte seine Merkmale und dessen letzter Wert sein Label sind. Das Lernen mittels ERM auf einem Teil der Blöcke und das Messen des durchschnittlichen Verlusts auf dem Rest liefert die Vergleichsgröße. Sind die Werte Realisierungen unabhängiger Zufallsvariablen, so ähnelt dieser Verlust dem Verlust einer Hypothese, die überhaupt keine Merkmale nutzt, nämlich der besten Konstante. Jeder Hypothesenraum ist hierfür geeignet: Ein lineares Modell mit $d = 1$ unter dem quadratischen Fehlerverlust, gelernt und ausgewertet auf denselben Blöcken, senkt den durchschnittlichen Verlust gegenüber dem der besten Konstante um den Faktor $1 - r^2$, wobei r die Autokorrelation zur Verzögerung eins ist, also die Pearson-Korrelationskoeffizient zwischen aufeinanderfolgenden Werten. Der Faktor folgt aus der Methode der kleinsten Quadrate: Die Steigung ist r mal das Verhältnis der Standardabweichungen von Label und Merkmal, und das durchschnittliche quadrierte Residuum beträgt dann $1 - r^2$ mal die Stichprobenvarianz der Labels. Ein Entscheidungsbaum wiederum erfasst Abhängigkeiten, die keine Pearson-Korrelationskoeffizient erkennen kann.

Abb. 2 führt dies auf den Kremser Messreihen aus. Ein Entscheidungsbaum der Tiefe drei mit den vorangegangenen Tagen als Merkmale erreicht einen durchschnittlichen Verlust von 14.5 auf Tagen, auf denen er nicht gelernt wurde, gegenüber 87.2 für die beste Konstante; die vorangegangenen Tage enthalten

also Informationen über den nächsten Tag, und die Datenpunkte sind keine Realisierungen unabhängiger Zufallsvariablen.

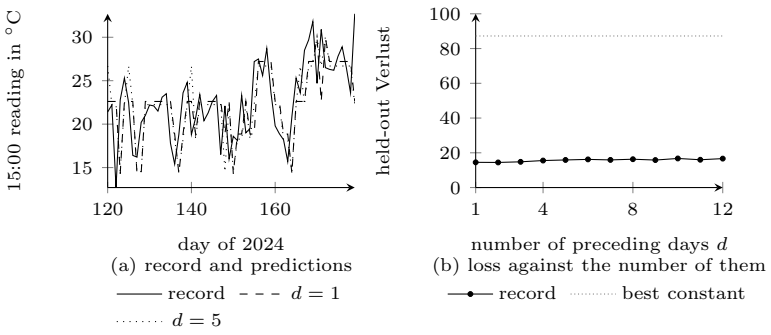

Abbildung 2: (a) Sechzig Tage der Messreihe aus Krems an der Donau und die Vorhersagen eines Entscheidungsbaums der Tiefe drei auf Basis der d vorangegangenen Tage. (b) Der durchschnittliche Verlust dieses Entscheidungsbaums auf Tagen, auf denen er nicht gelernt wurde, aufgetragen über d, mit dem Verlust der besten Konstante als gepunkteter Linie. Die Einbeziehung der vorangegangenen Tage senkt den Verlust auf etwa ein Sechstel davon. Daten erzeugt mit `pythondemos/iidasspt.py`

Diese Prüfung erkennt nur jene Abhängigkeit, die ihr Hypothesenraum und ihr Label ausdrücken können. Eine Folge kann abhängig sein und dennoch einen konstanten bedingten Erwartungswert besitzen, und dann hilft kein Hypothesenraum weiter. Für $x^{(r)} = \varepsilon^{(r)}\varepsilon^{(r-1)}$ mit u.i.v.en standardnormalverteilten Gaußsche Zufallsvariablen ε sagen Entscheidungsbäume der Tiefe eins bis vier den nächsten Wert allesamt schlechter voraus als die beste Konstante, und zwar monoton schlechter mit zunehmender Baumtiefe. Eine Vergrößerung des Hypothesenraums senkt keinen Verlust, wenn es im bedingten Erwartungswert nichts zu finden gibt, und verschlechtert die Generalisierung: Der größere Hypothesenraum senkt lediglich den Verlust auf jenen Tagen, auf denen die Hypothese trainiert wurde. Wird stattdessen der quadrierte Wert vorhergesagt, sinkt der Verlust unter den der Konstante. Das Vergrößern des Hypothesenraums deckt Abhängigkeiten folglich nur dort auf, wo sie im bedingten Erwartungswert auftreten; eine Änderung dessen, was die Hypothese vorhersagen soll, macht sie an anderer Stelle sichtbar. Keines von beidem allein macht die Prüfung zu einem vollständigen Test auf Unabhängigkeit (siehe Autokorrelation).

Die Anforderung identischer Verteilung muss wiederum gesondert überprüft werden. Dazu wird der Datensatz in gleich große Blöcke unterteilt und es werden die empirischen Verteilungen zweier Blöcke anhand des größten vertikalen Abstands zwischen ihren Verteilungsfunktionen verglichen. Dieser Abstand ist die Zweistichproben-Kolmogorov-Smirnov-Statistik (Gretton et al., 2012, Sect. 7.2.1). Sie beträgt höchstens eins und ist genau dann gleich eins, wenn jeder Wert des einen Blocks unter jedem Wert des anderen liegt.

Die Kremser Messreihe verletzt auch diese Anforderung deutlich. Abb. 3 wendet den Blockvergleich über Zeiträume wachsender Länge an. Innerhalb des Januars beträgt der größte Abstand zwischen den beiden 15-Tage-Blöcken 0.33; von Januar bis Juni sowie über das gesamte Jahr hinweg beträgt er 1, sodass jeder Wert eines der beiden Blöcke unter jedem Wert des anderen liegt.

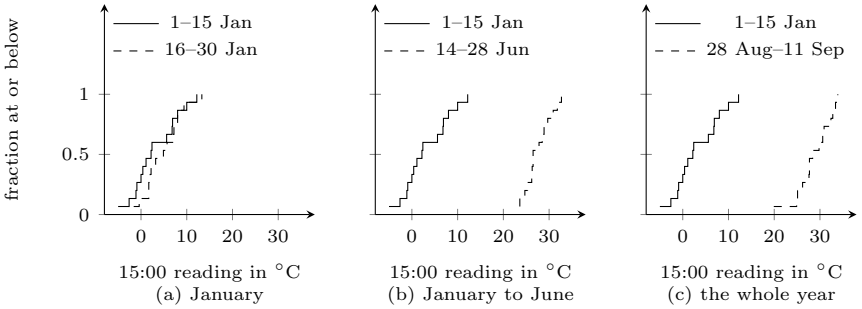

Abbildung 3: Empirische Verteilungsfunktionen der beiden am weitesten voneinander entfernten 15-Tage-Blöcke für drei Zeiträume der Messreihe aus Krems an der Donau. Daten erzeugt mit `pythondemos/iidasspt.py`

Für zwei Blöcke mit jeweils n Werten erzeugen u. i. v. Zufallsvariablen nach der Zweistichprobenform der Dvoretzky-Kiefer-Wolfowitz-Ungleichung mit einer Wahrscheinlichkeit von höchstens α einen Abstand über $\sqrt{\ln(2.17/\alpha)/n}$ (Wei and Dudley, 2012, Thm. 1). Die Konstante 2.17 ist der Wert, den zwei gleich große Blöcke mit $n \geq 4$ erfordern. Der für die Einstichproben-Ungleichung scharfe Wert 2 (Massart, 1990, p. 1270) gilt im Zweistichprobenfall erst ab $n = 458$ (Wei and Dudley, 2012, Thm. 1). Für $n = 15$ und $\alpha = 0.05$ liegt der Schwellenwert bei 0.50, den der Januar-Abstand von 0.33 nicht erreicht. Ein Abstand unterhalb seines Schwellenwerts beweist darüber hinaus nichts: Die Blöcke liefern dann keinen Beleg gegen eine gemeinsame Wahrscheinlichkeitsverteilung, während eine Abhängigkeit zwischen aufeinanderfolgenden Werten weiterhin bestehen kann. Beide Teile der Annahme erfordern daher eine eigene Überprüfung. Eine gröbere Prüfung vergleicht Monatsdurchschnitte: Der Januar liegt im Mittel bei 3.95 Grad und der Juli bei 27.21. Nimmt man die gemeinsame Wahrscheinlichkeitsverteilung als gaußsch an, wobei ihr Erwartungswert und ihre Varianz auf den Stichprobenmittel 16.30 und die Stichprobenvarianz des gesamten Jahres (eine Standardabweichung von 9.35 Grad) gesetzt werden, so hat die Differenz zwischen den Mittelwerten der 30 Januartage und der 31 Julitage den Erwartungswert null und eine Standardabweichung von 2.40 Grad. Der beobachtete Abstand von 23.27 Grad liegt damit 9.7 Standardabweichungen von null entfernt, was diese Wahrscheinlichkeitsverteilung mit einer Wahrscheinlichkeit von unter 10^{-21} hervorbringt. Das Abziehen des saisonalen Durchschnitts beseitigt diese Verletzung, während die Abhängigkeit bestehen bleibt.

Siehe auch: u.i.v., Autokorrelation, Ljung-Box-Test, Wahrscheinlichkeitsmodell, Datenpunkt, ZV, Risiko, ERM, Trainingsdatensatz, Validierungsdatensatz,

Verteilungsverschiebung, empirische Verteilung, Hypothesentest, Hypothesenraum, Entscheidungsbaum, Realisierung.

Querverweise

Literatur

1. Gretton et al. (2012). A Kernel Two-Sample Test. Journal of Machine Learning Research, 13(25), pp. 723–773.

2. Hoeffding (1963). Probability Inequalities for Sums of Bounded Random Variables. J. Amer. Statistical Assoc., 58(301), pp. 13–30.

3. Massart (1990). The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality. Ann. Probab., 18(3), pp. 1269–1283.

4. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

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

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

7. Vapnik (2000). The Nature of Statistical Learning Theory. 2nd ed., Springer, New York, NY, USA.

8. Wei and Dudley (2012). Two-sample Dvoretzky–Kiefer–Wolfowitz inequalities. Statist. Probab. Lett., 82(3), pp. 636–644.