Stichprobengröße

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 Stichprobengröße ist die Anzahl m der Datenpunkte in einer Stichprobe. Sie ist das grundlegende Maß dafür, wie viel eine Stichprobe enthält, jedoch kein universelles: Das zweifache Auflisten eines Datenpunkts erhöht die Anzahl, trägt aber nichts bei; korrelierte Datenpunkte führen dazu, dass die Anzahl den Gehalt einer Stichprobe überschätzt, und sie sagt nichts darüber aus, wie gut die Datenpunkte den Merkmalsraum abdecken. Viele Methoden des maschinelles Lernen (ML)s modellieren die Datenpunkte als Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) und bewerten eine Hypothese anhand ihres erwarteten Verlusts, des Risikos; da die Wahrscheinlichkeitsverteilung unbekannt ist, wird das Risiko durch den mittleren Verlust auf der Stichprobe geschätzt, für das Training mittels empirische Risikominimierung (ERM) sowie für die Validierung. Wie nahe dieser Mittelwert dem Risiko kommt, hängt von der Stichprobengröße ab. Eine Konzentrationsungleichung wie die Hoeffding-Ungleichung beschränkt die Abweichung des Mittelwerts vom Risiko und übersetzt eine geforderte Genauigkeit ϵ und ein Konfidenzniveau $1-\delta$ in eine Stichprobengröße der Größenordnung $\ln(1/\delta)/\epsilon^2$ für die Validierung; für das Training muss die Stichprobengröße zudem mit der Größe des Modells verglichen werden. Bei nacheinander aufgezeichneten Datenpunkte verringert die Autokorrelation die effektive Stichprobengröße. Dieselbe Überlegung liefert den Umfang der Stichprobe für einen statistischen Test, etwa darauf, ob zwei Merkmale korreliert sind: Die zum Nachweis einer Korrelation benötigte Stichprobengröße wächst umgekehrt quadratisch mit dieser Korrelation.

Definition

Eine Frostwarnung für Krems an der Donau (Österreich) sagt anhand der Lufttemperatur eines Tages um 06:00 (UTC) und um 15:00 (UTC) voraus, ob der nächste Tag um 06:00 Frost bringt. Eine Hypothese h für diese Aufgabe wird anhand ihres erwarteten Verlusts bewertet, des Risikos $\bar{L}\big(h\big) = \mathbb{E}\{L\left(\mathbf{z}, h\right)\}$, wobei der Erwartungswert über einen Datenpunkt $\mathbf{z}$ gebildet wird, der aus der Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{z})}$ der Tage gezogen wird. Viele Methoden des maschinelles Lernen (ML)s nutzen die Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme): Die Datenpunkte sind Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)} \sim$

$\mathbb{P}^{(\mathbf{z})}$. In der Praxis ist $\mathbb{P}^{(\mathbf{z})}$ unbekannt, und eine Methode beobachtet nur diese m Datenpunkte, eine Stichprobe. Deren Anzahl m ist die Stichprobengröße.

Da die Wahrscheinlichkeitsverteilung unbekannt ist, wird das Risiko durch den mittleren Verlust auf der Stichprobe geschätzt,

$$\widehat{L}\big(h|\mathcal{D}\big) := \frac{1}{m}\sum_{r=1}^{m} L\left(\mathbf{z}^{(r)}, h\right), \tag{1}$$

was dem empirisches Risiko entspricht. Die Empirische Risikominimierung (ERM) minimiert diesen Mittelwert über die Hypothesen eines Modells auf einem Trainingsdatensatz, und die Validierung vergleicht die gelernten Hypothesen anhand desselben Mittelwerts auf einem Validierungsdatensatz. Beide Verwendungen setzen voraus, dass der Mittelwert nahe am Risiko liegt, und wie nahe er ihm kommt, hängt von der Stichprobengröße ab: Nach dem Gesetz der großen Zahlen konvergiert der Mittelwert (1) mit wachsendem m gegen $\bar{L}\big(h\big)$ (Papoulis and Pillai, 2002, Abschn. 8-4; Shalev-Shwartz and Ben-David, 2014, Abschn. 4.2). Abb. 1 zeigt dies für die Frostwarnung: Das Stichprobenmittel des 0/1-Verlusts einer festen Hypothese, berechnet auf Stichproben von $m = 10$ bis 1585 Tagen, streut um das Risiko 0.099 mit einer Breite, die wie $1/\sqrt{m}$ abnimmt. Diese Streubreite ist die Standardabweichung des Mittelwerts über wiederholte Stichproben, und die nachfolgend angegebene Hoeffding-Ungleichung überführt sie in eine Abweichung, die mit einer vorgegebenen Wahrscheinlichkeit eingehalten wird.

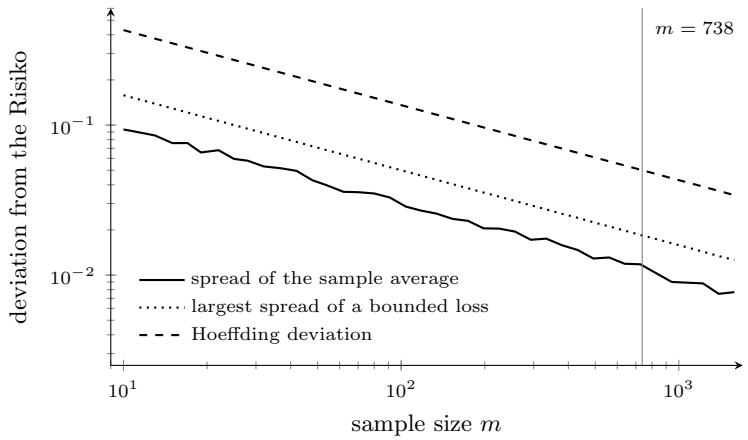

Abbildung 1: Das Stichprobenmittel des 0/1-Verlusts einer festen Frostwarnungs-Hypothese an m gleichverteilt mit Zurücklegen gezogenen Tagen des Jahres 2024 in Krems an der Donau in Abhängigkeit von der Stichprobengröße. Die durchgezogene Kurve ist die Standardabweichung (Streuung) des Mittelwerts über 400 Ziehungen, die gepunktete Kurve $1/(2\sqrt{m})$, die größte Standardabweichung, die ein Verlust in $[0, 1]$ zulässt; die gestrichelte Kurve ist die Abweichung ϵ, welche die Hoeffding-Ungleichung (2) mit einer Wahrscheinlichkeit von 0.95 garantiert, und die vertikale Linie markiert die Stichprobengröße 738, bei der diese Garantie $\epsilon = 0.05$ erreicht. Daten erzeugt mit `pythondemos/samplesize.py`

Die Stichprobe muss groß genug sein, und ein Ansatz zur Bestimmung der erforderlichen Größe nutzt eine Konzentrationsungleichung, welche die Wahrscheinlichkeit beschränkt, dass der Mittelwert um mehr als einen gegebenen Betrag vom Risiko abweicht. Für eine Verlustfunktion mit Werten in $[0, 1]$, wie etwa den 0/1-Verlust, liefert die Hoeffding-Ungleichung für jedes $\epsilon > 0$:

$$\mathbb{P}\left(\left|\widehat{L}\big(h|\mathcal{D}\big) - \bar{L}\big(h\big)\right| > \epsilon\right) \leq 2\exp\big(-2m\epsilon^2\big), \tag{2}$$

siehe Hoeffding (1963) und Shalev-Shwartz and Ben-David (2014), Lemma 4.5. Die Schranke gilt für eine Hypothese, die vor dem Ziehen der Stichprobe festgelegt wurde, was der Situation bei der Validierung entspricht: Die Hypothese wurde auf dem Trainingsdatensatz gelernt, und das Validierungsdatensatz ist eine neue Stichprobe. Setzt man die rechte Seite gleich einer tolerierten Fehler-Wahrscheinlichkeit δ und löst nach m auf, so werden eine geforderte Genauigkeit ϵ und ein Konfidenzniveau $1 - \delta$ in eine Stichprobengröße überführt:

$$m \geq \frac{\ln(2/\delta)}{2\epsilon^2}, \tag{3}$$

und bei so vielen Datenpunkte liegt der Validierungsfehler mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ innerhalb einer Umgebung von ϵ um das Risiko

(Shalev-Shwartz and Ben-David, 2014, Thm. 11.1). Für die Frostwarnung ergeben $\epsilon = 0.05$ und $\delta = 0.05$ den Wert $m \geq 738$ Tage; bei dieser Stichprobengröße liegen alle 2000 gezogenen Stichproben (`pythondemos/samplesize.py`) innerhalb von 0.05 um das Risiko. Die Schranke ist grob, da sie für jede Wahrscheinlichkeitsverteilung und jede Verlustfunktion in $[0, 1]$ gilt; die Stichprobengröße (3) ist daher hinreichend, aber nicht immer notwendig. Entscheidend ist ihre Gestalt: Eine Halbierung von ϵ vervierfacht die Stichprobengröße, während die Stichprobengröße in $1/\delta$ nur logarithmisch wächst.

Für das Training hat dieselbe Frage eine zweite Antwort. Die Hypothese wird dann unter Betrachtung der Stichprobe ausgewählt, sodass die Schranke (2) für eine feste Hypothese auf die gelernte nicht mehr anwendbar ist. Die Stichprobengröße muss dann mit der Größe des Modells verglichen werden. Für ein Modell der Vapnik-Chervonenkis-Dimension (VC-Dimension) d sorgt eine Stichprobengröße proportional zu d/ϵ^2 dafür, dass der Trainingsfehler jeder Hypothese im Modell gleichzeitig innerhalb von ϵ um ihr Risiko liegt (Shalev-Shwartz and Ben-David, 2014, Thm. 6.8). Eine im Vergleich zur effektive Dimension des Modells kleine Stichprobe ist die typische Ursache für Überanpassung, und die Lernkurve zeigt den Validierungsfehler der gelernten Hypothese als Funktion der Stichprobengröße (Hastie et al., 2009, Abschn. 7.10).

Die Stichprobengröße geht in eine dritte Aufgabe ein, den Hypothesentest: Sind zwei Merkmale eines Datenpunkts korreliert? Ist für die Tage des Jahres 2024 in Krems der Messwert eines Tages um 06:00 Uhr mit seinem Anstieg, dem Messwert um 15:00 Uhr abzüglich jenem um 06:00 Uhr, korreliert? Über das gesamte Jahr beträgt der Pearson-Korrelationskoeffizient der beiden Merkmale $\rho = 0.09$. Ein Test erhält eine Stichprobe von m Tagen und berechnet die Stichproben-Korrelation r, eine Funktion von Stichprobenmittel. Unter der Hypothese, dass die beiden Merkmale unabhängig sind, streut r um 0 mit einer Standardabweichung nahe $1/\sqrt{m}$. Für Merkmale, die Gaußsche Zufallsvariablen sind, beträgt die Varianz von r exakt $1/(m-1)$ (Anderson, 2003, Thm. 4.2.1). Für unabhängige Merkmale mit endlichen Varianzen jeder anderen Wahrscheinlichkeitsverteilung konvergiert $\sqrt{m}\, r$ nach dem zentraler Grenzwertsatz (ZGS), angewandt auf die Produkte der standardisierten Merkmale, deren Varianz unter Unabhängigkeit eins ist, gegen eine standardisierte Gaußsche Zufallsvariable. Für Merkmale, die Gaußsche Zufallsvariablen sind, ist die Fisher-Transformation $\operatorname{artanh}(r)$ näherungsweise eine Gaußsche Zufallsvariable mit Mittelwert $\operatorname{artanh}(\rho)$ und Standardabweichung $1/\sqrt{m-3}$ (Anderson, 2003, Thm. 4.2.5; Fisher, 1921). Der Test verwirft „unkorreliert“, wenn $|\operatorname{artanh}(r)|\sqrt{m-3}$ das Quantil $z_{1-\alpha/2}$ der standardisierten Gaußsche Zufallsvariable überschreitet, was die Wahrscheinlichkeit einer fälschlichen Ablehnung durch das Signifikanzniveau α beschränkt. Eine Korrelation ρ wird dann mit einer Wahrscheinlichkeit von mindestens $1 - \beta$, der Trennschärfe (Teststärke) des Tests, erkannt, sobald

$$m \geq 3 + \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\operatorname{artanh}(\rho)} \right)^2 . \tag{4}$$

Für $\rho = 0.09$, $\alpha = 0.05$ ($z_{0.975} = 1.96$) und eine Trennschärfe von 0.8 ($z_{0.8} =$

0.84) beträgt die Stichprobengröße (4) 968 Tage. Abb. 2 zeigt die Trennschärfe des Tests in Abhängigkeit von der Stichprobengröße: 0.82 bei $m = 968$ und 0.06 für eine Stichprobe von 30 Tagen. Der Verlauf entspricht wieder jenem von (3): Eine Halbierung der nachzuweisenden Korrelation vervierfacht die Stichprobengröße, sodass eine schwache Korrelation eine große Stichprobe benötigt, bevor „unkorreliert“ verworfen werden kann. Hier übersteigt die erforderliche Größe die 363 Tage des Jahres, die ein Label tragen, sodass eine derart schwache Korrelation aus einem einzigen Jahr an dieser Station überhaupt nicht nachgewiesen werden kann.

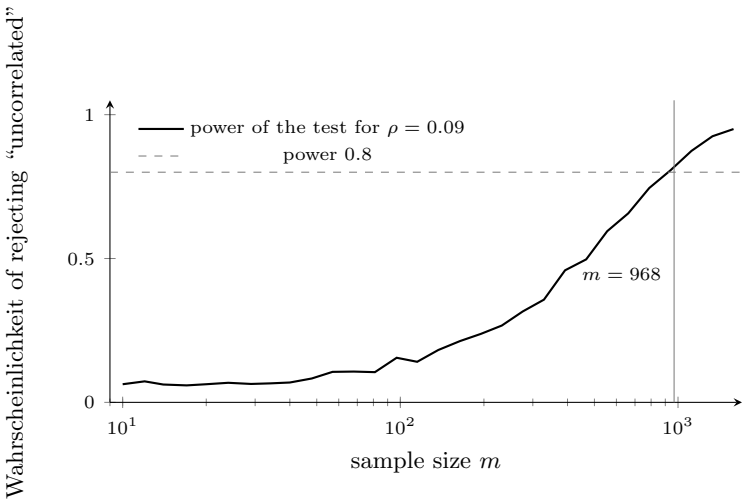

Abbildung 2: Die Trennschärfe des Tests auf Korrelation zum Signifikanzniveau $\alpha = 0.05$ für den Messwert eines Tages um 06:00 Uhr (UTC) in Krems an der Donau und seinen Anstieg bis 15:00 Uhr (UTC), deren Pearson-Korrelationskoeffizient über das Jahr 2024 bei $\rho = 0.09$ liegt, in Abhängigkeit von der Stichprobengröße (durchgezogene Kurve, der Anteil von 1000 Stichproben zu je m Tagen, bei denen der Test „unkorreliert“ verwirft). Die gestrichelte Linie markiert die Trennschärfe 0.8, die vertikale Linie die Stichprobengröße 968 aus (4). Daten erzeugt mit `pythondemos/samplesize.py`

Die Anzahl m ist für sich genommen kein Maß dafür, wie viel Information eine Stichprobe trägt. Wäre sie es, ließe sich Information durch Kopieren erzeugen: Führt man jeden Tag einer Stichprobe zweimal auf, ergibt dies eine Stichprobe der Stichprobengröße $2m$, die genau dasselbe aussagt wie das Original. Ihre empirische Verteilung ist unverändert, da jeder Tag nun zwei Einträge der Wahrscheinlichkeit $1/(2m)$ hält. Ebenso unverändert bleibt der mittlere Verlust (1) jeder Hypothese sowie die Streuung dieses Mittelwerts über wiederholte Stichproben. Nur $2m$ neu gezogene Tage verringern diese Streuung, und zwar um den Faktor $\sqrt{2}$. Läse man die Schranke (2) mit $2m$ anstelle von m, würde sie eine um genau diesen Faktor kleinere Abweichung versprechen. Dieses Ver-

sprechen gilt jedoch nicht, da eine Kopie eines Tages nicht unabhängig von dem Tag ist, den sie kopiert, während die Hoeffding-Ungleichung für Realisierungen von u.i.v. Zufallsvariablen gilt. Die Stichprobengröße zählt die Einträge einer Liste, und wie viel diese aussagen, hängt davon ab, wie die Liste erzeugt wurde.

Abhängigkeit ist selten so extrem wie eine direkte Kopie. Die obigen Stichprobengrößen setzen u.i.v. Zufallsvariablen voraus, und nacheinander aufgezeichnete Datenpunkte, wie etwa aufeinanderfolgende Tage, weisen oft eine Autokorrelation auf: Ein Frosttag ist nach einem Frosttag wahrscheinlicher, und ebenso eine falsche Vorhersage. Abhängige Datenpunkte tragen weniger Information als unabhängige. Für eine stationäre Folge mit der Autokorrelation ρ_ℓ bei der Verzögerung ℓ ist die Varianz des Mittelwerts von m aufeinanderfolgenden Werten gleich jener von m unabhängigen Werten multipliziert mit

$$1 + 2 \sum_{\ell=1}^{m-1} \Big(1 - \frac{\ell}{m}\Big) \rho_\ell, \tag{5}$$

siehe Brockwell and Davis (1991), Thm. 7.1.1. Der Mittelwert ist daher genauso genau wie jener von m geteilt durch (5) unabhängigen Werten, der effektiven Stichprobengröße. Für den 0/1-Verlust der Frostwarnung beträgt die Stichproben-Autokorrelation aufeinanderfolgender Tage 0.26 bei Verzögerung 1 und bleibt eine Woche lang positiv, und der Faktor (5) beträgt 5.6 für $m = 30$. Der Mittelwert über 30 aufeinanderfolgende Tage streut daher genauso stark wie jener über 5 zufällig gezogene Tage, und das gemessene Varianz-Verhältnis von 5.1 liegt bis auf ein Zehntel bei diesem Faktor. Der Faktor (5) setzt eine stationäre Folge voraus, was der 0/1-Verlust einer Frostwarnung nicht ist: Über 60 aufeinanderfolgende Tage sagt er ein Verhältnis von 8.4 voraus, während das gemessene bei 5.6 liegt, da Frost auf den Winter beschränkt ist. Eine Stichprobe aufeinanderfolgender Datenpunkte muss dann um den Faktor (5) größer sein, als (3) oder (4) angeben, oder zufällig aus der gesamten Aufzeichnung gezogen werden, wie es bei den Tagen in Abb. 1 der Fall ist.

Abgesehen von Abhängigkeiten sagt die Anzahl nichts darüber aus, wo die Datenpunkte liegen. Zwei Stichproben derselben Größe können den Merkmalsraum unterschiedlich abdecken: die eine über den gesamten Wertebereich der Merkmale verteilt, die andere in einem Teil davon gedrängt. Für eine Methode, deren Vorhersage an einem Merkmalsvektor von den Datenpunkte in seiner Nähe abhängt, bestimmt der größte Abstand von einem Punkt des interessierenden Bereichs im Merkmalsraum zum nächsten Datenpunkt den dortigen Fehler der Vorhersage, wie groß m auch sein mag (siehe Überdeckung). Eine Wetteraufzeichnung, die nur aus Sommertagen besteht, mag noch so lang sein: Sie lässt den Winterteil des Merkmalsraumes leer, und keine Stichprobengröße kann dies ausgleichen.

Siehe auch: Stichprobe, Datenpunkt, Datensatz, u.i.v.-Annahme, Risiko, empirisches Risiko, Gesetz der großen Zahlen, Konzentrationsungleichung, Hoeffding-Ungleichung, Validierung, Validierungsfehler, VC-Dimension, Lernkurve, Überanpassung, Autokorrelation, empirische Verteilung, Hypothesentest, Pearson-Korrelationskoeffizient.

Querverweise

Literatur

1. Anderson (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., Wiley, Hoboken, NJ.

2. Brockwell and Davis (1991). Time Series: Theory and Methods. 2nd ed., Springer-Verlag, New York, NY, USA.

3. Fisher (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1(4), pp. 3–32.

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

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

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

7. Papoulis and Pillai (2002). Probability, Random Variables, and Stochastic Processes. 4th ed., McGraw-Hill Higher Education, New York, NY, USA.