Mittelwert

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

Der Mittelwert eines Zufallsvektor $\mathbf{x}$ ist dessen Erwartungswert, das Lebesgue integral von $\mathbf{x}$ bezüglich seiner Wahrscheinlichkeitsverteilung. Der Begriff bezeichnet auch den Stichprobenmittel, also den Durchschnitt der Datenpunkte in einem Datensatz. Die beiden Verwendungen sind konsistent: Der Stichprobenmittel ist der Mittelwert des Zufallsvektor, der entsteht, wenn ein Datenpunkt gleichverteilt aus dem Datensatz gezogen wird. Beide Verwendungen teilen eine weitere Charakterisierung: Für einen Zufallsvektor mit endlichem zweitem Moment ist der Mittelwert die eindeutige Lösung eines Risiko- Minimierungsproblems unter der squared error loss. Für den Zufallsvektor aus einem Datensatz reduziert sich dieses Optimierungsproblem auf empirische Risikominimierung (ERM) im einfachsten regression-Szenario: die Vorhersage eines numerischen Label ohne Merkmale, wobei die gelernte Hypothese der Stichprobenmittel der Labels ist.

Definition

Eine Wetterstation misst die Lufttemperatur um 15:00 Uhr (UTC) über ein Jahr. Die 365 Werte zu addieren und durch 365 zu teilen ergibt einen einzelnen Wert, der die gesamte Messreihe zusammenfasst. Dieser Wert ist der Mittelwert der aufgezeichneten Zahlen, und dasselbe Wort bezeichnet die entsprechende Größe für eine Wahrscheinlichkeitsverteilung.

Der Mittelwert eines Zufallsvektor $\mathbf{x}$, der Werte im Euclidean space $\mathbb{R}^d$ annimmt, ist dessen Erwartungswert $\mathbb{E}\{\mathbf{x}\}$. Dies ist das Lebesgue integral von $\mathbf{x}$ bezüglich seiner Wahrscheinlichkeitsverteilung $\mathbb{P}$ (siehe z. B. (Billingsley, 1986, Thm. 16.12, Eq. (21.17)) oder (Rudin, 1987)), also

$$\mathbb{E}\{\mathbf{x}\} = \int_{\mathbb{R}^d} \mathbf{x} \, \mathrm{d}\mathbb{P}(\mathbf{x}).$$

Hier ist $\mathbb{P}$ die von $\mathbf{x}$ auf $\mathbb{R}^d$ induzierte Wahrscheinlichkeitsverteilung, sodass das Integral über die Werte läuft, die $\mathbf{x}$ annimmt, und nicht über den zugrundeliegenden sample space. Es muss keine Wahrscheinlichkeitsdichte existieren, damit der Mittelwert definiert ist; wenn eine existiert, bezeichnet mit p, gilt $\int_{\mathbb{R}^d} \mathbf{x}\, p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ (Billingsley, 1986, Thm. 16.10).

Der Durchschnitt aus dem einleitenden Absatz ist der Stichprobenmittel eines endlichen Datensatz der Stichprobengröße m, $\mathcal{D} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d\right\}$. Beide Verwendungen sind konsistent. Ein Datensatz definiert eine diskrete Zufallsvariable $\widetilde{\mathbf{x}}^{(\mathcal{D})} = \mathbf{x}^{(I)}$ auf dem sample space $\{1, \ldots, m\}$. Hierbei wird der Index I gleichverteilt gewählt, d. h. $\mathbb{P}(I = r) = 1/m$ für alle $r = 1, \ldots, m$. Der Mittelwert von $\widetilde{\mathbf{x}}^{(\mathcal{D})}$ ist genau der Stichprobenmittel

$$(1/m) \sum_{r=1}^{m} \mathbf{x}^{(r)}.$$

Die Wahrscheinlichkeitsverteilung von $\widetilde{\mathbf{x}}^{(\mathcal{D})}$, die jeweils Masse $1/m$ auf jeden Datenpunkt legt, ist die empirische Verteilung des Datensatz. Abb. 1 zeigt beide Lesarten für eine reellwertige Zufallsvariable (ZV) und ein Datensatz aus sieben Zahlen, wobei die empirische Verteilung als deren Verteilungsfunktion dargestellt ist.

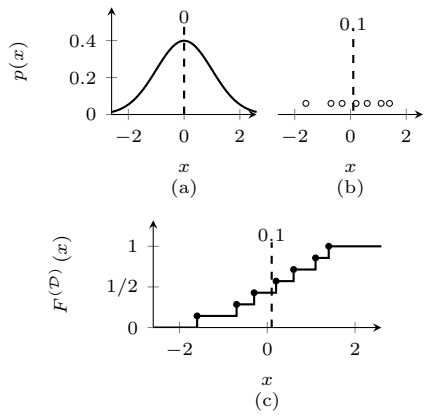

Abbildung 1: Die zwei Lesarten von *Mittelwert*. (a) Eine reellwertige ZV mit der Standardnormal- Wahrscheinlichkeitsdichte hat den Mittelwert $\mathbb{E}\{x\} = 0$, das Lebesgue integral von x bezüglich dieser Wahrscheinlichkeitsdichte. (b) Die $m = 7$ Datenpunkte mitteln sich zum Stichprobenmittel 0.1. (c) Die empirische Verteilung derselben Datenpunkte, dargestellt als deren Verteilungsfunktion $F^{(\mathcal{D})}(x)$: eine Treppe, die bei jedem Datenpunkt um $1/m$ ansteigt, von 0 bis 1. Sie ist die Wahrscheinlichkeitsverteilung der diskrete Zufallsvariable $\widetilde{\mathbf{x}}^{(\mathcal{D})}$ oben, und ihr Mittelwert ist der Stichprobenmittel aus (b). Der Stichprobenmittel 0.1 aus (b) und (c) unterscheidet sich vom Mittelwert 0 der ZV in (a), weil m endlich ist

Die beiden Lesarten sind durch eine gemeinsame Charakterisierung weiter verbunden. Betrachte einen beliebigen Zufallsvektor mit endlichem zweitem Moment, also $\mathbb{E}\{\|\mathbf{x}\|_2^2\} < \infty$. Dessen Mittelwert ist die eindeutige Lösung des fol-

genden Risiko-Minimierungsproblems (Bertsekas and Tsitsiklis, 2008):

$$\mathbb{E}\{\mathbf{x}\} = \arg\min_{\mathbf{c}\in\mathbb{R}^d} \mathbb{E}\big\{\|\mathbf{x}-\mathbf{c}\|_2^2\big\}.$$

Die Zielfunktion dieses Problems ist eine strictly convex Funktion der Optimierungsvariablen $\mathbf{c} \in \mathbb{R}^d$.

Für die diskrete Zufallsvariable, die ein Datensatz definiert, ist das Risiko der durchschnittliche quadratische Fehler über das Datensatz, sodass diese Risiko-Minimierungs-Charakterisierung direkt auf das einfachste machine learning (ML)-Problem anwendbar ist: die Vorhersage eines numerischen Label ohne jegliche Merkmale. Betrachte ein Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \big\{y^{(1)}, \ldots, y^{(m)}\big\}$ aus Labels $y^{(r)} \in \mathbb{R}$. Eine Hypothese ist dann eine einzelne Zahl $h \in \mathbb{R}$, die als Vorhersage für jeden Datenpunkt dient. Empirische Risikominimierung (ERM) mit der squared error loss sucht über Kandidaten-Hypothesen $h' \in \mathbb{R}$ und lernt

$$\widehat{h} = \arg\min_{h'\in\mathbb{R}} \frac{1}{m}\sum_{r=1}^{m}\big(y^{(r)}-h'\big)^2 = \frac{1}{m}\sum_{r=1}^{m} y^{(r)},$$

d. h., die gelernte Hypothese ist der Stichprobenmittel der Labels (siehe Abb. 2). Dieses Szenario entspricht der lineare Regression für Datenpunkte, die die konstante skalare Merkmal $x = 1$ tragen.

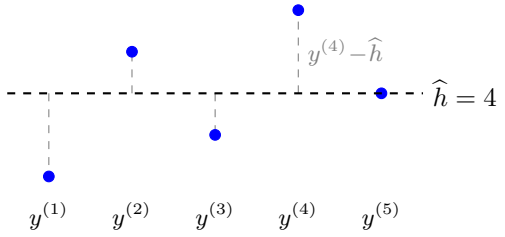

Abbildung 2: Ein Trainingsdatensatz aus $m = 5$ numerischen Labels $y^{(r)}$ (ausgefüllte Kreise) und die gelernte Hypothese $\widehat{h} = 4$ (horizontale gestrichelte Linie), die der Stichprobenmittel der Labels ist. Die vertikalen grauen Segmente sind die Fehler $y^{(r)} - \widehat{h}$; der Stichprobenmittel minimiert deren durchschnittliches Quadrat

Das Quadrieren der Abweichungen macht den Mittelwert auch empfindlich gegenüber Ausreißer. Wird einer von m Datenpunkte um einen Betrag δ verschoben, so ändert sich der Stichprobenmittel um δ/m, sodass ein einziger verfälschter Datenpunkt den Mittelwert beliebig weit verschieben kann. Der trimmed mean, der an jedem Ende einen Anteil γ verwirft, bleibt aber beschränkt, egal wie weit höchstens γm Datenpunkte verschoben werden, da die verschobenen Datenpunkte dann zu den verworfenen gehören. Der median bleibt beschränkt, solange weniger als die Hälfte der m Datenpunkte verschoben werden, da die unveränderten Datenpunkte dann die Mehrheit bilden und

die beiden mittleren Werte der sortierten Liste zwischen dem kleinsten und größten dieser liegen. Beide werden daher bevorzugt, wenn Merkmale oder Labels möglicherweise unzuverlässig sind (Huber, 1992; `pythondemos/mean.py`).

Synonyme: Erwartungswert, expected value.

Siehe auch: ZV, Zufallsvektor, Erwartungswert, Stichprobenmittel, Wahrscheinlichkeitsverteilung, Lebesgue integral, ERM, squared error loss, lineare Regression, median, trimmed mean.

Querverweise

Literatur

1. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.

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

3. Rudin (1987). Real and Complex Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

4. Huber (1992). Robust estimation of a location parameter. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 492–518. Springer-Verlag.