Wahrscheinlichkeitsverteilung

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 Wahrscheinlichkeitsverteilung einer Zufallsvariable (ZV) ist das probability measure, das die ZV auf der Menge ihrer Werte induziert: Sie weist jedem measurable Satz von Werten eine Wahrscheinlichkeit zu. Eine machine learning (ML)-Methode liest einen Datensatz ein und liefert eine Ausgabe, wie eine gelernte Hypothese oder die Vorhersage für einen gegebenen Datenpunkt, und sollte für jeden Datensatz gut funktionieren, der in der vorliegenden ML-Anwendung auftreten könnte. Die Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) modelliert einen solchen Datensatz als Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen mit einer gemeinsamen Wahrscheinlichkeitsverteilung, die aus Domänenwissen postuliert oder aus einem Datensatz geschätzt wird. Die Wahrscheinlichkeitsverteilung einer binären ZV wird vollständig durch eine einzelne Wahrscheinlichkeit spezifiziert, die einer diskrete Zufallsvariable durch eine probability mass function (pmf) und die einer reellwertigen ZV mit einer Wahrscheinlichkeitsdichte durch diese Wahrscheinlichkeitsdichte.

Definition

Man betrachte eine machine learning (ML)-Methode, die als Eingabe einen Datensatz $\mathcal{D} = \big\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\big\}$ erhält, wobei jeder Datenpunkt ein numerisches Merkmal gepaart mit einem numerischen Label ist, und eine Ausgabe liefert, wie eine gelernte Hypothese $\widehat{h}$ oder die Vorhersage für einen gegebenen Datenpunkt. Die Eingabe eines anderen Datensatz in dieselbe Methode kann eine andere Ausgabe liefern. Im Allgemeinen sollte die Methode für jeden Datensatz gut funktionieren, der in der vorliegenden ML-Anwendung auftreten könnte. Die Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) modelliert einen solchen Datensatz als unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)} \sim \mathbb{P}$ mit einer gemeinsamen Wahrscheinlichkeitsverteilung $\mathbb{P}$ (siehe Abb. 1). Die Ausgaben, die aus verschiedenen solchen Datensätze erhalten werden, sind selbst Realisierungen einer ZV, deren Wahrscheinlichkeitsverteilung von $\mathbb{P}$ und der Methode abhängt. Wie stark die Ausgaben streuen, d.h. wie stark sie voneinander abweichen, wird daher durch $\mathbb{P}$ und die Methode bestimmt.

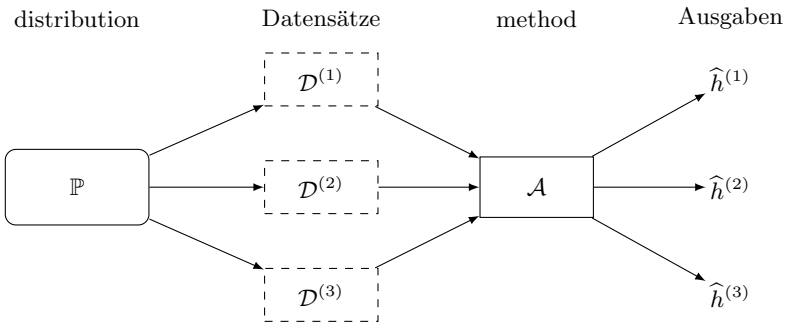
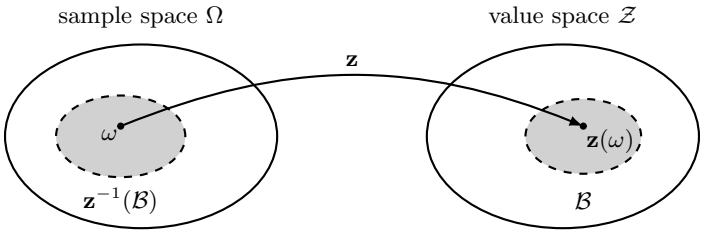


Abbildung 1: Drei Datensätze von Realisierungen von u.i.v. Zufallsvariablen, die aus einer gemeinsamen Wahrscheinlichkeitsverteilung $\mathbb{P}$ gezogen wurden, und die drei Ausgaben derselben ML-Methode $\mathcal{A}$. Diese drei Ausgaben sind Realisierungen einer anderen ZV $\widehat{h}$; die Wahrscheinlichkeitsverteilung von $\widehat{h}$ hängt von $\mathbb{P}$ und der ML-Methode ab. Wie stark die drei Ausgaben voneinander abweichen, wird daher durch $\mathbb{P}$ und die ML-Methode bestimmt

Formal ist eine ZV eine measurable Funktion $\mathbf{z} : \Omega \to \mathcal{Z}$ von einem probability space $(\Omega, \Sigma, \mathbb{P})$ zu einem measurable Raum $\mathcal{Z}$, der Menge der Werte, die $\mathbf{z}$ annehmen kann. Die Wahrscheinlichkeitsverteilung von $\mathbf{z}$ ist das probability measure $\mathbb{P}^{(\mathbf{z})}$, das $\mathbf{z}$ auf $\mathcal{Z}$ induziert. Dieses probability measure weist jeder measurable Teilmenge $\mathcal{B} \subseteq \mathcal{Z}$ die Wahrscheinlichkeit

$$\mathbb{P}^{(\mathbf{z})}(\mathcal{B}) := \mathbb{P}\left(\mathbf{z} \in \mathcal{B}\right) = \mathbb{P}\big(\{\omega \in \Omega : \mathbf{z}(\omega) \in \mathcal{B}\}\big),$$

die Wahrscheinlichkeit des Ereignis, dass $\mathbf{z}$ einen Wert in $\mathcal{B}$ annimmt, zu (siehe Abb. 2; Billingsley, 1986, Sect. 13, Eq. (13.7), and Sect. 14, Eq. (14.1); Gray, 2009; Kallenberg, 1997). Dieses Ereignis ist das preimage $\mathbf{z}^{-1}(\mathcal{B}) = \{\omega \in \Omega : \mathbf{z}(\omega) \in \mathcal{B}\}$, die Menge der Ergebnisse ω des sample space Ω, die $\mathbf{z}$ in $\mathcal{B}$ abbildet. Die Messbarkeit von $\mathbf{z}$ garantiert, dass $\mathbf{z}^{-1}(\mathcal{B})$ zur σ-Algebra Σ gehört, so dass $\mathbb{P}\ \mathbf{z}^{-1}(\mathcal{B})$ eine Wahrscheinlichkeit zuweist. Der hochgestellte Index von $\mathbb{P}^{(\mathbf{z})}$ benennt die ZV, deren Verteilung gemeint ist, wie in $\mathbb{P}^{(\mathbf{x},y)}$ für einen Datenpunkt mit Merkmale $\mathbf{x}$ und Label y. Wenn nur eine ZV im Spiel ist, wird der hochgestellte Index weggelassen, und $\mathbf{z} \sim \mathbb{P}$ besagt, dass $\mathbb{P}$ die Wahrscheinlichkeitsverteilung von $\mathbf{z}$ ist.



$$\mathbb{P}^{(\mathbf{z})}(\mathcal{B}) = \mathbb{P}(\mathbf{z} \in \mathcal{B}) = \mathbb{P}\big(\mathbf{z}^{-1}(\mathcal{B})\big)$$

Abbildung 2: Eine ZV $\mathbf{z}$ bildet jedes Ergebnis ω des sample space Ω auf einen Wert $\mathbf{z}(\omega)$ im Wertebereich $\mathcal{Z}$ ab. Ihre Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{z})}$ weist einer measurable Menge $\mathcal{B} \subseteq \mathcal{Z}$ (rechts, schattiert) die Wahrscheinlichkeit zu, die $\mathbb{P}$ dem preimage $\mathbf{z}^{-1}(\mathcal{B}) \subseteq \Omega$ (links, schattiert) zuweist, dem Ereignis, dass $\mathbf{z}$ einen Wert in $\mathcal{B}$ annimmt

Unter der u.i.v.-Annahme bestimmt die Wahrscheinlichkeitsverteilung $\mathbb{P}$ der Datenpunkte, wie wahrscheinlich jeder Datensatz ist: Ein Datensatz von Realisierungen von u.i.v. Zufallsvariablen ist wahrscheinlich, wenn seine Datenpunkte dort liegen, wo $\mathbb{P}$ eine große Wahrscheinlichkeit zuweist. Diese Verteilung ist im Allgemeinen unbekannt; nur der Datensatz wird beobachtet. In der Praxis wird $\mathbb{P}$ basierend auf Domänenwissen postuliert oder aus einem Datensatz geschätzt, z.B. durch seine empirische Verteilung oder sein histogram, oder durch maximum likelihood innerhalb einer parametrischen Familie. Eine parametrische Familie ist eine Menge von Wahrscheinlichkeitsverteilungen, die durch einige Zahlen indiziert sind, wie die Normalverteilungen, die durch Mittelwert und Varianz indiziert sind.

Eine probability mass function (pmf) oder eine Wahrscheinlichkeitsdichte ist eine kompakte Art, eine Wahrscheinlichkeitsverteilung zu spezifizieren. Die Wahrscheinlichkeitsverteilung einer binären ZV $y \in \{0, 1\}$ wird vollständig durch die einzelne Wahrscheinlichkeit $\mathbb{P}(y = 0)$ spezifiziert, da $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = 0)$; allgemeiner wird die Wahrscheinlichkeitsverteilung einer diskrete Zufallsvariable durch ihre pmf spezifiziert. Die Wahrscheinlichkeitsverteilung einer reellwertigen ZV $x \in \mathbb{R}$, die eine Wahrscheinlichkeitsdichte $p^{(x)}(\cdot)$ besitzt, wird durch diese Wahrscheinlichkeitsdichte spezifiziert. Die Wahrscheinlichkeit eines Wertes in einem Intervall $[a, b]$ ist das Integral

$$\mathbb{P}(x \in [a, b]) = \int_a^b p^{(x)}(x) \, \mathrm{d}x,$$

welches für ein kurzes Intervall $[a, b]$ ungefähr $p^{(x)}(a) \, |b - a|$ ist.

Da die Wahrscheinlichkeitsdichte die Wahrscheinlichkeit eines kurzen Intervalls um jeden Wert angibt, kann sie gezeichnet werden. Für Datenpunkte $\mathbf{z} = (x, y)$ mit einem einzelnen numerischen Merkmal x und einem numerischen Label y kann eine Wahrscheinlichkeitsverteilung in der Ebene $\mathbb{R}^2$ visualisiert

werden. Abb. 3 zeigt eine solche als Graustufenbild über der Merkmal–Label-Ebene: Je dunkler eine Region, desto größer ist ihre Wahrscheinlichkeitsdichte, d.h. desto wahrscheinlicher fällt ein aus $\mathbb{P}$ gezogener Datenpunkt in sie. Die dargestellte Wahrscheinlichkeitsdichte ist eine gewichtete Summe zweier Gaußscher Wahrscheinlichkeitsdichten, d.h. ein Gaußsches Mischmodell (GMM). Parallel zu Abb. 1 werden drei Datensätze $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \mathcal{D}^{(3)}$, die jeweils aus $m = 20$ Realisierungen von u.i.v. Zufallsvariablen mit dieser Verteilung bestehen, zusammen mit den drei Hypothesen $\widehat{h}^{(1)}, \widehat{h}^{(2)}, \widehat{h}^{(3)}$ gezeigt, die von derselben polynomial regression-Methode aus ihnen gelernt wurden. Wie in Abb. 3 gezeigt, unterscheiden sich die drei gelernten Hypothesen am stärksten dort, wo die Wahrscheinlichkeitsdichte klein ist.

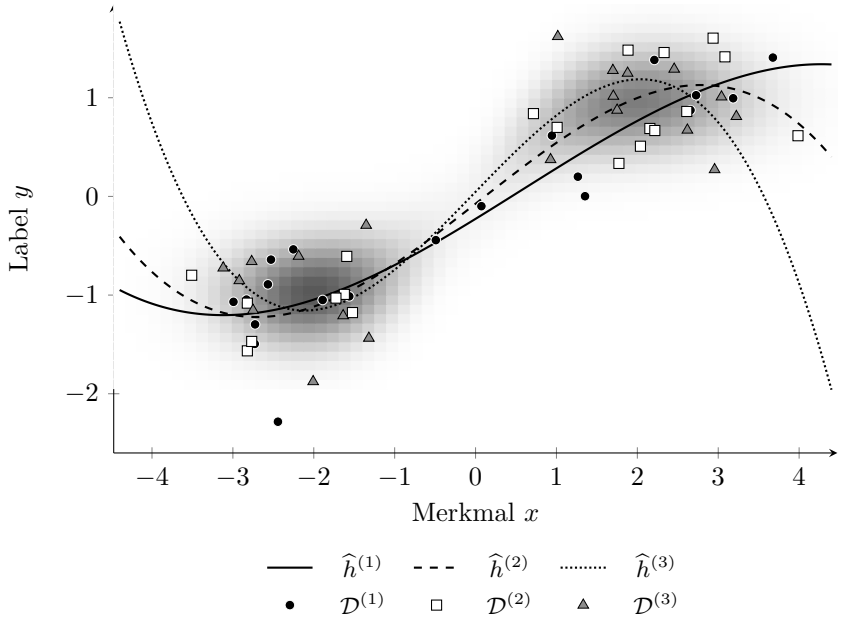


Abbildung 3: Eine Wahrscheinlichkeitsverteilung über Datenpunkte $\mathbf{z} = (x, y)$, als Graustufenbild gezeichnet: Je dunkler eine Region, desto größer ist ihre Wahrscheinlichkeitsdichte. Die Wahrscheinlichkeitsdichte ist ein GMM mit zwei Komponenten. Parallel zu Abb. 1 ergeben drei Datensätze $\mathcal{D}^{(1)}$ (gefüllte Kreise), $\mathcal{D}^{(2)}$ (offene Quadrate) und $\mathcal{D}^{(3)}$ (gefüllte Dreiecke), die jeweils aus $m = 20$ Realisierungen von u.i.v. Zufallsvariablen mit dieser Verteilung bestehen, die drei Hypothesen $\widehat{h}^{(1)}$ (durchgezogene Kurve), $\widehat{h}^{(2)}$ (gestrichelte Kurve) und $\widehat{h}^{(3)}$ (gepunktete Kurve), die von derselben polynomial regression-Methode gelernt wurden. Die Hypothesen unterscheiden sich am stärksten dort, wo die Wahrscheinlichkeitsdichte klein ist. Data generated by `pythondemos/probdist.py`

Das Beispiel veranschaulicht auch, was empirische Risikominimierung (ERM) aus einem Datensatz bestimmen kann. Die ERM wählt die Hypothese

anhand ihres durchschnittlichen Verlust über den Datensatz. Sie gewichtet daher eine Region, in der der Datensatz wenige oder keine Datenpunkte enthält, gering. Folglich unterscheidet die ERM nicht zwischen Hypothesen, die sich nur dort unterscheiden und nahezu denselben Trainingsfehler aufweisen.

Synonyme: Verteilung, Gesetz.

Siehe auch: u.i.v., u.i.v.-Annahme, Realisierung, ZV, Wahrscheinlichkeit, measurable, Wahrscheinlichkeitsdichte, pmf, diskrete Zufallsvariable, probability measure, probability space, empirische Verteilung, Erwartungswert, Generalisierung.

Querverweise

Literatur

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

2. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.

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