Imputation

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 Imputation ersetzt fehlende Attributwerte von Datenpunkte durch Schätzungen, die aus den aufgezeichneten Attributwerten berechnet werden. In tabellarische Daten ist ein fehlender Wert eine leere Zelle der Tabelle, und die Imputation füllt die Zelle. Sie ist eine Reaktion auf missing data, neben dem Verwerfen der unvollständigen Datenpunkte. Das Füllen ist ein machine learning (ML)-Problem: Der fehlende Wert ist das Label eines Datenpunkt, die daneben aufgezeichneten Werte sind seine Merkmale, und die vollständigen Datenpunkte bilden einen Trainingsdatensatz, auf dem eine Hypothese-Abbildung (ein Imputer) gelernt wird.

Definition

Die Wetterstation Krems an der Donau zeichnet für jeden Tag die minimale Temperatur, den Niederschlag und die maximale Temperatur auf. Das Jahr 2024 ist daher eine Tabelle mit einer Zeile pro Tag und einer Spalte pro Größe (GeoSphere Austria, 2026). Diese Tabelle ist tabellarische Daten. Ein Tag, an dem das Thermometer ausfiel, lässt die Zelle für die maximale Temperatur seiner Zeile leer, während die Tabelle ihre Form behält (Abb. 1). Die Imputation füllt eine solche Zelle: Sie ersetzt fehlende Attributwerte von Datenpunkte, Merkmale oder Labels durch Schätzungen, die aus den aufgezeichneten Attributwerten berechnet wurden (Little and Rubin, 2019). Sie ist eine Reaktion auf missing data, neben dem Verwerfen der unvollständigen Datenpunkte oder der Verwendung einer Methode, die die Lücken toleriert (Little and Rubin, 2019).

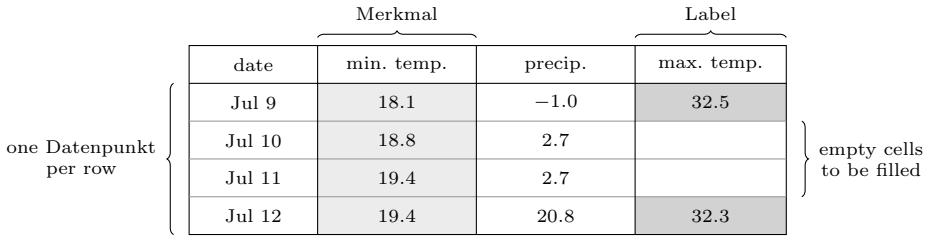

Abbildung 1: Vier der 366 Zeilen der Tabelle der täglichen Wetteraufzeichnungen aus dem Jahr 2024, 9.–12. Juli. Jede Zeile ist ein Datenpunkt, nämlich ein Tag; die Kopfzeilen kürzen die minimale Temperatur (Min.-Temp.), den Niederschlag (Niederschl.) und die maximale Temperatur (Max.-Temp.) ab, und der spezielle Wert -1.0 der Niederschlagsspalte zeigt an, dass kein Niederschlag beobachtet wurde. Die Zellen für die maximale Temperatur des 10. und 11. Juli sind leer. Die hellgraue Spalte enthält das Merkmal und die dunkelgraue Spalte das Label eines Tages, sodass die beiden gezeigten vollständigen Zeilen und die anderen vollständigen Tage beides liefern, während die beiden leeren Zellen die zu imputierenden Labels sind

Das Füllen ist ein machine learning (ML)-Problem (Abayomi et al., 2008). Die leere Zelle ist das Label ihrer Zeile, und die aufgezeichneten Zellen derselben Zeile sind ihre Merkmale. Die vollständigen Zeilen bilden einen Trainingsdatensatz, da jede sowohl einen Merkmalsvektor als auch das dazugehörige Label liefert. Eine Hypothese-Abbildung, ein Imputer, wird aus ihnen mittels empirische Risikominimierung (ERM) gelernt. Das Anwenden dieser Hypothese auf eine unvollständige Zeile liefert die Schätzung, die ihre Zelle füllt. Die einfachste Imputation füllt jede leere Zelle einer Spalte mit dem Stichprobenmittel der aufgezeichneten Zellen dieser Spalte, was die Hypothese ist, die den Merkmalsvektor ignoriert (Little and Rubin, 2019).

Ein numerisches Experiment nimmt 15 über das Jahr verteilte Tage und leert die Zellen für die maximale Temperatur von 3 davon. Aus den 12 vollständigen Zeilen lernt es eine lineare Regression-Hypothese, die die minimale Temperatur eines Tages auf seine maximale Temperatur abbildet. Die geleerten Zellen werden dann mit ihren Vorhersagen gefüllt, dargestellt als scatterplot in Abb. 2. Die entfernten Werte sind bekannt, sodass das Füllen bewertet werden kann. Der durchschnittliche squared error loss auf den 3 Zellen, in Quadratgrad Celsius, beträgt 17.6 für die gelernte Hypothese und 150.2 für den Spaltenmittelwert.

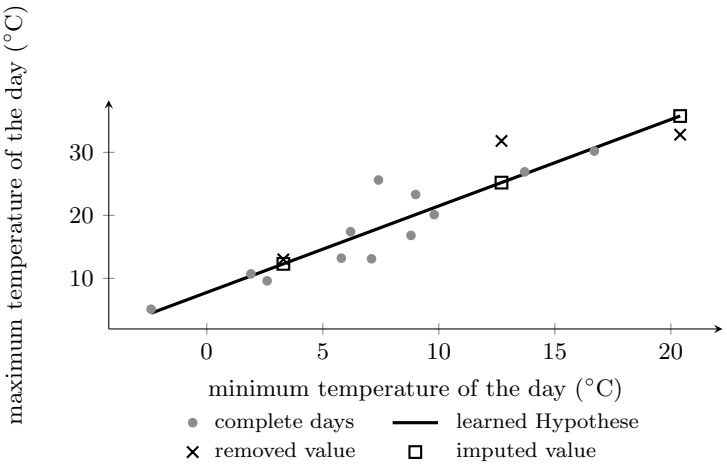

Abbildung 2: Scatterplot von 15 Tagen der Tabelle in Abb. 1, über das Jahr verteilt, mit der minimalen Temperatur eines Tages als Merkmal und der maximalen Temperatur als Label. Die 12 vollständigen Tage sind graue Kreise, und die durchgezogene Linie ist die lineare Regression-Hypothese, die aus ihnen gelernt wurde. Für jede der 3 geleerten Zellen markiert ein Kreuz den entfernten Wert und ein offenes Quadrat den imputierten Wert

Ein fehlender Wert kann sich unter anderen fehlenden Werten befinden, sodass der Merkmalsvektor seiner Zeile selbst unvollständig ist. Rechteckige Bereiche, die aus einem Luftbild verloren gegangen sind, sind ein solcher Fall. Jeder Pixel ist ein Datenpunkt, seine Merkmale sind die Farben der acht umgebenden Pixel, und sein Label ist seine eigene Farbe, sodass ein verlorener Pixel innerhalb eines Bereichs auch Nachbarn verloren hat. Das Füllen erfolgt dann in Iterationen: Die fehlenden Werte beginnen beim Stichprobenmittel der aufgezeichneten Werte, und jede Iteration schätzt sie aus den Werten neu, die ihre Nachbarn aktuell haben, bis sich die Werte nicht mehr ändern. Für ein lineares Modell bildet eine Iteration den Vektor $\mathbf{x}$ der fehlenden Werte auf $\mathbf{M}\mathbf{x}+\mathbf{b}$ ab, wobei eine Matrix $\mathbf{M}$ aus den Gewichte der Hypothese gebildet wird. Zwei Läufe, die bei $\mathbf{x}$ und $\mathbf{x}'$ gestartet wurden, unterscheiden sich nach t Iterationen um $\mathbf{M}^t(\mathbf{x}-\mathbf{x}')$, sodass die Iterationen von jedem Start aus konvergieren, genau dann, wenn $\mathbf{M}^t \to \mathbf{0}$, d.h. genau dann, wenn der spectral radius $\rho(\mathbf{M})$ kleiner als eins ist (Horn and Johnson, 2013, Thm. 5.6.12), und der Grenzwert ist dann der eindeutige Fixpunkt der Iteration. Ein numerisches Experiment füllt die verlorenen Pixel durch solche Iterationen. Für die Hypothese, die aus den intakten Pixeln mittels der Methode der kleinsten Quadrate gelernt wurde, ist $\rho(\mathbf{M}) = 2.72$ und die Iterationen divergieren. Der Strafterm der Ridge-Regression schrumpft die Gewichte: Mit dem Regularisierung-Parameter $\alpha = 0.03$ ergibt die durch Ridge-Regression gelernte Hypothese $\rho(\mathbf{M}) = 0.84$, und die Iterationen konvergieren. Mit dieser Hypothese ist der durchschnittliche squared error loss auf den

verlorenen Pixeln halb so groß wie der der Baseline, die jeden von ihnen mit der durchschnittlichen Farbe füllt, und die Grenzen, an denen sich die Farben stark ändern, erscheinen unscharf.

Ein imputierter Wert ist eine Schätzung und keine Messung, und eine Methode, die auf imputierten Datenpunkte trainiert wurde, kann die beiden nicht unterscheiden. Ein Teil des Trainingsdatensatz wurde durch eine Hypothese erzeugt und nicht beobachtet, sodass ein Trainingsfehler, der über imputierte Zellen mittelt, eine Hypothese teilweise gegen die Vorhersagen des Imputers bewertet und nicht gegen Messungen; für den Imputer selbst trägt jede gefüllte Zelle null Verlust bei. Ein einzelner gefüllter Wert enthält auch keine Informationen über die uncertainty der Schätzung (Hastie et al., 2009, Sect. 9.6; Little and Rubin, 2019). Das Aufzeichnen, welche Werte imputiert wurden, und das Heraushalten dieser aus dem Testdatensatz, hält die Schätzung des Risiko frei von Werten, die eine Hypothese erzeugt hat.

Synonyme: Imputation, Imputation fehlender Werte.

Siehe auch: missing data, tabellarische Daten, Datenpunkt, Merkmal, Label, Merkmalsvektor, Trainingsdatensatz, ERM, lineare Regression, squared error loss, Baseline, Testdatensatz.

Querverweise

Literatur

1. Abayomi et al. (2008). Diagnostics for Multivariate Imputations. J. Roy. Statist. Soc.: Ser. C (Appl. Statist.), 57(3), pp. 273–291.

2. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/hist v2-1d.

3. Horn and Johnson (2013). Matrix Analysis. 2nd ed., Cambridge Univ. Press, New York, NY, USA.

4. Little and Rubin (2019). Statistical Analysis with Missing Data. 3rd ed., Wiley, Hoboken, NJ.

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