mittlerer quadratischer Fehler (MSE)

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 mittlere quadratische Fehler (engl. *mean squared error*, MSE) einer Hypothese ist der Durchschnitt ihres quadratischer Verlusts, also der quadrierten Differenz zwischen dem Label eines Datenpunkts und der zugehörigen Vorhersage. Der Durchschnitt ist entweder ein Erwartungswert unter einem Wahrscheinlichkeitsmodell für die Daten, der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme), wodurch der MSE zum Risiko unter dem quadratischer Verlust wird, oder der Durchschnitt über die Datenpunkte eines Datensatzs, wodurch er zum empirisches Risiko, zum Trainingsfehler oder zum Validierungsfehler wird. Die beiden Lesarten fallen für die empirische Verteilung des Datensatzs zusammen, die jedem seiner Datenpunkte dieselbe Wahrscheinlichkeit zuweist. Wenn die u.i.v.-Annahme gilt, schätzt der Durchschnitt über einen Datensatz den Erwartungswert und verbessert sich mit dem Stichprobengröße. Die Wurzel des MSE (engl. *root MSE*), seine Quadratwurzel, wird in der Einheit des Labels gemessen.

Definition

Betrachtet sei eine Hypothese h, die die Temperatur um 15:00 Uhr (UTC) eines Tages in Krems an der Donau (Österreich) aus der Temperatur um 06:00 Uhr (UTC) desselben Tages vorhersagt. Ein Wetterdienst, der sie nutzt, gibt eine einzige Zahl dafür an, wie weit die Vorhersagen im Durchschnitt danebenliegen, und der MSE ist diese Zahl: der Durchschnitt des quadratischer Verlusts, also der quadrierten Differenz zwischen dem Label eines Datenpunkts und der zugehörigen Vorhersage. Das Wort *Durchschnitt* hat hier zwei Lesarten, und der MSE ist in beiden definiert.

Die erste Lesart benötigt ein Wahrscheinlichkeitsmodell dafür, wie die Daten zustande kommen. Die Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) ist das einfachste: Sie behandelt die täglichen Wettermessungen als Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV), sodass jeder Tag, ein Datenpunkt mit Merkmalsvektor $\mathbf{x}$ und Label y, eine unabhängige Ziehung aus einer Wahrscheinlichkeitsverteilung $\mathbb{P}$ ist, die sich von Tag zu Tag nicht ändert. Unter diesem Modell ist der MSE von h der Erwartungswert ihres quadratischer Verlusts,

$$\bar{L}\big(h\big) := \mathbb{E}\big\{\big(y - h(\mathbf{x})\big)^2\big\}, \tag{1}$$

was dem Risiko von h unter dem quadratischer Verlust entspricht (Hastie et al., 2009, Abschn. 2.4). Der Erwartungswert (1) wird über die Wahrscheinlichkeitsverteilung $\mathbb{P}$ des Datenpunkts gebildet, welche die u.i.v.-Annahme zwar postuliert, aber nicht näher bestimmt: In einer maschinelles Lernen (ML)-Anwendung ist $\mathbb{P}$ unbekannt. Berechnet werden kann der MSE auf einem Datensatz $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$ vom Stichprobengröße m, der Durchschnitt des quadratischer Verlusts über dessen Datenpunkte,

$$\widehat{L}\big(h|\mathcal{D}\big) := \frac{1}{m}\sum_{r=1}^{m}\big(y^{(r)} - h(\mathbf{x}^{(r)})\big)^2, \tag{2}$$

was das empirisches Risiko von h auf $\mathcal{D}$ unter dem quadratischer Verlust ist. Er ist der Trainingsfehler, wenn $\mathcal{D}$ der Trainingsdatensatz ist, und der Validierungsfehler, wenn $\mathcal{D}$ das Validierungsdatensatz ist. Für ein lineares Modell, das mittels der Methode der kleinsten Quadrate auf den 364 Tagen des Jahres 2023 in Krems gelernt wurde, für die beide Messwerte vorliegen, beträgt der MSE (2) an diesen Tagen 18.0 und an den 365 entsprechenden Tagen des Jahres 2024 15.3, gemessen in Quadratgrad. Seine Quadratwurzel, die Wurzel des MSE von 3.9 Grad, hat die Einheit des Labels.

Die beiden Definitionen sind ein und dieselbe Definition, angewandt auf zwei Wahrscheinlichkeitsverteilungen. Die empirische Verteilung $\mathbb{P}^{(\mathcal{D})}$ des Datensatzs $\mathcal{D}$ weist jedem seiner Datenpunkte die Wahrscheinlichkeit $1/m$ zu. Der Erwartungswert des quadratischer Verlusts unter $\mathbb{P}^{(\mathcal{D})}$ ist daher eine Summe von m Termen mit dem Gewicht $1/m$ für jeden Term,

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}^{(\mathcal{D})}}\big\{\big(y - h(\mathbf{x})\big)^2\big\} = \widehat{L}\big(h|\mathcal{D}\big). \tag{3}$$

Gemäß der Identität (3) ist der Durchschnitt (2) der Erwartungswert (1) mit der empirische Verteilung von $\mathcal{D}$ anstelle von $\mathbb{P}$. Jede Aussage über den MSE als Erwartungswert gilt dann auch für den Durchschnitt mit $\mathbb{P}^{(\mathcal{D})}$ in dieser Rolle, wie etwa seine Zerlegung in die Varianz des Labels um dessen bedingten Mittelwert und die quadrierte Abweichung von h von diesem Mittelwert (Bishop, 2006, Abschn. 1.5.5). Abb. 1(a) zeigt die empirische Verteilung der quadrierten Fehler der 365 Tage des Jahres 2024. Ihr Mittelwert ist der MSE von 15.3, während ihr Median bei 8.3 liegt und zwei Drittel der Tage einen Verlust unterhalb des MSE aufweisen: Einige wenige stark abweichende Tage fallen in einem quadrierten Durchschnitt stark ins Gewicht. Wenn umgekehrt die Datenpunkte von $\mathcal{D}$ Realisierungen von u.i.v. Zufallsvariablen mit der Wahrscheinlichkeitsverteilung $\mathbb{P}$ sind, ist der Durchschnitt (2) ein Schätzwert für den Erwartungswert (1), der sich nach dem Gesetz der großen Zahlen mit dem Stichprobengröße m verbessert. Abb. 1(b) zeigt dies für die Tage des Jahres 2024 in zufälliger Reihenfolge: Über die ersten 30 Tage liegt der Durchschnitt in neun von zehn Reihenfolgen zwischen 10.2 und 21.1, über die ersten 180 Tage zwischen 13.8 und 16.8. Wie weit der Trainingsfehler vom Risiko abweichen kann, ist Gegenstand der Generalisierung.

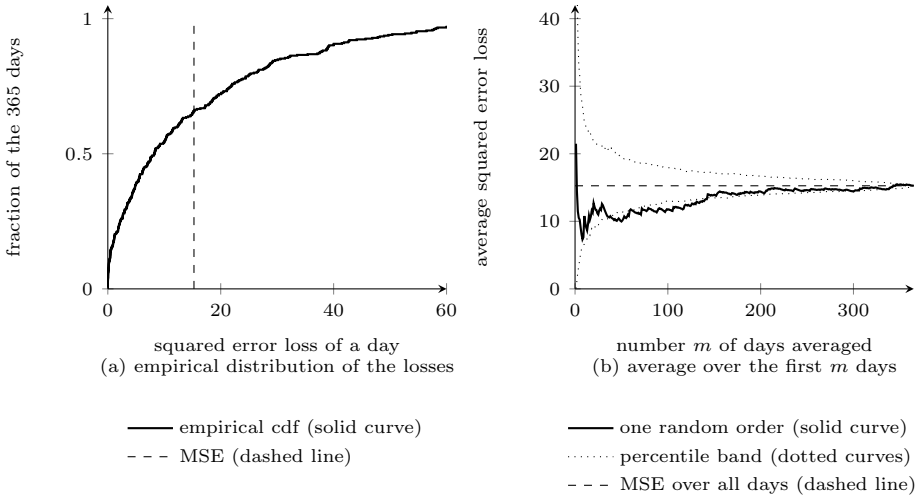


Abbildung 1: Der quadratische Fehlerverlust eines lineares Modells, gelernt mittels der Methode der kleinsten Quadrate auf den Tagen des Jahres 2023 in Krems, ausgewertet auf den 365 Tagen des Jahres 2024. (a) Die empirische Verteilungsfunktion der 365 Verluste (durchgezogene Kurve), abgeschnitten bei 60 Quadratgrad mit 12 Tagen jenseits dieser Grenze, sowie ihr Mittelwert, der MSE (2) (gestrichelte Linie). (b) Der Durchschnitt der Verluste über die ersten m Tage des Jahres 2024 in einer zufälligen Reihenfolge (durchgezogene Kurve), das Band zwischen dem 5. und dem 95. Perzentil über 500 Reihenfolgen, das diesen Durchschnitt für neun von zehn Reihenfolgen enthält (gepunktete Kurven), sowie der MSE über alle 365 Tage (gestrichelte Linie). Daten erzeugt mit `pythondemos/mse.py`

Das Quadrieren der Abweichung führt dazu, dass ein stark abweichender Tag schwer ins Gewicht fällt, und genau darin liegt die Hauptschwäche des MSE. Ein einziges falsch gemessenes Label geht proportional zum Quadrat seiner Abweichung in den Durchschnitt ein, sodass es den für ein ganzes Jahr ausgewiesenen Wert dominieren kann. Der mittlerer absoluter Fehler (MAE), also der Durchschnitt des absoluter Verlusts anstelle des quadratischer Verlusts, gewichtet jeden Tag lediglich proportional zu seiner einfachen Abweichung. Der Unterschied zeigt sich darin, welchen Wert jeder Durchschnitt von einer konstanten Vorhersage verlangt: Der MSE ist am kleinsten beim Mittelwert der Labels, der MAE bei deren Median (Hastie et al., 2009, Abschn. 10.6). An den 365 Tagen des Jahres 2024 in Krems liegen diese Werte bei 16.30 und 15.80 Grad. Ein Kommafehler verschiebt eines dieser 365 Labels um 5, 10, 20 oder 50 Grad. Der Mittelwert verschiebt sich dann um 0.014, 0.027, 0.055 bzw. 0.137 Grad, also um den Fehler geteilt durch den Stichprobengröße. Der Median verschiebt sich um höchstens 0.1 Grad, den Schritt zum benachbarten Label, ganz gleich, wie groß der Fehler wird. Der MSE ist daher die falsche Zusammenfassung für einen Datensatz mit Ausreißern oder einer Fehlerverteilung mit schweren Rändern, und der MAE

oder eine Kleinste-Absolute-Abweichungs-Regression ist hier die übliche Wahl (Hastie et al., 2009, Abschn. 10.6). Was der MSE im Gegenzug bietet, ist der Mittelwert, auf den er abzielt: Er ist überall differenzierbar, was der Methode der kleinsten Quadrate überhaupt erst eine geschlossene Form verleiht. Siehe auch: quadratischer Verlust, Risiko, empirisches Risiko, Trainingsfehler, Validierungsfehler, empirische Verteilung, Erwartungswert, Wahrscheinlichkeitsmodell, u.i.v.-Annahme, Gesetz der großen Zahlen, Generalisierung, MAE, absoluter Verlust, Ausreißer, lineare Regression, Verlust.

Querverweise

Literatur

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

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