outlier

Author and editors

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

Abstract

An outlier is a data point that deviates so far from the rest of a dataset that it is suspected of having come about in a different way. Being an outlier is a judgment made against a stated probabilistic model or dataset and not a property a data point carries on its own: a day is typical for the year and extreme for its month, and two features can deviate jointly while neither deviates alone. Two sources have to be told apart, a probability distribution with a heavy tail and a mistaken or coded reading, because only the second calls for a repair of the data. With more than two features, dimensionality reduction to two learned features makes the deviating data points visible in a scatterplot, which is why it is a common first step. How much an outlier changes the learned hypothesis depends on the loss function: the squared error loss weighs a residual by its square and is far more sensitive to a mismeasured label than the absolute error loss. Least absolute deviation regression, Huber regression, and robust principal component analysis (robust PCA) are built on that difference, and a contamination model states how outliers enter the sample.

Definition

A weather station at Krems an der Donau (Austria) recorded 81.7 millimeters of rain on 14 September 2024, more than forty times the average daily amount of 1.9 millimeters over that year. The same archive writes a day with no precipitation not as 0 but as -1 millimeters, a code that 199 of the 366 days carry. The two numbers are the largest and the smallest value of their column, and only one of them marks a day worth a second look. An outlier is a data point that deviates so far from the rest of a dataset that it is suspected of having come about in a different way (Hawkins, 1980). What counts as far is set by a stated notion of ordinary behavior, so being an outlier is a judgment made against a probabilistic model or a dataset and never a property the data point carries on its own (Chandola et al., 2009, Sect. 1.1).

Changing the probabilistic model changes the answer while the data stay as they are. Consider the screen that calls a day unusual when its maximum daytime temperature lies more than two standard deviations from the mean. Under a probabilistic model in which that temperature is a Gaussian random variable (Gaussian RV) with mean 17.9 and standard deviation 9.2 degrees,

estimated from all 366 days of 2024, the screen flags three days. With the mean and the standard deviation estimated from the day's calendar month alone, it flags ten days, and only 31 December under both. On 17 May the maximum reached 16.6 degrees, which lies 0.14 standard deviations below the mean of the year and 2.60 below the mean of May: typical for the year, extreme for its month. The coldest maximum of 2024, the −3.6 degrees of 9 January, runs the other way, 2.34 standard deviations below the mean of the year and 1.94 below the mean of January, an ordinary cold winter day. Fig. 1 draws both screens over the year.

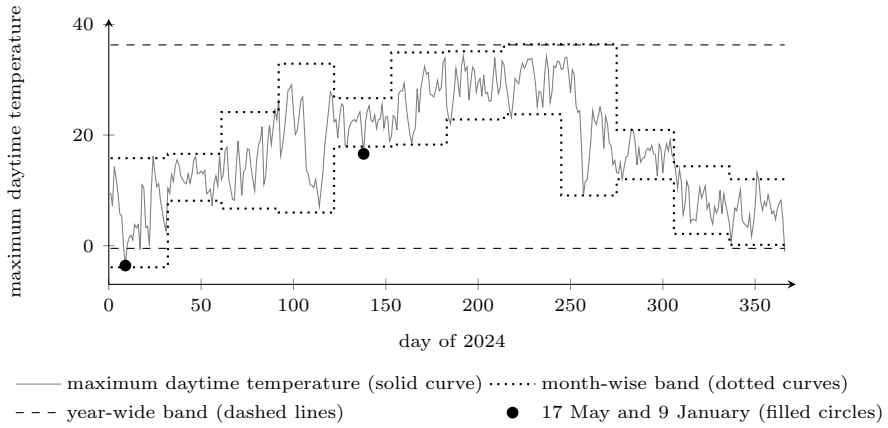


Figure 1: The maximum daytime temperature of the 366 days of 2024 at Krems an der Donau (solid curve), with two screens that call a day unusual when the temperature lies more than two standard deviations from the mean: the band of the probabilistic model estimated from all 366 days (dashed lines), and the band of the probabilistic model estimated from the day's calendar month (dotted curves). The filled circles mark 17 May, inside the year-wide band and below the month-wise one, and 9 January, below the year-wide band and inside the month-wise one. Data generated by `pythondemos/outlier.py`

A deviation can also sit in the relation between two features rather than in either one. Each day is a data point with the feature vector $\mathbf{x} = (x_1, x_2)^\top$ holding its morning minimum and its maximum daytime temperature. Over 2024, the two have correlation 0.88, so a probabilistic model that gives $\mathbf{x}$ a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$ expects them to rise and fall together. It scores a day by the squared Mahalanobis distance (Bishop, 2006, Sect. 2.3)

$$\big(\mathbf{x} - \boldsymbol{\mu}\big)^\top \mathbf{C}^{-1} \big(\mathbf{x} - \boldsymbol{\mu}\big), \tag{1}$$

of which the probability density function (pdf) of the multivariate normal distribution is a decreasing function, so a large squared distance is a small likelihood

function. The largest value of the squared Mahalanobis distance (1) over 2024, 6.7, belongs to 29 March: a morning minimum of 0.5 and a maximum of 21.2 degrees, 0.91 and 0.36 standard deviations from the respective means. Neither feature is remarkable on its own, and jointly they are the daily range of 20.7 degrees that is the fourth largest of the year.

The weather station records eight measurements on each day, and eight numbers are more than a scatterplot can show. Dimensionality reduction to two features is therefore a common first step in finding outliers, because two numbers per data point are what a scatterplot does show. Principal component analysis (PCA) learns a linear map from the eight standardized measurements of a day to $d' = 2$ features, which keep 63.6 percent of their variance; the first is a temperature direction and the second a wetness direction. In those two features, the deviating days are visible without any screen at all, as Fig. 2 shows. Evaluating the squared Mahalanobis distance (1) in those two features rather than in the two temperatures puts 14 September at 50.7, against 22.9 for the second farthest day and below 8.3 for each of the remaining 363 days.

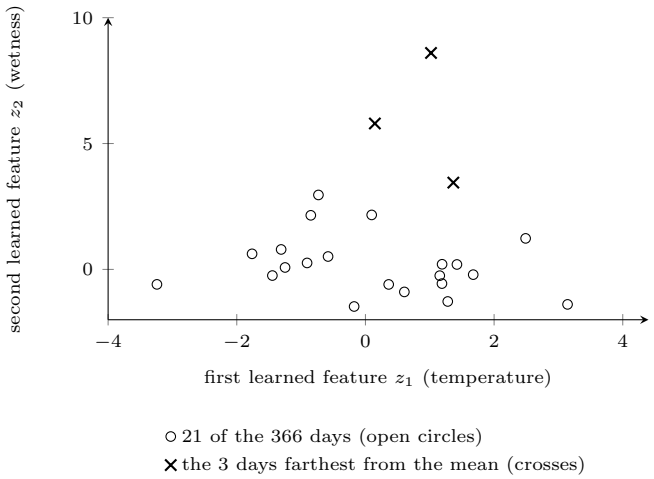


Figure 2: Every 18th of the 366 days of 2024 at Krems, 21 days, placed at the two features that PCA learns from the eight standardized measurements of a day (open circles), together with the 3 days farthest from the mean in those two features (crosses). The farthest of the three is 14 September, the day of 81.7 millimeters of rain. Data generated by `pythondemos/outlier.py`

The code −1 of the opening is of a different kind. The archive writes a day with no precipitation as −1 millimeters and a day with less than 0.1 millimeters as 0, so −1 is not an amount at all. A plausibility screen that rejects a negative rainfall flags 199 of the 366 days, 54 percent of the year, and not one of those days is unusual; read as an amount, the code lowers the annual total from 712.0 to 513.0 millimeters. The deviation is from the assumption about the coding of

the tabular data, not from the process that produced the data, and the repair is to recode the column rather than to drop the days. The two kinds have to be told apart before anything is removed, since an outlier arises either from a probability distribution with a heavy tail or from a mistaken reading (Bishop, 2006, Sect. 2.3.7), and a coding convention is neither.

Removal is not the only response. How much an outlier changes the learned hypothesis depends on the loss function: the squared error loss weighs a residual by its square and the absolute error loss weighs it linearly, which makes least squares far more sensitive to a grossly mismeasured label than empirical risk minimization (ERM) with the absolute error loss (Hastie et al., 2009, Sect. 10.6). On the Krems record, least squares over the linear model learns a hypothesis that predicts the maximum daytime temperature of a day from its morning minimum. The maximum of 21 January 2024 is 3.5 degrees. A misplaced decimal point that records it as 35 moves the predictions of that hypothesis by up to 0.56 degrees and leaves the one learned with the absolute error loss where it was. Raising the same label to 1000 moves the least squares predictions by 17.7 degrees, and the other hypothesis still does not move at all. How far a single replaced data point can move the learned hypothesis is a question about the stability of the method. Least absolute deviation regression, Huber regression with the Huber loss, and robust principal component analysis (robust PCA) for a feature matrix are built on this difference, and a contamination model states how outliers enter the sample and what can then be guaranteed.

Cross References

anomaly detection, contamination model, robustness, stability, Huber regression, least absolute deviation regression, robust PCA, absolute error loss, squared error loss, probabilistic model, dimensionality reduction, PCA

References

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

2. Chandola et al. (2009). Anomaly Detection: A Survey. ACM Comput. Surv., 41(3), pp. 1–58.

3. Hawkins (1980). Identification of Outliers. Chapman and Hall, London.

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