

feature matrix

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Oliumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

The feature matrix of a dataset collects the feature vectors of its data points as its rows, so it has one row per data point and one column per feature. Reading it along a row returns one data point, and reading it along a column returns one feature across the whole dataset. A few numbers computed from it summarize the dataset. Its shape is the pair of the number of data points and the number of features, and linear regression has a unique solution only when the feature matrix has full column rank. Its singular values are at most as many numbers as its smaller dimension, the count of the nonzero ones is its rank, and their squares are the eigenvalues of $\mathbf{X}^\top \mathbf{X}$, which divided by the number of data points is the sample covariance matrix. The condition number, the ratio of the largest singular value to the smallest nonzero one, bounds how far the linear regression solution moves when the labels are perturbed.

Definition

A weather station at Krems an der Donau records eight measurements on each day of a year: the lowest, highest and mean air temperature, the precipitation, the sunshine duration, the humidity, the air pressure and the wind speed. Writing the eight measurements of one day as a row, and placing the rows of the 366 days under one another, gives a table of numbers. That table is the feature matrix of the dataset.

Consider a dataset $\mathcal{D}$ of m data points with feature vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$, each in $\mathbb{R}^d$. It is convenient to collect the individual feature vectors into the feature matrix:

$$\mathbf{X} := (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)})^\top = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_d^{(m)} \end{pmatrix} \in \mathbb{R}^{m \times d}.$$

Note that the feature matrix is of size $m \times d$, i.e., it has m rows and d columns. Each row holds one data point and each column holds one feature, so reading the feature matrix along a row or along a column answers two different questions (see Fig. 1).

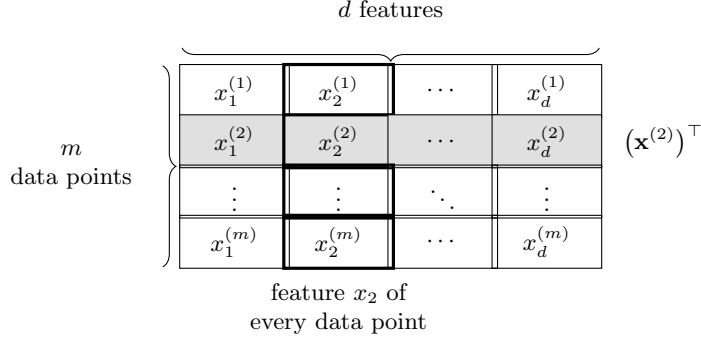


Figure 1: The feature matrix $\mathbf{X}$ of a dataset with m data points and d features. The shaded row is the transposed feature vector $(\mathbf{x}^{(2)})^\top$ of the second data point; the heavily framed column collects the second feature x_2 of every data point

The feature matrix holds md numbers, and a few summaries computed from it already say much about the learning task. The first is its shape, the pair (m, d) . It says whether there are more data points than features. Linear regression has a unique solution only when $\mathbf{X}$ has full column rank, which requires $m \geq d$ (see normal equations). More features than data points is the regime in which overfitting is unavoidable without regularization (Hastie et al., 2009, Sect. 3.4).

The next summaries come from the singular value decomposition (SVD) of $\mathbf{X}$. Its singular values $\lambda_1 \geq \dots \geq \lambda_{\min\{m, d\}} \geq 0$ are at most $\min\{m, d\}$ numbers, and the count of the nonzero ones is the rank of $\mathbf{X}$ (Golub and Van Loan, 2013, Sect. 2.4). A rank below d means that some feature is a linear combination of the others and carries nothing new. The squares of the singular values are the eigenvalues of $\mathbf{X}^\top \mathbf{X}$, and this matrix divided by m is the sample covariance matrix of features with zero sample mean. The spectrum is therefore what principal component analysis (PCA) uses to decide how many features a dimensionality reduction can keep.

One number summarizes how far apart the singular values lie, the condition number $\kappa(\mathbf{X}) = \lambda_1 / \lambda_{\text{rank}(\mathbf{X})}$. It is the ratio of the largest singular value to the smallest nonzero one. It bounds how far the linear regression solution moves when the labels are perturbed, and it governs how fast gradient descent (GD) converges on the least squares objective function (Boyd and Vandenberghe, 2004, Sect. 9.3).

Fig. 2 shows these summaries for the weather year at Krems an der Donau. Each data point is a day and carries the eight measurements as its features, so the feature matrix is of shape $(366, 8)$ and holds 2928 numbers. Each feature is first scaled to zero sample mean and unit sample variance, because the measurements carry different units. Eight singular values then summarize the whole year: the largest is 36.4 and the smallest 0.05, the two largest carry 64 percent of the sum of the squared singular values, and the four largest carry 89 percent.

The smallest singular value lies far below the others, and it names a redundancy among the features. The mean temperature of a day differs from the midpoint of its lowest and its highest temperature by 0.02 degrees Celsius on average. Dropping the mean temperature raises the smallest singular value to 4.19 and lowers the condition number from 748 to 7.6. The rank is 8 either way, so the rank alone does not see this near-dependency and the singular values do.

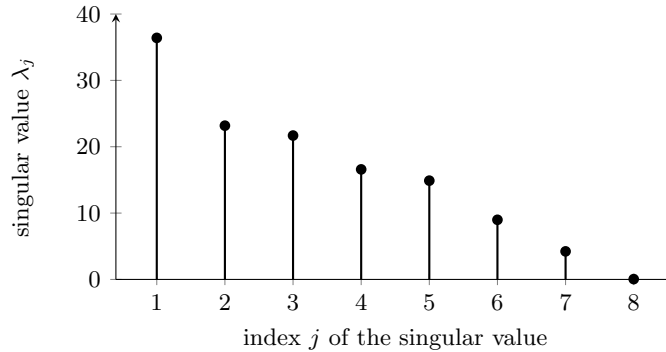


Figure 2: The singular values of the feature matrix of the 366 days of 2024 at Krems an der Donau, each data point carrying eight measurements scaled to zero sample mean and unit sample variance. Eight numbers summarize a matrix of 2928 entries. The eighth is 0.05, far below the seventh, which is the near-dependency among the three temperature measurements. Data generated by `pythondemos/featuremtx.py`

Cross References

dataset, data point, feature vector, feature, matrix, SVD, singular value, rank, condition number, spectrum, sample covariance matrix, PCA, dimensionality reduction

References

1. Boyd and Vandenberghe (2004). Convex Optimization. Cambridge Univ. Press, Cambridge, U.K.
2. Golub and Van Loan (2013). Matrix Computations. 4th ed., The Johns Hopkins Univ. Press, Baltimore, MD, USA.
3. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.