feature space

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

Synonyms

input space, instance space

Abstract

The feature space of a machine learning (ML) application is the set of all values that the feature vector of a data point can take. It is the range of the feature transformation that maps a data point to its feature vector. For data points with d numeric features the feature space is usually the Euclidean space $\mathbb{R}^d$, whose Euclidean distance and inner product are what nearest neighbor methods and linear models use. The same data points placed in two feature spaces have different nearest neighbors and yield different hypotheses, so the choice of the feature space is part of the model, and dropping a feature is model pruning. A feature space can also be a finite set, such as the months of a year, or a set of strings or graphs, which a kernel makes accessible. A kernel brings a second feature space with it, a Hilbert space of possibly infinite dimension in which the kernel is the inner product and which a kernel method uses without ever forming its elements.

Definition

A day at Krems an der Donau can be described by a single number, its morning minimum temperature, or by two numbers, that minimum and the maximum temperature of the previous day (GeoSphere Austria, 2026). The first description turns every day into a point of the real line $\mathbb{R}$, the second into a point of the plane $\mathbb{R}^2$. The set of all values that the feature vector of a data point can take is the feature space $\mathcal{X}$ of a machine learning (ML) application. A feature transformation

$$\mathbf{\Phi} : \mathcal{Z} \to \mathcal{X}, \quad \mathbf{x} = \mathbf{\Phi}(\mathbf{z}), \tag{1}$$

maps a data point $\mathbf{z}$ from the set $\mathcal{Z}$ of all data points of the application to its feature vector, and the feature space is the range of this map (Shalev-Shwartz and Ben-David, 2014, Sect. 16.1). The two descriptions of a day are two feature transformations (1) of the same data points, with the feature spaces $\mathbb{R}$ and $\mathbb{R}^2$.

For data points with d numeric features, the default feature space is the Euclidean space $\mathbb{R}^d$ (Axler, 2024, Sect. 1A). The word *space*, rather than *set*,

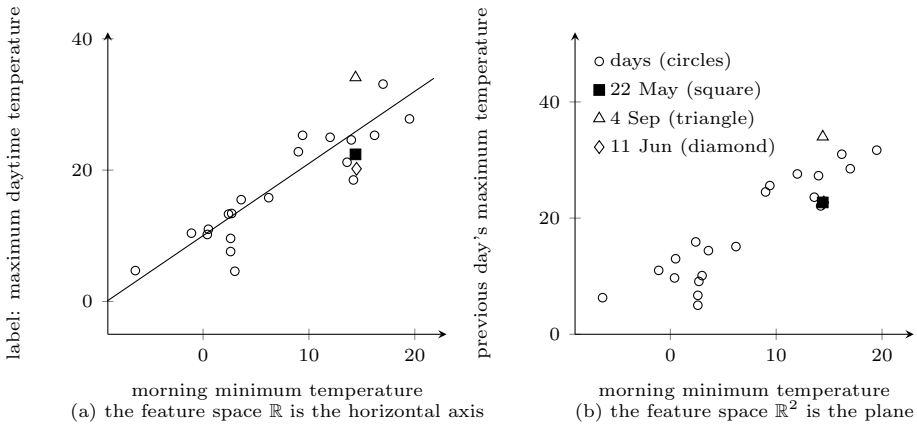
says that $\mathcal{X}$ carries structure that an ML method uses: the Euclidean distance $\|\mathbf{x} - \mathbf{x}'\|_2$ makes $\mathbb{R}^d$ a metric space, and the inner product $\mathbf{x}^\top \mathbf{x}'$ makes it a Hilbert space. The k-nearest neighbors (k-NN) method uses the distance: the nearest neighbor of a data point with feature vector $\mathbf{x}$ in a training set with feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ is $\mathbf{x}^{(r^\star)}$ with

$$r^\star := \operatorname*{arg\,min}_{r=1,\ldots,m} \left\| \mathbf{x} - \mathbf{x}^{(r)} \right\|_2, \tag{2}$$

and the label of that neighbor is the prediction for $\mathbf{x}$ (Cover and Hart, 1967; Hastie et al., 2009, Sect. 2.3.2). A linear model uses the inner product: its hypotheses are the maps $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ with model parameters $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ (Hastie et al., 2009, Sect. 2.3.1).

Which feature space the 366 days of 2024 at Krems are placed in changes what both methods deliver. The label of a day is its maximum daytime temperature. On 22 May 2024 the morning minimum was 14.4 degrees Celsius, the maximum of the previous day 22.7, and the maximum of the day itself 22.4. In the feature space $\mathbb{R}$, the nearest neighbor (2) of this day among the other 365 days is 4 September, with the same morning minimum 14.4 but a maximum of 34.1. In $\mathbb{R}^2$ it is 11 June, with the feature vector $(14.5, 22.7)^\top$ and a maximum of 20.2 (Fig. 1). The prediction by the nearest neighbor misses the label by 11.7 degrees in the first feature space and by 2.2 degrees in the second, and 346 of the 366 days have a different nearest neighbor in $\mathbb{R}$ than in $\mathbb{R}^2$. A linear model learned by least squares from all 366 days has a training mean squared error (MSE) of 19.2 in $\mathbb{R}$ and of 9.2 in $\mathbb{R}^2$.

The choice of the feature space is therefore part of the model. A hypothesis space is a set of maps $h : \mathcal{X} \to \mathcal{Y}$ into the label space $\mathcal{Y}$, so changing $\mathcal{X}$ changes which maps it contains. The hypotheses $h(x) = wx + b$ on $\mathbb{R}$ are, read as maps of the pair $(x_1, x_2)^\top$, the hypotheses $w_1 x_1 + w_2 x_2 + b$ on $\mathbb{R}^2$ with $w_2 = 0$. Dropping a feature thus shrinks the hypothesis space, which is model pruning, one of the three routes that regularization distinguishes: the training error of empirical risk minimization (ERM) cannot fall (19.2 against 9.2), while the smaller hypothesis space may generalize better. Feature learning methods and feature standardization change the feature space in the other direction, by learning or rescaling the feature transformation (1). Rescaling one feature changes the distances in the nearest neighbor rule (2), and with them the nearest neighbors.



(a) the feature space $\mathbb{R}$ is the horizontal axis

(b) the feature space $\mathbb{R}^2$ is the plane

Figure 1: The same days of 2024 at Krems in two feature spaces: every 18th of the 366 days (circles), 22 May (filled square), its nearest neighbor in $\mathbb{R}$, 4 September (triangle), and its nearest neighbor in $\mathbb{R}^2$, 11 June (diamond), 23 days in all. (a) The feature space $\mathbb{R}$ of the morning minimum temperature, drawn against the label; the distance between two days is the horizontal gap between their markers, and the solid line is the linear model hypothesis learned by least squares. (b) The feature space $\mathbb{R}^2$ of the morning minimum and the previous day's maximum; the filled square and the diamond almost coincide, while the triangle is far from both. Data generated by `pythondemos/featurespace.py`

A feature space need not be a Euclidean space. Describing a day by its month gives the finite feature space $\mathcal{X} = \{1, \ldots, 12\}$. A hypothesis on it is a table of 12 values, and the table that minimizes the training MSE holds the monthly means of the maximum temperature, with an MSE of 17.0. One-hot encoding (see feature vector) maps the 12 months to the 12 unit vectors of $\mathbb{R}^{12}$, and on these every table is a linear model: least squares learns the same 12 means. Under the discrete metric, which is 0 for two equal and 1 for two different months, each of the 30 other days of May is a nearest neighbor of 22 May. Other feature spaces without arithmetic are the strings over an alphabet, for text documents and DNA sequences, and the graphs, for molecules. A method then accesses the feature space through a kernel that measures the similarity of two of its elements (Schölkopf and Smola, 2002, Sect. 13.2; Vishwanathan et al., 2010). The hours of a day, placed on a cycle graph, are the feature space of kernel ridge regression.

A kernel brings a second feature space with it. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is an inner product of transformed feature vectors $\phi(\mathbf{x})$ in a Hilbert space $\mathcal{H}$, $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$ (Bishop, 2006, Eq. (6.1)), and $\mathcal{H}$ is also called the feature space of the kernel. For the polynomial kernel of degree two on $\mathbb{R}^2$, this space is explicit,

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \mathbf{x}^\top \mathbf{x}'\right)^2 = \phi(\mathbf{x})^\top \phi(\mathbf{x}'), \tag{3}$$

with the feature transformation $\phi(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2\right)^\top \in \mathbb{R}^6$, so a linear model on the transformed feature vectors is a quadratic hypothesis on the original feature space (Shalev-Shwartz and Ben-David, 2014, Example 16.1). The 366×366 matrix of the polynomial kernel values (3) of the days at Krems has rank 6, the dimension of $\mathbb{R}^6$. The feature space of the Gaussian kernel $\exp\left(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\sigma^2)\right)$ has infinite dimension (Shalev-Shwartz and Ben-David, 2014, Example 16.2): for $\sigma = 1$ degree, the matrix of its values on the 366 days has rank 366, so no feature space of smaller dimension produces these values. A kernel method never forms the transformed feature vectors; it uses the feature space $\mathcal{H}$ only through kernel values, which is what makes a space of infinite dimension usable (Bishop, 2006, Sect. 6.2; Shalev-Shwartz and Ben-David, 2014, Sect. 16.2). Fig. 2 draws three further shapes of a feature space: the real line, a bounded convex subset of the plane, and the nodes of an undirected graph. The second is the feature space of a sensor with a bounded range, and in the third the distance between two feature vectors is the length of the shortest path between their nodes.

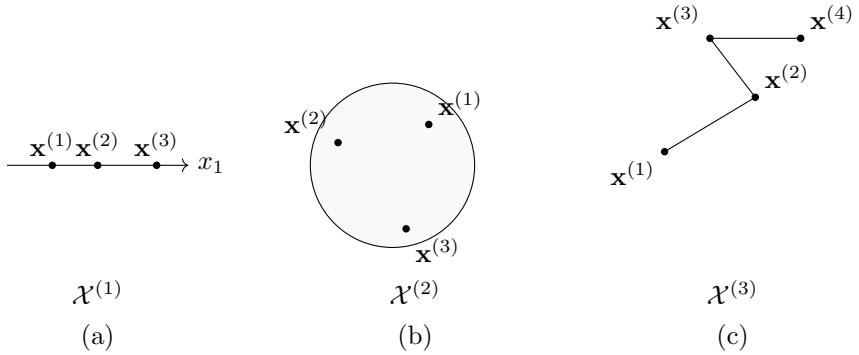


Figure 2: Three feature spaces. (a) The real line $\mathcal{X}^{(1)} = \mathbb{R}$. (b) A bounded convex set $\mathcal{X}^{(2)} \subseteq \mathbb{R}^2$. (c) A finite feature space $\mathcal{X}^{(3)}$ whose elements are the nodes of an undirected graph

Cross References

feature vector, feature, feature transformation, data point, label space, Euclidean space, Euclidean distance, metric space, Hilbert space, inner product, k-NN, linear model, hypothesis space, model pruning, feature learning, feature standardization, kernel, kernel method, kernel ridge regression, RKHS, graph

References

1. Axler (2024). Linear Algebra Done Right. 4th ed., Springer, Cham, Switzerland.

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

3. Cover and Hart (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp. 21–27.

4. 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/v2-1d.

5. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

6. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

7. Vishwanathan et al. (2010). Graph Kernels. Journal of Machine Learning Research, 11, pp. 1201–1242.

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