hypothesis space

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

hypothesis class, model

Abstract

A hypothesis space $\mathcal{H}$ is a set of hypothesis maps $h : \mathcal{X} \to \mathcal{Y}$ from a feature space into a label space. Every machine learning (ML) method uses an underlying hypothesis space, which is a subset of the set $\mathcal{Y}^{\mathcal{X}}$ of all possible maps from $\mathcal{X}$ into $\mathcal{Y}$. Available computational resources limit the size of $\mathcal{H}$, which in turn shapes the method's computational cost and predictive behavior. The hypothesis space typically carries geometric structure: real-valued hypotheses can be compared via the deviation in their predictions on a reference dataset, and a parametric model inherits a geometric structure from the underlying parameter space $\mathcal{W}$.

Definition

A weather service predicting tomorrow's maximum daytime temperature from today's morning temperature must commit to a set of candidate prediction maps: for example, all linear functions of the morning temperature. This set of candidates is the hypothesis space of the method; more generally, every machine learning (ML) method commits to a specific hypothesis space $\mathcal{H}$, the set from which it learns a single hypothesis $\hat{h}$. Formally, $\mathcal{H}$ is a subset of the set $\mathcal{Y}^{\mathcal{X}}$ of all possible maps from the feature space into the label space (see Fig. 1).

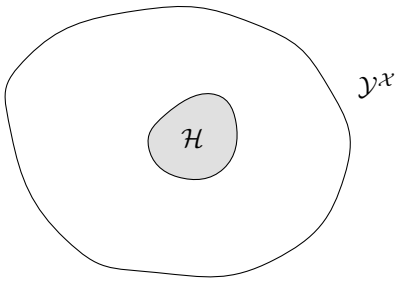

Figure 1: Hypothesis space $\mathcal{H}$ of an ML method as a restricted subset of the much larger set $\mathcal{Y}^{\mathcal{X}}$ of all possible maps from the feature space $\mathcal{X}$ into the label space $\mathcal{Y}$

Linear regression and other linear methods use the linear model as $\mathcal{H}$, i.e., the set of all linear maps $\mathbb{R}^d \to \mathbb{R}$ (Shalev-Shwartz and Ben-David, 2014, Ch. 2). Another canonical example is the set of input–output maps realizable by an artificial neural network (ANN) of fixed architecture, as the model parameters vary.

The choice of $\mathcal{H}$ is a central design decision for an ML method. From an ML engineering perspective, it is guided by two factors. The first is the available computational resources, such as memory, processing time, and communication bandwidth. The second factor is the number of available data points, which limits the maximum size of $\mathcal{H}$ that can be effectively trained without overfitting.

The size of the underlying $\mathcal{H}$ is an important characteristic of an ML method. In principle, $\mathcal{H}$ could be as large as $\mathcal{Y}^{\mathcal{X}}$ itself. In practice, finite computational resources restrict an ML method to a much smaller subset $\mathcal{H}$. The size of $\mathcal{H}$ allows the anticipation of generalization before any training is carried out. A small $\mathcal{H}$ typically incurs a small generalization gap, while a large $\mathcal{H}$ is more prone to overfitting. For a finite $\mathcal{H}$ we can use the cardinality as effective size. For an infinite $\mathcal{H}$, an effective size can be quantified by the Vapnik–Chervonenkis dimension (VC dimension) or the Rademacher complexity.

The same $\mathcal{H}$ can be used by different ML algorithms, such as empirical risk minimization (ERM) on a fixed training set, online learning on a stream of data points, or Bayesian inference that returns a posterior distribution over $\mathcal{H}$ rather than a single hypothesis (Bishop, 2006, Sect. 1.2.3).

The word *space* (rather than *set*) reflects the fact that $\mathcal{H}$ typically carries some geometric structure. In general, $\mathcal{H}$ is not a vector space: the sum of two hypotheses in $\mathcal{H}$ need not belong to $\mathcal{H}$. What is more, there are hypothesis spaces that are not equipped with an algebraic structure at all, i.e., there is no meaningful way to add two hypotheses in $\mathcal{H}$ or to multiply a hypothesis by a scalar.

For real-valued hypotheses, two $h, h' \in \mathcal{H}$ can be compared by their average squared prediction deviation on a reference dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$,

$$\frac{1}{m} \sum_{r=1}^{m} \big(h(\mathbf{x}^{(r)}) - h'(\mathbf{x}^{(r)})\big)^2.$$

Fig. 2 illustrates this comparison: the two hypotheses are evaluated on the data points in $\mathcal{D}$ and the squared differences in their predictions are averaged. This construction is not meaningful for hypotheses with a finite label space, as used in classification methods. Instead, two such hypotheses can be compared by their disagreement rate on $\mathcal{D}$.

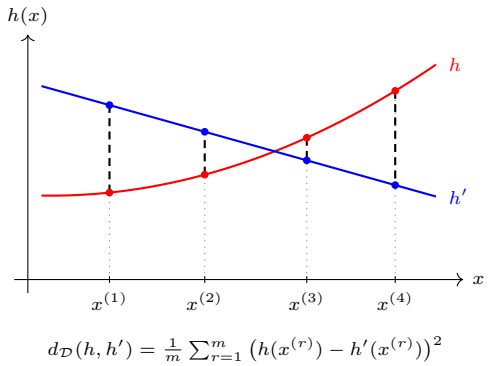

Figure 2: Comparing two hypotheses h (red) and h' (blue) by their predictions at four scalar features $x^{(1)},\ldots,x^{(4)}$. The dashed segments mark the gap $h(x^{(r)}) - h'(x^{(r)})$ for each data point; the average of its square defines the distance $d_{\mathcal{D}}(h,h')$ between h and h'

A parametric model $\mathcal{H} = \{h^{(\mathbf{w})} : \mathbf{w} \in \mathcal{W}\}$ inherits a geometric structure from the underlying parameter space $\mathcal{W} \subseteq \mathbb{R}^d$. A norm on $\mathbb{R}^d$ induces the candidate distance $d(h^{(\mathbf{w})}, h^{(\mathbf{w}')}) := \|\mathbf{w} - \mathbf{w}'\|$ on $\mathcal{H}$. When the parameterization $\mathbf{w} \mapsto h^{(\mathbf{w})}$ is injective, distinct weight vectors give distinct hypotheses, and this candidate distance is a genuine metric. The linear model $\mathcal{H} = \{h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$ is a canonical example of an injective parameterization: distinct weight vectors $\mathbf{w}$ produce distinct linear maps, so $\|\mathbf{w} - \mathbf{w}'\|$ is a metric on $\mathcal{H}$. When the parameterization is not injective, two distinct weight vectors $\mathbf{w} \neq \mathbf{w}'$ can represent the same hypothesis $h^{(\mathbf{w})} = h^{(\mathbf{w}')}$. The distance $\|\mathbf{w} - \mathbf{w}'\|$ is then strictly positive even though the two hypotheses coincide, so it is only a pseudo-metric. Permuting the hidden units of an ANN, for instance, changes $\mathbf{w}$ but leaves $h^{(\mathbf{w})}$ unchanged.

Cross References

hypothesis, model, map, linear model, parametric model, parameter space, metric, norm, ERM, online learning, Bayesian inference, posterior distribution, VC dimension, Rademacher complexity

References

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

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