



Why

Linear predictors are simple and we know how to select the parameters. The main downside is that there may not be a linear relationship between inputs and outputs.

Definition

A *feature map* (or *regression function*) for outputs A is a mapping $\phi : A \rightarrow \mathbf{R}^d$. In this setting, we call $a \in A$ the *raw input record* and we call $\phi(a)$ an *embedding*, *feature embedding* or *feature vector*. We call the components of a feature vector the *features*. We call $\phi(A)$ the *regression range*.

A feature map is *faithful* if, whenever records a_i and a_j are in some sense “similar” in the set A , the embeddings $\phi(a_i)$ and $\phi(a_j)$ are close in the vector space \mathbf{R}^d .

Since it is common for raw input records $a \in A$ to consist of many fields, it is regular to have several feature maps ϕ_i which operate component-wise on the fields of a . These are sometimes called *basis functions*, by analogy with real function approximators (see **Real Function Appproximators**). We concatenate these field feature maps and commonly add a constant feature 1. Since \mathbf{R}^d is a vector space, it is common to refer to it in this case as the *feature space*.

Given a dataset $a = (a^1, \dots, a^n)$ in A and a feature map $\phi : A \rightarrow \mathbf{R}^d$, the *embedded dataset* of a with respect to ϕ is the dataset $(\phi(a^1), \dots, \phi(a^n))$ in \mathbf{R}^d .

Featurized consistency: a route around $X \neq \mathbf{R}^d$

Recall that a dataset is parametrically consistent with the family $\{h_\theta : X \rightarrow Y\}_\theta$ if there exists θ^* so that the dataset is consistent with θ^* . We saw how to pick θ if we use a linear model with a squared loss (see **Least Squares Linear Regressors**).

Let $\mathcal{G} = \{g_\theta : \mathbf{R}^d \rightarrow \mathbf{R}\}_\theta$. A dataset is *featurized parametrically consistent* with respect to the family \mathcal{G} and the feature map $\phi : X \rightarrow \mathbf{R}^d$ if it is parametrically consistent with respect to $\mathcal{G} \circ \phi = \{g \circ \phi \mid g \in \mathcal{G}\}$.

The interpretation is that we have transformed the problem of selecting a predictor on an arbitrary space X to the problem of selecting a predictor on the space \mathbf{R}^d . In so doing, we can continue to use simple predictors, such as those that are linear and minimize the squared error on the dataset.¹

In other words, we have “shifted emphasis” from the model function $h : X \rightarrow \mathbf{R}$ to the *regression function* from $\mathbf{R}^d \rightarrow \mathbf{R}$. If we know the features and the input x , then we know the *regression vector* $\phi(x)$. The *regression range* is the set $\{\phi(x) \mid x \in X\}$. In this case linearity pertains to the parameters $\theta \in \mathbf{R}^d$ instead of the inputs (or experimental conditions) $x \in X$.

¹Future editions are likely to modify this section.

