high-dimensional regime

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

A machine learning (ML) method operates in the high-dimensional regime when the effective dimension of its hypothesis space exceeds the sample size m of the training set. For a linear model, the regime begins where the number d of features passes the number m of data points. The feature matrix then has rank at most m, so the least squares solution is not unique and every solution interpolates the training set. A training error of zero therefore carries no information about the risk, so the regime removes the usual evidence about generalization. Regularization and dimensionality reduction restore a usable method, the first by adding a penalty term to the training error and the second by shrinking the hypothesis space. High-dimensional statistics develops theory and methods for this regime, and its guarantees need a structural assumption such as sparsity.

Definition

Consider a weather service at Krems an der Donau (Austria) that predicts the maximum daytime temperature of a day from the maxima of the previous d days. Each day is a data point with feature vector $\mathbf{x} \in \mathbb{R}^d$ and label $y \in \mathbb{R}$, and a linear model $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ is learned by least squares on a training set $\mathcal{D}^{(\text{train})}$ of $m = 30$ days of 2023. For $d = 1$, the training error is 12.4 and the validation error on the validation set $\mathcal{D}^{(\text{val})}$ of the 366 days of 2024 is 10.1 squared degrees. For $d = 29$, the two numbers are 1.7 and 100.4, and for $d = 30$ they are 0 and 13444. Once the number of features reaches the number of data points in $\mathcal{D}^{(\text{train})}$, the training error carries no information about the loss on days outside $\mathcal{D}^{(\text{train})}$. Fig. 1 draws both errors for every d from 1 to 60.

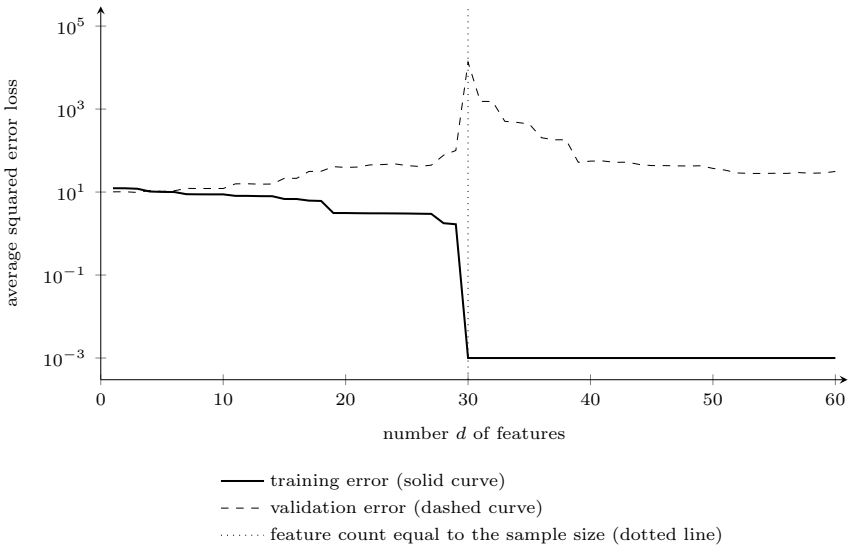


Figure 1: The training error (solid curve) and the validation error (dashed curve) of a linear model that predicts the maximum daytime temperature of a day at Krems from the maxima of the previous d days, learned by least squares on $m = 30$ days of 2023 and evaluated on the 366 days of 2024. The training error is drawn at the axis floor 10^{-3} wherever it is zero. The dotted line marks the feature count $d = m = 30$ at which the high-dimensional regime begins. Data generated by `pythondemos/highdimregime.py`

A machine learning (ML) method operates in the high-dimensional regime when the effective dimension of its hypothesis space $\mathcal{H}$ exceeds the sample size m of the training set,

$$d_{\text{eff}}\left(\mathcal{H}\right) > m. \tag{1}$$

For the linear model above, the condition (1) reads $d > m$. Empirical risk minimization (ERM) with the squared error loss delivers a solution of the least squares problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{r=1}^{m}\left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right)^2. \tag{2}$$

Here, the feature matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ stacks the feature vectors of $\mathcal{D}^{(\text{train})}$ as its rows and the vector $\mathbf{y} \in \mathbb{R}^m$ stacks their labels.

The rank of $\mathbf{X}$ is at most m, so for $d > m$ the nullspace of $\mathbf{X}$ holds a nonzero vector and the solution of the least squares problem (2) is not unique: its minimizers form an affine set of dimension at least $d-m$. The pseudoinverse picks the one of smallest Euclidean norm, $\widehat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y}$. Every minimizer drives the training error to zero, which is interpolation of $\mathcal{D}^{(\text{train})}$. The curse of dimensionality is a separate consequence of a large d: the data points spread thin over the feature

space $\mathbb{R}^d$ however many of them there are, while the high-dimensional regime compares two counts.

Generalization is the performance of a learned hypothesis on data points outside $\mathcal{D}^{(\text{train})}$, and the regime removes the evidence that a small training error usually provides about it. A hypothesis that interpolates $\mathcal{D}^{(\text{train})}$ has a training error of zero whichever labels the data points of $\mathcal{D}^{(\text{train})}$ carry, so a zero training error leaves the risk undetermined and overfitting cannot be diagnosed from $\mathcal{D}^{(\text{train})}$ alone. A large artificial neural network (ANN) is trained in the same regime, with far more trainable model parameters than data points (Zhang et al., 2017, Sect. 1). A deep network reaches a training error of zero even on a training set whose labels have been replaced by random ones (Zhang et al., 2017, Sect. 2). Learning theory measures the size of $\mathcal{H}$ by its Vapnik–Chervonenkis dimension (VC dimension) and by its Rademacher complexity. The sample size that a generalization guarantee needs grows in proportion to the VC dimension (Shalev-Shwartz and Ben-David, 2014, Thm. 6.8).

Two routes restore a usable method on the same features, and both are forms of regularization. Adding the penalty term $\alpha\|\mathbf{w}\|_2^2$ to the training error gives ridge regression. With $\alpha = 100$, the validation error of the weather example stays between 10.2 and 22.9 squared degrees for every d from 1 to 60. Dimensionality reduction shrinks $\mathcal{H}$ instead: least squares on the first $d' = 5$ principal directions of the feature vectors of $\mathcal{D}^{(\text{train})}$, which is principal component analysis (PCA), keeps the validation error between 9.8 and 22.1. At $d = m$ neither learned hypothesis interpolates $\mathcal{D}^{(\text{train})}$: the training error is 4.2 with the penalty term and 11.2 after the reduction. Fig. 2 draws the validation error of both against the validation error of least squares on all d features.

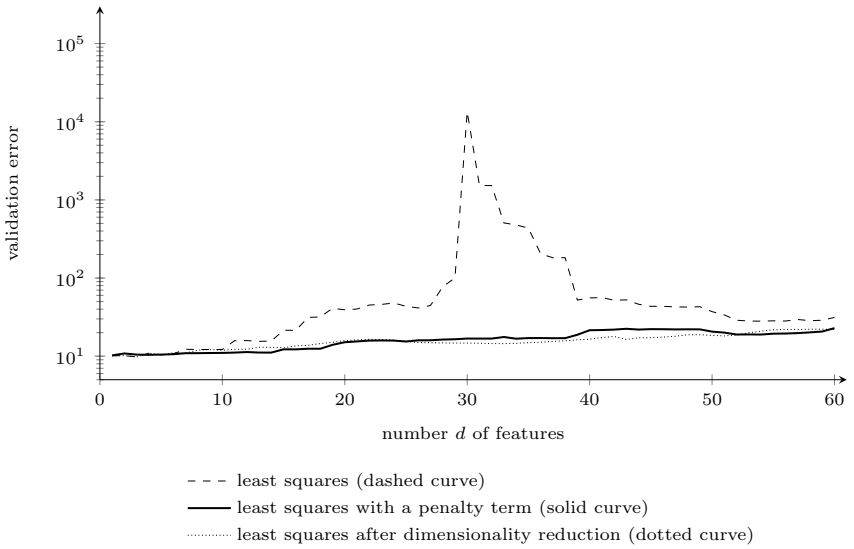


Figure 2: The validation error of the same linear model at Krems against the number d of features, for least squares (dashed curve), for least squares with the penalty term at $\alpha = 100$ (solid curve), and for least squares on the first $d' = 5$ principal directions (dotted curve). Data generated by `pythondemos/highdimregime.py`

High-dimensional statistics is the branch of mathematical statistics that develops theory and methods for this regime (Bühlmann and Geer, 2011; Wainwright, 2019, Sect. 1.1). Without a structural assumption on the model parameters, no method estimates them consistently once d exceeds m (Wainwright, 2019, Sect. 7.1.1). One such assumption is sparsity, which says that few model parameters are nonzero. The least absolute shrinkage and selection operator (Lasso) exploits it with a penalty term built from the ℓ_1-norm (Tibshirani, 1996). The regime is common wherever features are cheap and data points are not: a gene-expression dataset with 2308 features and 63 data points needs regularization before any classifier can be learned on it (Hastie et al., 2009, Sect. 18.2).

Cross References

effective dimension, curse of dimensionality, interpolation, overfitting, generalization, regularization, dimensionality reduction, ridge regression, Lasso, VC dimension

References

1. Bühlmann and Geer (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer-Verlag, Berlin, Germany.

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

3. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. J. Roy. Statist. Soc.: Ser. B (Methodological), 58(1), pp. 267–288.

4. Wainwright (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Univ. Press, Cambridge, U.K.

5. Zhang et al. (2017). Understanding deep learning requires rethinking generalization. In: Int. Conf. Learn. Represent. (ICLR).

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