A UNIFIED FRAMEWORK FOR UNSUPERVISED LEARNING

GUILLE D. CANAS[♯,†], TOMASO POGGIO[♯,†], LORENZO ROSASCO[♯,†]
♯ - *CBCL, MCGOVERN INSTITUTE, BCS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY*
† - ISTITUTO ITALIANO DI TECNOLOGIA
{GUILLEDC, LROSASCO}@MIT.EDU, TP@AI.MIT.EDU

We describe a common framework under which a number of widely used, and seemingly disparate algorithms for unsupervised learning can be cast. In this framework, an unsupervised algorithm simply estimates the support of the distribution generating the data. The formulation makes clear the separation between the approximation and statistical components of the algorithms, leading to bounds on their performance. The analysis further allows to set the free parameters of the algorithms to optimize the performance bounds. We show a precise relation between the support estimation view, and the alternative (generative) view of unsupervised learning as approximating the data-generating distribution, establishing a formal relation between two interpretations of unsupervised learning.

Interpretation of unsupervised learning. The problem of unsupervised learning is widely described as that of recovering, or finding structure (patterns) in data sampled from a probability distribution [6]. The importance of unsupervised learning is clear both from the availability of unlabeled data, as well as from its central role in practical supervised learning systems. In particular, due to the recent progress in sensory systems and other data-acquisition technology, the amount of unlabeled data is typically much more abundant, and often much more cheaply acquired, than labeled data. Even for supervised learning tasks, many state-of-the-art systems use a preliminary unsupervised stage, which is often of central importance [7, 13, 15]. From a theoretical perspective, it is widely accepted that an unsupervised algorithm that is able to accurately approximate the input data while reducing its dimensionality can have a significant impact on downstream supervised stages, whose finite sample performance may be subject to some form of curse of dimensionality [4].

A unified view. The structure inferred in unsupervised learning has been variously interpreted to be, possibly depending on the application, a clustering or partitioning of the data [9], a, possibly sparse, encoding of the data [12], or learning a lower-dimensional approximation (manifold learning) [1, 14]. All these interpretations have in common that they produce: 1) a compact representation, or estimate of the support of the data, in the form an *approximating set*, and 2) an associated *encoding* with respect to this representation. For instance, the k-means algorithm approximates the data by a discrete set of points (means), and encodes a sample by the index of its closest mean (in way that resembles vector quantization [5]). Similarly, in PCA, the computed approximating set is a lower-dimensional affine space, where each data point is encoded using its local coordinates in this space. Several other unsupervised learning algorithms, such as sparse coding [12], k-flats [2], or non-negative matrix factorization [8], can be easily shown to conform to this model [10, 11].

Empirical and expected error. In practice, an unsupervised learning algorithm typically has a tunable parameter that controls the size, or complexity of the approximating set. For example, this parameter may be the number of means in k-means, the dimension of the affine spaces in PCA, or the number of dictionary elements for sparse coding. Once the parameter is chosen, most algorithms, including all the ones mentioned above, proceed by computing the set that best approximates the available samples. In other words, they find the set that minimizes the distance from the samples to their projection onto the set (in some cases, with some additional constraint in the encoding, such as sparsity). Clearly, as the size of the set is increased, the distance from the *available* samples to the approximating set cannot increase, and typically decreases. However, since we are ultimately interested in the expected performance of the algorithm with respect to a random sample drawn from the true distribution, a question naturally arises of whether a good approximation of the samples necessarily implies a good approximation with respect to the distribution.

Regularization in unsupervised learning and bias-variance tradeoff. In this work, we show that, in general, the answer to this question is negative, and thus that there is in general an optimal size/complexity for the approximating set. In other words there is in general a tradeoff in which small sets may be of insufficient complexity to accurately approximate the input, while sufficiently large sets that reproduce the empirical samples with high fidelity may fail to perform well on the true distribution. Indeed, we show that, for widely-used unsupervised algorithms, there is an inherent tradeoff between approximation accuracy and model size, which can be precisely

analyzed. We illustrate these ideas using k-means and k-flats as examples, and demonstrate this tradeoff both in theory and empirically. While there exist previous theoretical studies related to our analysis, see for example [3], to the best of our knowledge, this inherent tradeoff in the model complexity of unsupervised learning algorithms, very much akin to the classical bias-variance tradeoff of supervised learning, and its implications for their analysis and practical use, has not been previously analyzed. We conclude noting that the setup we consider is very much related to the problem of dictionary learning, where one is interested in finding compact/parsimonious data representations. In this view, our results prove that for a given dataset (and a corresponding distribution) there exists an optimal dictionary size defined by a suitable bias variance tradeoff.

REFERENCES

[1] M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, The University of Chicago, 2003.

[2] P. S. Bradley and O. L. Mangasarian. k-plane clustering. *J. of Global Optimization*, 16:23–32, January 2000.

[3] Joachim M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical report, 1998.

[4] Jerome H. Friedman. On bias, variance, 01 loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.*, 1:55–77, January 1997.

[5] Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2000.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[7] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18:1527–1554, July 2006.

[8] D. D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[9] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[10] A. Maurer and M. Pontil. K–dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839 –5846, nov. 2010.

[11] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems 23*, pages 1786–1794. 2010.

[12] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[13] S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. Mathematics of the neural response. *Found. Comput. Math.*, 10:67–91, January 2010.

[14] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[15] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV (5)*, pages 141–154, 2010.