The Curse of Dimensionality for Local Kernel Machines

Yoshua Bengio, Olivier Delalleau and Nicolas Le Roux
Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc, H3C 3J7, Canada
{*bengioy,delallea,lerouxni*}*@iro.umontreal.ca*
http://www.iro.umontreal.ca/∼lisa

We present a series of arguments supporting the claim that a large class of modern learning algorithms based on local kernels are highly sensitive to the curse of dimensionality. These algorithms include in particular local manifold learning algorithms such as Isomap (Tenenbaum, de Silva and Langford, 2000) and LLE (Roweis and Saul, 2000), and support vector classifiers (Boser, Guyon and Vapnik, 1992) with Gaussian or other local kernels. The results show that these algorithms are local in the sense that crucial properties of the learned function at x depend on the neighbors of x in the training set. This makes them highly sensitive to the curse of dimensionality, well studied for classical non-parametric statistical learning algorithms. There is a large class of data distributions for which non-local solutions could be expressed compactly and potentially be learned with few examples, but which will require a large number of local bases and therefore a large number of training examples when using a local learning algorithm.

The curse of dimensionality for classical non-parametric models can be traced to the bias-variance dilemma. When the kernel bandwidth is small, the effective number of examples around x that influence the prediction is small, making the prediction highly sensitive to that particular training sample, i.e. yielding to high variance. When the kernel bandwidth is large, the prediction becomes overly smooth, i.e. yielding to high bias. Indeed the prediction at x can be seen in classical non-parametric predictors as an average of empirical observations over the (possibly weighted) volume of the neighbors. The problem is that in high dimension d (or more precisely when the data lie on or near a manifold of dimension d), that volume (where one expects to find enough effective neighbors) grows exponentially, making the bias very large just to keep variance constant. As shown many times in the statistics litterature (Härdle et al., 2004), this yields to extremely slow convergence, with the number of examples required to reach a given generalization error scaling as n_1^{cd} where n_1 (> 1) and c (> 0) are constants. Note that the d that matters here is the data manifold's dimension, since these methods are essentially based on the Euclidean distance between near neighbors and these converge to the geodesic distance on the manifold as the number n of training points augments.

We present a series of arguments that suggest that modern kernel methods such as SVMs (for supervised learning) and spectral manifold learning methods (for unsupervised learning) suffer also from the curse of dimensionality (although to a lesser extent in the case of SVMs, since they can become well regularized linear classifiers when these generalize better than too local predictors). In the arguments below we assume that the predictor function is $f(x) = b + \sum_{i=1}^{n} \alpha_i K(x, x_i)$. Note this includes spectral clustering and spectral dimensionality reduction (Bengio et al., 2004) when seen as inductive algorithms. We also assume that $K(u, v)$ is local (converging to 0 when $||u - v|| \to \infty$) or that its derivative is local (in the sense that its derivative w.r.t v only depends on the near neighbors of v). This is true for the Gaussian kernel as well as all the spectral manifold learning algortithms studied for example in (Bengio et al., 2004). Some of these statements are specific to the Gaussian kernel: $K(u, v) = e^{-||u-v||^2/\sigma^2}$ with global bandwidth hyper-parameter σ. The training examples are $\{x_1, \ldots, x_n\}$, with +1 or -1 labels $\{y_1, \ldots, y_n\}$ in the classification case. Note that $\frac{\partial f(x)}{\partial x}$ represents a tangent vector (or the set of vectors that span the **tangent plane**, when $f(x)$ is multi-dimensional) for spectral manifold learning (Bengio and Monperrus, 2005), while it represents the decision surface's **normal vector** in the case of classification tasks.

1. *When the test example x is far from all the x_i, the predictor either converges to a constant or a nearest neighbor classifier.* Neither of these are good in high dimension.

2. *$\frac{\partial f(x)}{\partial x}$ is constrained to be approximately a linear combination of the vectors $(x - x_i)$ with x_i a near neighbor of x.* In high dimension (when the number of effective neighbors is significantly smaller than the dimension), this is a very strong constraint (either on the shape of the manifold or of the decision surface).

3. *As $\sigma \to \infty$ with Gaussian kernels, a regularized SVM becomes linear and eventually turns into a*

constant classifier.

4. *As $\sigma \to 0$ with Gaussian kernels, the SVM classifier converges to a prediction that only depends on the one nearest neighbor.*

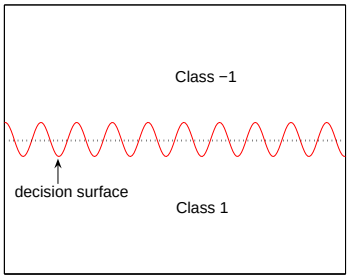
The previous two statements suggest that a better σ is intermediate, i.e. with some examples such that $||x - x_i||$ is on the same order as σ.

5. *When there are examples with $||x - x_i||$ near σ, with x on the decision surface, changes in x small w.r.t. σ yield only small changes in the normal vector of the decision surface.*

The above statement is one about bias: within a ball of radius σ, the decision surface is constrained to be smooth (small changes in x yield small changes in the shape of the surface).

6. *If there exists a line in $\mathbb{R}^d$ that intersects m times with the decision surface S (and is not included in S), then one needs at least $\lceil \frac{m}{2} \rceil$ Gaussians (of same width) to learn S.*

This says that in order to represent a function that varies many times over the range of the data, one needs a number of examples that scales linearly with the number of these "variations". Hence a function that varies a lot necessarily requires a large number of examples with a local kernel, although it might be representable efficiently otherwise. For instance in the case of the figure on the right, one would need at least 10 Gaussians to learn the decision surface, because the dotted line crosses it 19 times (even though it is a simple sinusoidal function).



7. *At least 2^{d-1} examples are required to represent the d-bit parity (the function from $\{0,1\}^d$ to $\{-1,1\}$ which is 1 iff the sum of the bits is even), when using Gaussians with fixed σ centered on data points.*

More formal statements and their proofs can be found in a technical report (Bengio, Delalleau and Le Roux, 2005). One could clearly hope for stronger results, but the above are already strong indications that when the data lie on a high-dimensional and curved manifold, local learning methods such as SVMs, spectral dimensionality reduction and spectral clustering methods are doomed to scale poorly, in the sense that the number of examples required to capture the essential structure in the distribution might grow exponentially with the manifold dimension. We hope that these results will stimulate research on non-local learning methods, and experiments on high-dimensional complex tasks such as those involved in vision and language.

(Incomplete) References

Bengio, Y., Delalleau, O., and Le Roux, N. (2005). The curse of dimensionality for local kernel machines. Technical Report 1258, Département d'informatique et recherche opérationnelle, Université de Montréal.

Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J.-F., Vincent, P., and Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219.

Bengio, Y. and Monperrus, M. (2005). Non-local manifold tangent learning. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*. MIT Press.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer, http://www.xplore-stat.de/ebooks/ebooks.html.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Schmitt, M. (2002). Descartes' rule of signs for radial basis function neural networks. *Neural Computation*, 14(12):2997–3011.

Tenenbaum, J., de Silva, V., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.