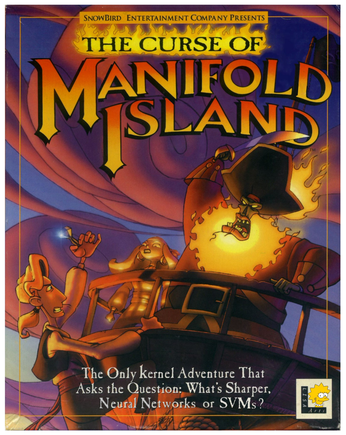
Université de Montréal

The Curse of Highly Variable Functions for Local Kernel Machines

Yoshua Bengio
Olivier Delalleau
Nicolas Le Roux

Outlook



- Smoothness prior is the basis for many algorithms, that learn only from local neighborhoods of training data
- Makes them unfit to learn functions that vary a lot, e.g. needed for complex AI tasks
- Ex: SVMs with local kernel, manifold learning algorithms (LLE, ISOMAP, kernel PCA), graph- based semi-supervised learning, ...
- Need for non- locallearning!

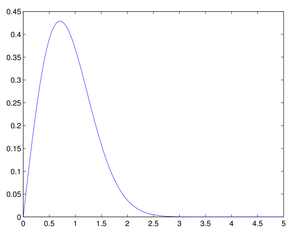
General Curse Idea

Setting: learned function expressed as a linear function of (possibly *data- dependent*) kernel functions:

$$f(x) = b + \sum_{i=1}^{n} \alpha_i K_D(x, x_i) \qquad (1)$$

1. Locality: if the kernel is local in some sense, important properties of $f(\cdot)$ at point x will depend mostly on training examples x_i in neighborhood $\mathcal{N}(x)$ of x.

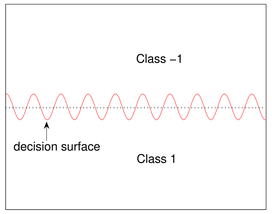
   Ex: derivative of the Gaussian kernel w.r.t. distance $\|x - x_i\|^2$. The derivative of $f(\cdot)$ at x depends on those x_i in $\mathcal{N}(x)$ $\Rightarrow$ e.g. shape of the decision surface in a kernel classifier.



2. Smoothness: important properties of $f(\cdot)$ at x vary slowly in $\mathcal{N}(x)$ (i.e. $f(\cdot)$ is smooth, in some sense, within $\mathcal{N}(x)$).
3. Complexity: if the target function for $f(\cdot)$ is "complex", i.e. it varies a lot, smoothness $\Rightarrow$ need to consider many neighborhoods ("tiling" the space with local patches).
   CURSE! Locality $\Rightarrow$ *need training examples in each patch* (whose number may grow in $O(const^d)$).

Minimum Number of Bases

Corollary of (Schmitt, 2002): if K_D is the Gaussian kernel and $f(\cdot)$ changes sign at least $2k$ times along some straight line (i.e. that line crosses the decision surface at least $2k$ times), then there must be at least k bases (non- zeroα_i's).



Ex: This "complex" sinusoidal decision surface requires a minimum of 10 Gaussians to be learned with a Gaussian kernel classifier.

Parity problem: learning the d- bits parity function

$$\text{parity}: (b_1, \ldots, b_d) \in \{0,1\}^d \mapsto \begin{cases} 1 \text{ if } \sum_{i=1}^{d} b_i \text{ is even} \\ -1 \text{ otherwise} \end{cases}$$

with a Gaussian kernel classifier requires **at least** 2^{d-1} **bases** (when centered on training points).

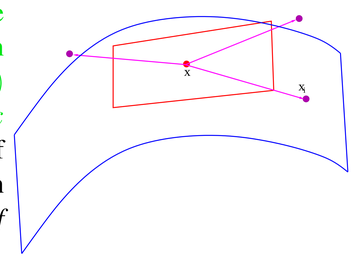
Bottom-line: with a purely local kernel, it can be difficult to learn a simple function that varies a lot.

Manifold Learning

Many manifold learning algorithms can be written as in eq. 1 with K_D having some locality property such that

$$\frac{\partial f(x)}{\partial x} \simeq \sum_{x_i \in \mathcal{N}(x)} \hat{K}_D(x, x_i)(x_i - x)$$

i.e. the tangent plane of the manifold is approximately in the span of the vectors $(x_i - x)$ with x_i a near neighbor of x $\Rightarrow$ **high-variance** estimators if not enough training points in neighborhood of x (*curse of the manifold dimensionality*).



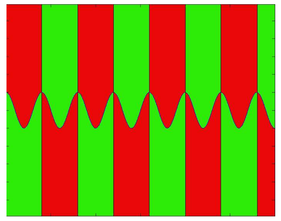
Ex: kernel PCA with a Gaussian kernel, Locally Linear Embedding, ISOMAP, ...

Graph- Based Semi- Supervised Learning

A number of semi- supervised learning algorithms are based on the idea of *propagating labels* on the nodes of a neighborhood graph, starting from known labels, until some convergence is obtained. Typical cost function optimized this way:

$$C(\hat{Y}) = \|\hat{Y}_l - Y_l\|^2 + \mu \hat{Y}^\top L \hat{Y} + \mu\epsilon \|\hat{Y}\|^2$$

Proposition: *the number of regions with constant estimated label is less than (or equal to) the number of labeled examples.*



"Double curse": with a Gaussian kernel, we would need many labeled examples (one in each colored region), and many unlabeled ones (along the sinusoidal line).

$\Rightarrow$ a neighborhood graph may not be appropriate for the task (e.g. parity problem).

Non- Local Learning: We are not doomed!

- complex $\neq$ non- smooth (e.g. Kolmogorov complexity weak prior that can buy a lot of power)
- similar $\neq$ close in vector space $\Rightarrow$ task- specific similarity
- nonlocal learning algorithms $\Rightarrow$ can generalize even with few training samples (Bengio and Larochelle, 2006), can guess density shape near never seen examples, even though no explicit and specific prior knowledge is used.

Partial 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. and Larochelle, H. (2006). Non-local manifold parzen windows. In *Advances in Neural Information Processing Systems 18*. MIT Press.

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

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

Snapp, R. R. and Venkatesh, S. S. (1998). Asymptotic derivation of the finite-sample risk of the k nearest neighbor classifier. Technical Report UVM-CS-1998-0101, Department of Computer Science, University of Vermont.