About the locality of kernels in high-dimensional spaces

Damien Francois[1], Vincent Wertz[1], and Michel Verleysen[2]

[1] UCL - CESAME
Avenue G. Lemaitre, 4
B-1348 Louvain-la-Neuve, Belgium
(e-mail: {francois,wertz}@auto.ucl.ac.be)

[2] UCL - Machine Learning Group
Place du Levant, 3,
B-1348 Louvain-la-Neuve, Belgium
(e-mail: verleysen@dice.ucl.ac.be)

Abstract. Gaussian kernels are widely used in many data analysis tools such as Radial-Basis Function networks, Support Vector Machines and many others. Gaussian kernels are most often deemed to provide a local measure of similarity between vectors. In this paper, we show that Gaussian kernels are adequate measures of similarity when the representation dimension of the space remains small, but that they fail to reach their goal in high-dimensional spaces. We suggest the use of p-Gaussian kernels that include a supplementary degree of freedom in order to adapt to the distribution of data in high-dimensional problems. The use of such more flexible kernel may greatly improve the numerical stability of algorithms, and also the discriminative power of distance- and neighbor-based data analysis methods.

Keywords: High dimensional spaces, Local Models, Gaussian Kernels.

1 Introduction

Data analysis is one of the areas where artificial neural networks and machine learning techniques in general, have the most impact. During the last twenty years, there has been a considerable effort to develop data analysis techniques that are adapted to the abundance of data in nowadays information society. Although those tools are different in many aspects, be it from theoretical, technical or historical point of view, many of them share a common characteristic: for one reason or another, they use kernels. This is for example the case for Radial-Basis Function Networks (RBFN) [Bishop, 1995], for Support Vector Machines (SVM) [Cristianini and Shawe-Taylor, 2000], but also

[1] The work of D. Francois is funded by a grant from the Belgian FRIA. Part of the work of D. Francois and of the work of V. Wertz is supported by the Interuniversity Attraction Pole (IAP), initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.

[2] M. Verleysen is a Senior Research Associate of the Belgian F.N.R.S. (National Fund For Scientific Research).

for the more traditional Parzen estimators of probability densities [Parzen, 1962], for mixtures of Gaussians [McLachlan and Peel, 2000], etc.

Kernels can be defined in various ways. In most cases however, Kernel means a function whose value only depends on a distance between the input and a constant, named center; the input and the center may be vectors. Naturally, the most often used kernel is the Gaussian one. There are several good justifications to using Gaussian kernels. The first one is that the Gaussian function is a natural one: by the Central Limit Theorem, the sum of independent variables having the same distribution, whatever the distribution is, tends to a Gaussian distribution as the number of terms in the sum tends to infinity. The Gaussian function or distribution is also the only one that can be described without loss of information by its two first moments; it is therefore of particular interest for second-order statistics, including all linear data analysis methods.

Besides these general considerations, Gaussian kernels are most often used for their locality property: it is obvious that the Gaussian output may be considered as high when the input is close from the center and low (or even negligible) when the argument is far from the center. Locality is a primary importance concept for many reasons that range from the interpretability of the models to their numerical stability, through experimentally observed advantages with specific types of data.

This paper aims to show that the use of Gaussian kernels may be valid when the data are represented in low-dimensional spaces, but fails to reach its objectives in high-dimensional spaces. It is shown that high-dimensional Gaussian kernels are usually not local, and cannot be made local through scaling factors. This paper suggests using the so-called Generalized p-Gaussian kernel, which can be made local in any-dimensional space through the adaptation of a supplementary parameter.

This paper is organized as follows. Section 2 briefly recalls why the concept of locality is important in data analysis methods. Section 3 shows that Gaussian kernels are not local functions in high-dimensional spaces. Finally, in Section 4 Generalized p-Gaussian kernels are introduced as a possible alternative to Gaussian kernels for high-dimensional data analysis methods.

2 Why is locality so important?

While the locality property seems important in many algorithms, few papers address the reasons why it is indeed important. In the following, some intuitive arguments in the favor of local kernels are developed, without any attempt to be exhaustive.

2.1 Interpretability

The main argument for locality is interpretability. In most if not all applications, practitioners are not happy about responses given by blind models, i.e.

models that do not provide interpretability of their outputs. Nevertheless several algorithms are mostly blind, or at least have the reputation to be blind; examples are feed-forward artificial neural networks such as the Multi-Layer Perceptron (MLP) and RBFN. Interpretation in the latter models can however come from an examination of their hidden units outputs.

Indeed, the Kernel function can be seen as measure of similarity. The range of the kernel is between zero and one (note that when kernels are used for density estimation they are normalized so that their integral equals one ; this is not the case here. In any case, scaling does not change the arguments below). An input may be considered as close to the kernel center when the kernel output is near 1, and far when the output is near 0 ; indeed the ouptu of a Gaussian kernel decreases from 1 to 0 according to a negative exponential of the squared Euclidean distance between the vectors. Kernels may then be used to express in a numerical form the intuitive notion of closeness, i.e. similarity, with the continuity and derivability properties that are necessary in most algorithms. Regions spanned by kernels up to the limits defined (in afuzzy way) by the notion of closeness may help to the interpretation of the model.

The closeness concept is essential in local mmodels. For instance, RBFN and SVM models build the output corresponding to a new input x as a weighted sum of the output values associated to certain entities living in the input space (respectively called centroids and support vectors) ; while the weight is the similarity measurement between x and those entities. In other words, the more similar the new input is to a given entity, the more importance that entity has in computing the predicted value. Many Lazy Learning methods can be interpreted this way too.

2.2 Numerical stability

For RBFN as for SVM, the values of each kernel at each data point is gathered into a matrix which is used to formulate the corresponding optimization problem. The conditioning, and thus the sensitivity and numerical stability of the problem, depends on the condition number of that matrix. This section illustrates the fact that building a kernel-based model leads to an ill-formed optimization problem when locality of the kernels is not ensured.

Suppose N points randomly drawn according to a uniform distribution in the $[0,1]^d$ d-dimensional cube. A vector quantization is then performed on these N points to obtain M centroids, representative on the initial distribution. A traditional RBFN learning consists in placing Gaussian kernels on each of the M centroids, and evaluating the scalar RBFN output as a linear combination of the kernel outputs [Hwang and Bang, 1997]. The M linear coefficients are found by least squares. The matrix of the system is the $N \times M$ matrix built by evaluating each kernel on each data point. It is known that the numerical stability of the system depends on the condition

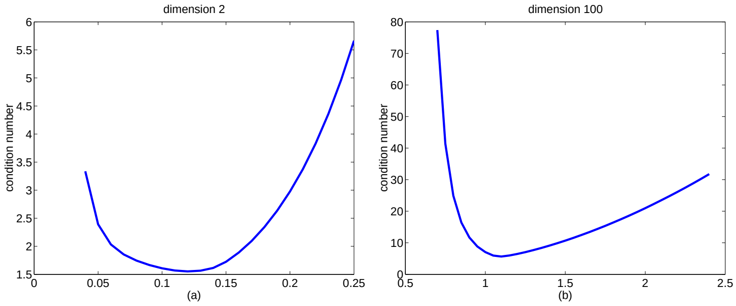

Fig. 1. Condition number of the system matrix in a RBF network, with respect to the standard deviation (width) of the kernels.

number of the matrix, which is defined as the ratio between the largest and smallest singular value of the matrix [Golub and van Loan, 1996].

Figure 1 shows an example of this condition number, for a system with 200 data points and 10 centroids. The condition number is plotted versus the common standard deviation (width) of the kernels. As the centroids learned by vector quantization have a distribution equal to the distribution of the initial data, i.e. they are uniformly distributed, it is natural to assume that all kernel standard deviations are equal.

One can see on Figure 1 that an optimum exists in the condition number of the system matrix, corresponding to an optimal standard deviation. While the exact value does not matter here, one can easily see that deviations much smaller or larger than the optimal lead to ill-conditioned matrices.

If the standard deviation is too small, the Gaussian kernels will not reach (with a significant value) the data points, even those that are close to the centroids. Very large coefficients will thus result from the system solution, both in positive and negative values, in order to both include all data into the radius of attraction of at least one Gaussian kernel, and at the same time keeping a weighted sum into a small range (corresponding to a smooth function to approximate).

On the contrary, if the standard deviation is too large, the Gaussian kernels will be very flat, leading to having most or all points into their respective radius of attraction. Approximating a smooth but non-flat (constant) function therefore also results in very large, both positive and negative, model coefficients.

Both situations therefore lead to ill-defined systems. Locality (not too large standard deviation) is thus also important for the numerical stability of the algorithms. Of course, too narrow kernels should be avoided too, as this corresponds to a kind of overfitting.

3 Gaussian kernels are not adequate in high-dimensional spaces

At first sight, the objective, i.e. measuring the similarity between two vectors, and the way to reach the goal, i.e. using a Gaussian kernel, perfectly match. However, without reference to Gaussian kernels, one could define an ideal kernel as a kernel whose output gives an acceptable measure of the similarity between two vectors; acceptable means for example that among a finite distribution, the closest vectors to a query should be evaluated as similar to the query, while vectors that are far from the query should be evaluated as non similar. In other words, among a finite distribution, the selected similarity measure should be able to find in acceptable proportions both similar and non similar vectors to a query point. In the next section, it will be shown that Gaussian kernels fit with this definition in low-dimensional spaces, while they do not fit it in high-dimensional spaces. To illustrate this problem, let us imagine that data have a Gaussian distribution centered at C (the following is qualitatively valid for any distribution though). We will compare the distribution of distances between any point and C, to the shape of a kernel centered on C too. As the kernel will be used to assess if points are close or not from C, this experiment allows to verify that the kernel is discriminative (is not too flat) in the effective range of the distance distribution. On Figure 2, the thick line represents the kernel value, while the thin line (and grayed area) represent the distance distribution. One easily sees on graphs (a) and (b) that, in low dimension, for a well-chosen kernel width value, the small (resp. large) distances in the distribution will be mapped onto kernel values close to one (resp. zero). This matches the definition of an ideal kernel as detailed in the previous paragraph.

However when the space dimension increases, the correspondence between the range of distances in the histogram, and the range of the decreasing slope in the Gaussian kernel cannot be guaranteed anymore. Graphs (c) and (d) refer to space dimensions 10 and 100 respectively, for several kernel width values. It is seen that it is more difficult to adjust the value of the kernel width is in order to cope with the ideal kernel definition: in all cases, there is a large part of the Gaussian kernel decreasing slope that falls out of the range of distances in the histogram. This means that close distances (left queue of the distribution) and large distances (right queue of the distribution) are hardly distinguishable from their kernel values; the notion of similarity itself (are data close or far one from another) looses its significance. Needless to say, the consequences in methods based on nearest neighbors are dramatic.

Another view of the same phenomenon comes from the following experiment. Let us imagine a d-dimensional uniform distribution, quantized into a predefined number M of centroids. A Gaussian kernel is centered on each initial point of the distribution; the kernel is evaluated on the furthest and closest centroids. Then the difference between the two Gaussian outputs is taken, and averaged over all points of the distribution. The result is repre-

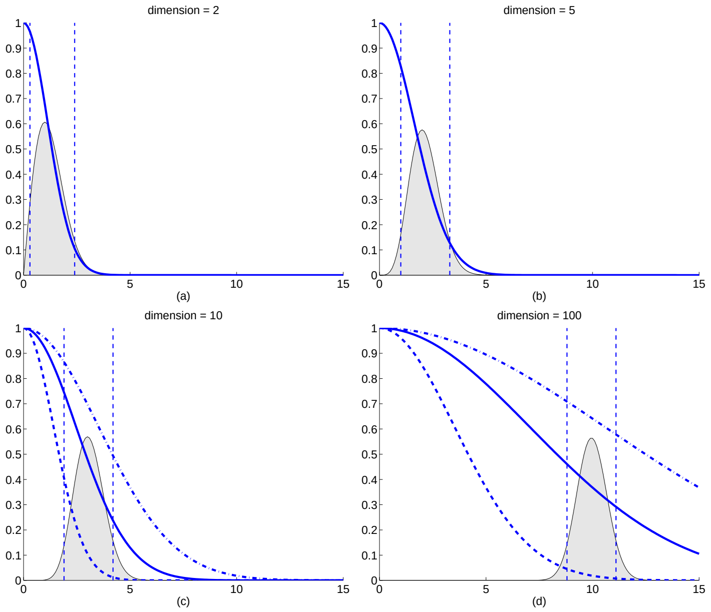

Fig. 2. Kernel values as a function of the distance to their centers for several space dimensions, along with the distribution of distances for normally distributed data. Vertical lines correspond to 5 and 95 percentile resp.

sentative of the contrast between the similarity of a point to its closest and furthest away centroids; if the contrast is large, a model built with such a kernel can be considered 'local' ; if it is small, the notion of neighborhood looses its significance.

Figure 3 shows this contrast with respect to the width of the kernels. In dimension 2 (left), the contrast is close to 1 for a well-chosen value of the kernel; distances are easily distinguishable. Note that the ideal kernel standard deviation is relatively small, which corresponds to a kernel having a local character. In dimension 100, the contrast hardly reaches 0.2; distances are far less distinguishable, whatever the kernel standard deviation is.

4 Recovering locality in HD spaces

The necessity to more or less span the effective range of distances between data in a real distribution setting, by the effective part of the kernel (i.e. the part with the decreasing slope), requires to add a parameter with respect to the Gaussian kernel. Besides the width that controls the slope of the kernel, there is a need for a supplementary parameter that controls the smallest distance corresponding to the decreasing part of the kernel. An example of

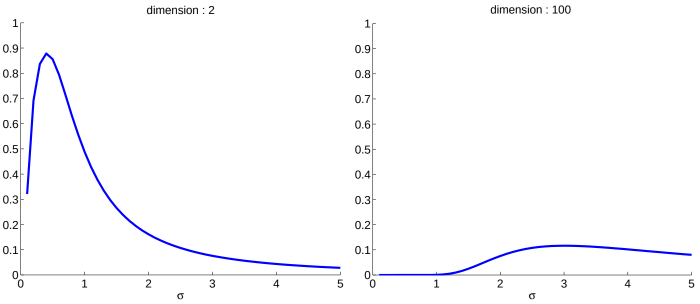

Fig. 3. Contrast (see definition in text) in a 2-dimensional (left) and a 100-dimensional (right) uniform distribution, with respect to the kernel standard deviation.

kernel that fulfills this requirement is the p-Gaussian kernel :

$$K(x, y) = \exp(-d(x, y)^p / \sigma^p),$$

where p and σ are the two parameters. Normalizing coefficient for density estimation can be found in [Kassam, 1988], but once again this is not needed for measuring similarities. Figure 4 (left) shows an example of p-Gaussian kernel, width $p = 11$ and $\sigma = 4.3$. It is seen that the kernel slope effectively covers the range of distances, according to the definition of ideal kernel

The method to set adequate values to p and σ can easily be deduced from the same requirements. As the decreasing slope of the kernel has to cover the effective range distances in the histogram built on the sample distribution, two equations can be deduced once this range is known: one for the lowest value of the range, one for the highest one. Of course, as we are speaking about distributions, taking extreme values is not a good idea; rather, for example, the 5% and 95% percentiles of the distribution should be estimated. Let d_N and d_F be these two values respectively. Then two equations can be written by making the p-Gaussian kernel evaluated at d_N (resp. d_F) equal to 95% (resp. 5%) of the full kernel range :

$$p = \frac{\ln\left(\frac{\ln(0.05)}{\ln(0.95)}\right)}{\ln \frac{d_F}{d_N}} \quad ; \quad \sigma = \frac{d_N}{(-\ln(0.05))^{1/p}} = \frac{d_F}{(-\ln(0.95))^{1/p}}$$

Figure 4 (right) shows the results of the experiment described earlier to estimate the contrast, with respectively, the Gaussian kernel and a kernel with optimized p and σ values.

5 Conclusion

Local kernels or functions are used in many data analysis paradigms and algorithms, such as Radial-Basis Function networks, Support Vector Machines,

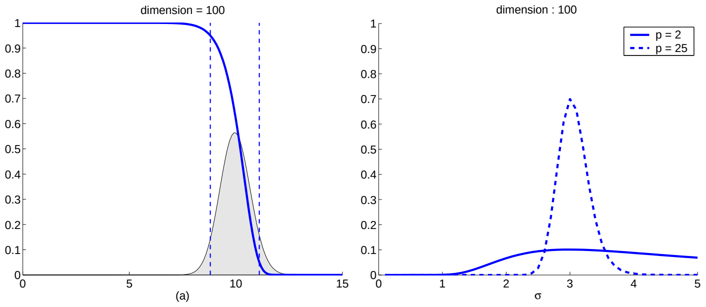

Fig. 4. (left) Kernel values along with distance distribution for the ideal kernel, $p = 11$; (right) contrast for Gaussian kernel and ideal kernel in dimension 100.

some Vector Quantization methods, etc. Locality is used as a way of interpretation, and also to provide measures of similarities between data. In this paper, we show that the widely used Gaussian Kernel is appropriate to represent similarities in low-dimensional spaces, but fails to fulfill this goal in high-dimensional ones. When similarities cannot be expected anymore to be measured adequately, many problems may be expected, for example in nearest neighbor search. The numerical stability of the methods may be lost.

p-Gaussian kernels are presented as an alternative to Gaussian kernels. An additional parameter makes it possible to keep the effective part of the Gaussian slope in the effective part of the distribution of distances between data. In this way, p-Gaussian kernels will adequately discriminate small and large distances between pairs of data even in a high-dimensional setting, a task that Gaussian kernel fails to fulfill. A methodology is presented to set the parameters according to a specific data sample. Future work will consist in using such flexible kernels in learning algorithms for high-dimensional data.

References

[Bishop, 1995]C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford university press, 1995.

[Cristianini and Shawe-Taylor, 2000]N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[Golub and van Loan, 1996]G. H. Golub and C. F. van Loan. *Matrix Computations.* Johns Hopkins University Press, 3rd edition, 1996.

[Hwang and Bang, 1997]Y.-S. Hwang and S.-Y. Bang. An efficient method to construct a radial basis function neural network classifier. *Neural Networks*, 10(8):1495–1503, 1997.

[Kassam, 1988]S. A. Kassam. *Signal Detection in Non-Gaussian Noise.* Springer-Verlag, 1988.

[McLachlan and Peel, 2000]G. J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, 2000.

[Parzen, 1962]E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.