Scale-Invariance of Support Vector Machines based on the Triangular Kernel

François Fleuret Hichem Sahbi

IMEDIA Research Group

INRIA

Domaine de Voluceau

78150 Le Chesnay, France

Abstract

This paper focuses on the scale-invariance of support vector machines using the triangular kernel, and on their good performance in artificial vision. Our main contribution is the analytical proof that, by using this kernel, if both training and testing data are scaled by the same factor, the response of the classification function remains the same. We compare the performance of this kernel to those of a standard Gaussian on face detection and handwritten character recognition.

1. Introduction

For a decade now, Support Vector Machines [3] have proven to be generic and efficient tools for classification and regression. SVMs got their popularity both from

a solid theoretical support [2, 11], and because they clearly untie the specification of the model from the training. The former corresponds to the choice of the underlying kernel, and the later can be done optimally with classical quadratic optimization methods.

We study in this paper the invariance to scale of the induction process based on SVM using the triangular kernel. What we prove theoretically is that, if one scales the training set by a certain factor, the learned classification function is scaled by the same factor. Such an invariance property is not true in general. The estimation of a scaling parameter for the Gaussian or other standard kernels is done through a tedious cross-validation process. This implies repeating several times the complete training of an SVM, usually on very large databases.

Note that the invariance to scale we consider in this paper neither requires or implies the invariance of the kernel itself. Actually, as we will see, the triangular kernel is trivially not scale-invariant.

The standard feature space used in artificial vision (raw pixel representation or wavelet coefficients) are linear transformations of the pixel intensities. Thus invariance to scale ensures an invariance to linear transformations of the gray levels of the picture population.

Beside those theoretical results, we also give experimental evidences of the good performances of an SVM based on the triangular kernel.

In §2 we summarize the standard formalization of SVMs, and we present the triangular kernel. In §3 we show analytically how induction based on an SVM using the triangular kernel is invariant to scaling. Then, we give in §4 results on a simple 2D problem and on real-world tasks to illustrate what this invariance means and to show how it improves the generalization performance.

2. Support Vector Machines

2.1. Standard Formalization

We will focus on SVMs for classification. Basically, SVM methods project data to classify in a space of large (possibly infinite) dimension, where a linear criterion is used. For any training set, one can choose an appropriate projection Ψ so that linear separability may be achieved. Computation is done without an explicit form of the projection, but only with the kernel corresponding to the scalar product between projections.

The model is thus specified by choosing the kernel k:

$$k(x,\ x') = \langle \Psi(x),\ \Psi(x') \rangle$$

And a function f which sign is the predicted class:

$$f(x) = \langle \omega,\ \Psi(x) \rangle\ +\ b$$

Let X be a random variable on R^N standing for the feature distribution of data to classify (in the original space), and Y on $\{-1,\ 1\}$ the real class of the data. We will denote $\mathcal{T}\ =\ \{(x_1, y_1), \ldots, (x_n, y_n)\}$ a training set generated i.i.d according to (X, Y). The computation of ω is achieved by minimizing $||\omega||$ under correct classification of the training set, i.e. $\forall i,\ y_i\ f(x_i)\ \geq\ 1$. This is equivalent to maximizing the margin between training points and the separating hyper-plan (in the high-dimensional space), and ensures good generalization property on the real population [11]. It can be proven [2] that ω is of the form $\sum_i \alpha_i y_i\ \Psi(x_i)$, where the α_i come from the following quadratic optimization problem:

Maximize

$$L(\alpha) = \sum_i \alpha_i \; - \; \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \; y_i y_j \; k(x_i, \; x_j) \tag{1}$$

under

$$\forall i, \quad \alpha_i \; \geq \; 0 \quad \text{and} \quad \sum_i \alpha_i y_i \; = \; 0 \tag{2}$$

The value of b does not appear in the optimization, and it has to be computed, given the α_i:

$$b = \frac{\min\limits_{y_i=-1} \sum\limits_j \alpha_j k(x_i, x_j) - \min\limits_{y_i=+1} \sum\limits_j \alpha_j k(x_i, x_j)}{2}$$

Finally, using the expansion of ω, we can write the classification function as:

$$f(x) = \sum_i \alpha_i \; y_i \; k(x_i, \; x) \; + \; b$$

2.2. Triangular kernel

The classification power of SVMs comes directly from the complexity of the underlying kernel. Many kernels can be used[9, 13], the most standard being the Gaussian [10]:

$$k(x, \; x') = \exp\left(-||x - x'||^2 / \sigma^2\right)$$

The parameter σ in this kernel is directly related to scaling. If it is overestimated, the exponential behaves almost linearly and it can be shown that the projection into the high-dimensional space is also almost linear and useless[4]. On the contrary, when underestimated, the function lacks any regularization power and the decision boundary is jagged and irregular, highly sensitive to noisy training data (see figure 2, middle row). Several methods have been developed to estimate an optimal σ, so that the whole process would be invariant to scaling [4].

We focus here on a less standard kernel referred to as the unrectified triangular kernel:

$$k_{\mathrm{T}}(x,\ x') = -||x - x'||$$

It has been shown in [1] that this defines a conditionally positive-definite kernel. This means that for any $x_1, \ldots, x_n$ and any $c_1, \ldots, c_n$ such that $\sum_i c_i = 0$, then $\sum_{i,j} c_i\, c_j\, k_{\mathrm{T}}(x_i,\ x_j) \geq 0$. Due to the equilibrium constraint $\sum_i \alpha_i\, y_i = 0$ in (2), this ensures that k_{T} can be used as a SVM Kernel [8].

3. Scale-invariance of the classification

3.1. Scaling of the triangular kernel

Even if the triangular kernel is not invariant to scaling, it still has an interesting weak property of invariance that we could describe as an invariance "in shape" (see figure 1). Given a scaling factor $\gamma > 0$ this weak invariance can be formally expressed as:

$$k_{\mathrm{T}}(\gamma\ x,\ \gamma\ x') = -\gamma\ ||x - x'||$$

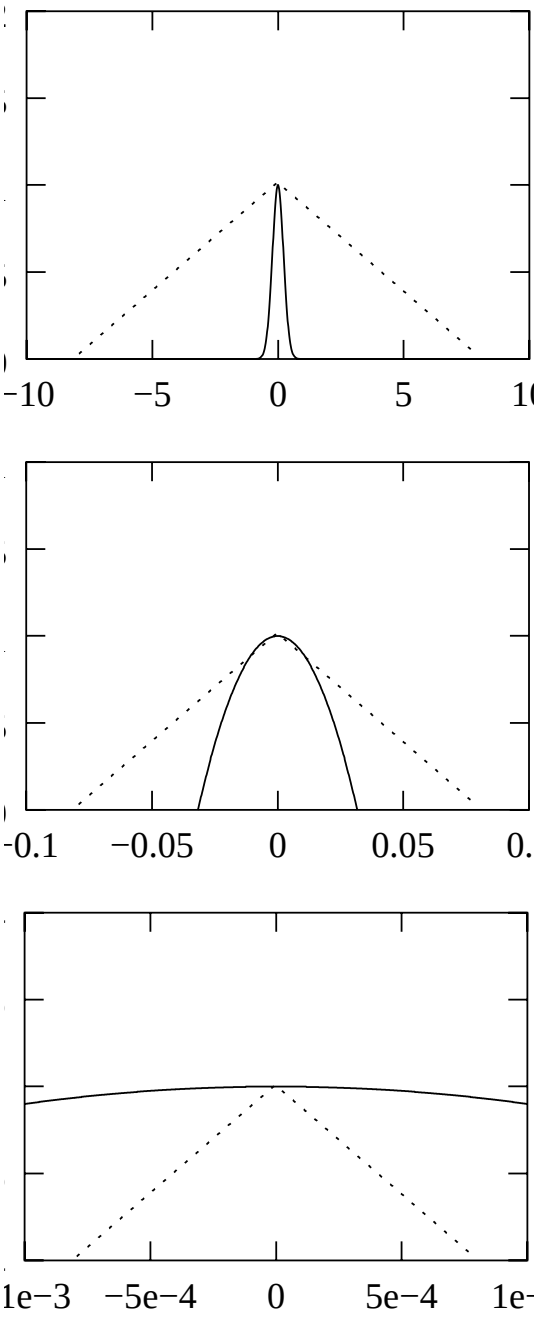

Figure 1: *Gaussian kernel (continuous line) and triangular kernel (dashed line) at various scales (top to bottom, respectively* $\times 10^0$*,* $\times 10^2$ *and* $\times 10^4$*). Intuitively, whereas the triangular kernel is identical in shape at all scales, the Gaussian kernel has different shapes, from a Dirac-like to a uniform weighting of the neighborhood.*

$$= \gamma \; k_{\mathrm{T}}(x, \; x')$$

Thus, when the points are scaled by a certain factor γ, the value of the kernel scales by γ.

3.2. Invariance of the classifier

In the following, we consider a situation where we scale the data by a factor $\gamma > 0$. Let's denote $\mathcal{T}^\gamma = \{\gamma \; x_1, \; \ldots, \; \gamma \; x_n\}$ a training set for that population. We denote f^γ the classification function obtained by training the SVM on $\mathcal{T}^\gamma$ (thus, f^1 is the classifier built from the data at the original scale). We will show the main following result:

$$\forall x, \; f^\gamma(\gamma \; x) = f^1(x)$$

Let α^γ, ω^γ and b^γ be the parameters of the classification function estimated on $\mathcal{T}^\gamma$. We have:

$$f^\gamma(x) = \sum_i \alpha_i^\gamma \; y_i \; k_{\mathrm{T}}(\gamma \; x_i, \; x) \; + \; b^\gamma$$

Thus, the α_i^γ come from the minimization system corresponding to $\mathcal{T}^\gamma$:

Maximize

$$L^\gamma(\alpha^\gamma) = \sum_i \alpha_i^\gamma - \frac{1}{2} \sum_{i,j} \alpha_i^\gamma \; \alpha_j^\gamma \; y_i y_j \; k_{\mathrm{T}}(\gamma \; x_i, \; \gamma \; x_j)$$

under

$$\forall i, \quad \alpha_i^\gamma \;\geq\; 0 \quad \text{and} \quad \sum_i \alpha_i^\gamma \; y_i \;=\; 0$$

It follows:

$$\begin{aligned} L^\gamma(\alpha^\gamma) &= \sum_i \alpha_i^\gamma - \frac{\gamma}{2} \sum_{i,j} \alpha_i^\gamma \; \alpha_j^\gamma \; y_i y_j \; k_{\mathrm{T}}(x_i, \; x_j) \\ &= \frac{1}{\gamma} \left(\sum_i \gamma \alpha_i^\gamma - \frac{1}{2} \sum_{i,j} \gamma \alpha_i^\gamma \; \gamma \alpha_j^\gamma \; y_i y_j \; k_{\mathrm{T}}(x_i, \; x_j) \right) \\ &= \frac{1}{\gamma} \; L^1(\gamma \; \alpha^\gamma) \end{aligned}$$

Which leads to: $\forall i, \; \alpha_i^\gamma \;=\; \frac{1}{\gamma} \; \alpha_i^1$, and to the following equality, $\forall x$:

$$\begin{aligned} \sum_j \alpha_j^\gamma y_j k_{\mathrm{T}}(\gamma x, \; \gamma x_j) &= \sum_j \frac{1}{\gamma} \alpha_j^1 y_j (\gamma k_{\mathrm{T}}(x, \; x_j)) \\ &= \sum_j \alpha_j^1 y_j k_{\mathrm{T}}(x, \; x_j) \end{aligned}$$

Thus, we can easily show that $b^\gamma = b^1$. Finally we obtain our main result:

$$\begin{aligned} f^\gamma(\gamma x) &= \sum_i \alpha_i^\gamma \; y_i \; k_{\mathrm{T}}(\gamma \; x_i, \; \gamma \; x) \; + \; b^\gamma \\ &= \sum_i \alpha_i^1 \; y_i \; k_{\mathrm{T}}(x_i, \; x) \; + \; b^1 \\ &= f^1(x) \end{aligned}$$

3.3. Simple 2D classification problem

To illustrate the invariance to scale of the induction process, we have set up a simple classification task in two dimension. The original training population is a set of 512

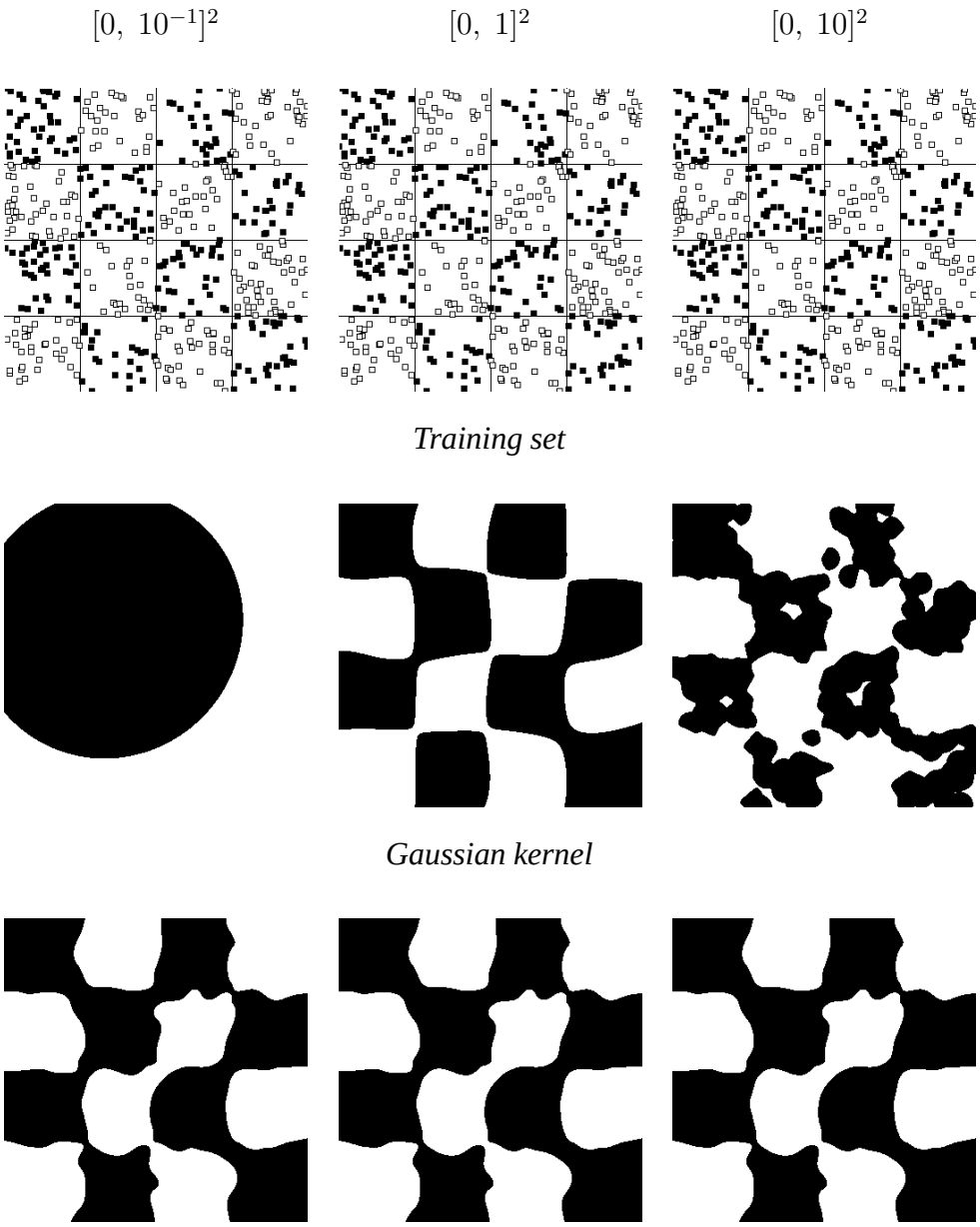

Figure 2: *A simple classification task in 2D. The upper row shows the training set scaled by three different factors. The figures are zoomed according to the same factors for ease of representation. The middle row shows the results of the classifications with a Gaussian kernel, and the lower row shows results with the triangular kernel.*

points, uniformly distributed in the unit square. The class of each of those samples is a deterministic function of its location in the square (see figure 2, upper row). From this sample, we have produced two others, one scaled down by a factor of 10, and the other scaled up by the same factor. We have built three SVMs based on a Gaussian kernel with $\sigma = 0.2$ on those three samples, and three SVMs based on the triangular kernel. Results are shown on figure 2.

As expected, the Gaussian kernel either smoothes too much (figure 2, middle row, left), is accurate (figure 2, middle row, center) or overfits (figure 2, middle row, right), while the triangular kernel behaves similarly at all scales.

4. Experiments

4.1. Face detection

The initial motivation for this study was to understand the good generalization performance of the triangular kernel in the context of face detection. Sahbi and Geman have developed a highly efficient detector based on a hierarchy of SVMs using the triangular kernel [7]. Their approach consists in building several SVMs dedicated to population of face pictures more and more constrained in position in the image plan.

We focus here on the generalization performances of individual classifiers dedicated to constrained populations of face pictures. Figure 3 shows some examples from two of them, the first less constrained than the second. Both are synthetically generated by doing affine bitmap transformations of the original pictures which are taken from the ORL database of faces[5].

Each picture is a 64×64 pixel in 256 gray levels and contains a face roughly centered. The distance between the eyes of each face is between 10 and 20 pixels. We

Figure 3: *Training samples of two face populations. The upper row shows samples from a loosely constrained population while the lower row shows a sample from a highly constrained population.*

Table 1: *Performance comparison between the triangular and the Gaussian kernel on the face vs. non-face classification problem.*

Kernel	Weak constraints	Hard constraints
Triangular	6.88%	0.69%
Gaussian ($\sigma = 10^3$)	7.36%	1.56%
Gaussian ($\sigma = 6.10^2$)	7.83%	0.90%
Gaussian ($\sigma = 10^2$)	21.14%	37.73%
Gaussian ($\sigma = 10$)	41.80%	37.73%

use as features a vector of 256 Haar wavelet coefficients. These simple Haar features allow us to capture the main facial details at various orientations and resolutions and can be computed efficiently using the integral image [12].

The results given here correspond to SVMs trained with 400 face pictures and 600 background images. Error rates are estimated on 400 other face pictures, verifying the same pose constraints, and 600 other background pictures.

As expected, the more the faces are constrained in pose, the easier is the classification, since tolerance to translation and rotation is no more expected from the SVM. Results on table 1 show the performance of both the triangular and the Gaussian kernel. While the Gaussian kernel relies heavily on the choice of σ, the triangular kernel achieves the same order of performances without tuning of any scale parameter.

Table 2: *Performance comparison between the triangular and the Gaussian kernel on handwritten digit recognition.*

Kernel	Error rate
Triangular	3.93%
Gaussian ($\sigma = 10^{-1}$)	35.87%
Gaussian ($\sigma = 1$)	5.18%
Gaussian ($\sigma = 10$)	6.89%
Gaussian ($\sigma = 100$)	20.68%

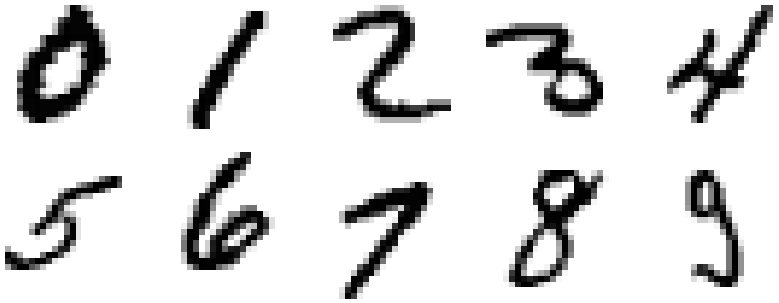

Figure 4: *Some handwritten digits from the MNIST database.*

4.2. Character recognition

This last experiment is a classical problem of handwritten digit recognition on the MNIST database [6]. This database contains $70,000$ black and white digit pictures of size 28×28 pixels (see figure 4). The features we use for that experiment are 64 Haar-wavelet coefficients, similar to the ones used for the face detection experiment.

We train ten SVMs, $f^{(0)}, \ldots, f^{(9)}$, each one dedicated to one of the digits. The training for each of them is done on $60,000$ examples and the testing is done on $10,000$ other images. For each picture, we consider as features 64 simple Haar-wavelet coefficients to gain local invariance to deformations. The final classifier F is based on a winner-take all rule: the result of the classification is the index of the

SVM with the highest response.

$$F(x) = \arg\max_i f^{(i)}(x)$$

Results are shown on table 2 for the Gaussian kernel at various σ and for triangular kernel.

5. Discussion

5.1. Ideal invariant training set

Another interesting property appears when we consider an hypothetical infinite two-dimensional spiral-shaped training set. Such an infinite set $\mathcal{T}$ could be built to be invariant to a certain mapping ρ, composition of a rotation and a scaling (this set would be an union of orbits of that mapping cf. figure 5). The training of an SVM with the triangular kernel would be also invariant under that transformation. So if we denote $f_{\mathcal{T}}$ (respectively $f_{\rho(\mathcal{T})}$) the classification function obtained by training on $\mathcal{T}$ (respectively on $\rho(\mathcal{T})$), as $\mathcal{T} = \rho(\mathcal{T})$ we would have

$$\forall x,\ f_{\mathcal{T}}(x) = f_{\rho(\mathcal{T})}(\rho(x)) = f_{\mathcal{T}}(\rho(x))$$

which means that the learned boundary itself would be invariant. That implies it would possess details at several scales at while. We do not have such an example in real, but we can still approximate that result by considering a finite spiral-shaped set. As it can be seen on figure 6, the boundary at the center has a finer scale far smaller than at the outer area.

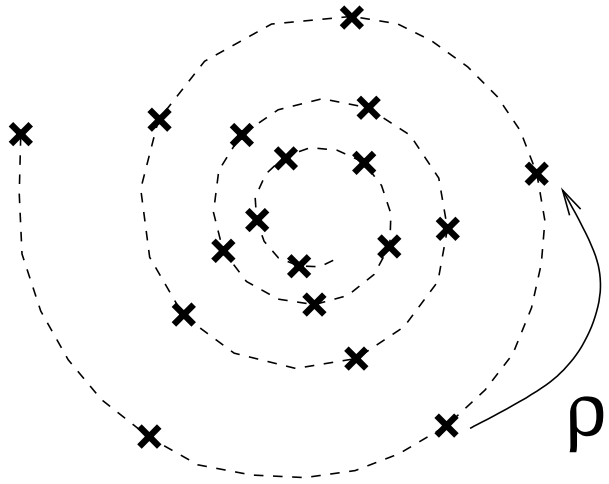

Figure 5: *The iterations of a mapping ρ, composition of a rotation and a scaling, generate an infinite set of point invariant under ρ.*

5.2. Soft margins

Soft margin SVM training consists in bounding the Lagrange coefficients α_i to control the influence of outliers on the learning process. Also, in many concrete situations, the range of values allowed for the coefficients is fixed by computer representations of real numbers. With such a bounding, the theoretical invariance to scale would not hold anymore.

Nevertheless, our main result shows that, with the triangular kernel k_{T}, the coefficients are proportional to the inverse of the scaling of the population. Such a linear increase is very reasonable and lead to values that could be handled without bounding in all our experiments.

6. Conclusion

We have shown in this article that classification with SVMs based on the triangular kernel is invariant to the scaling of the data. Therefore, using this kernel avoids

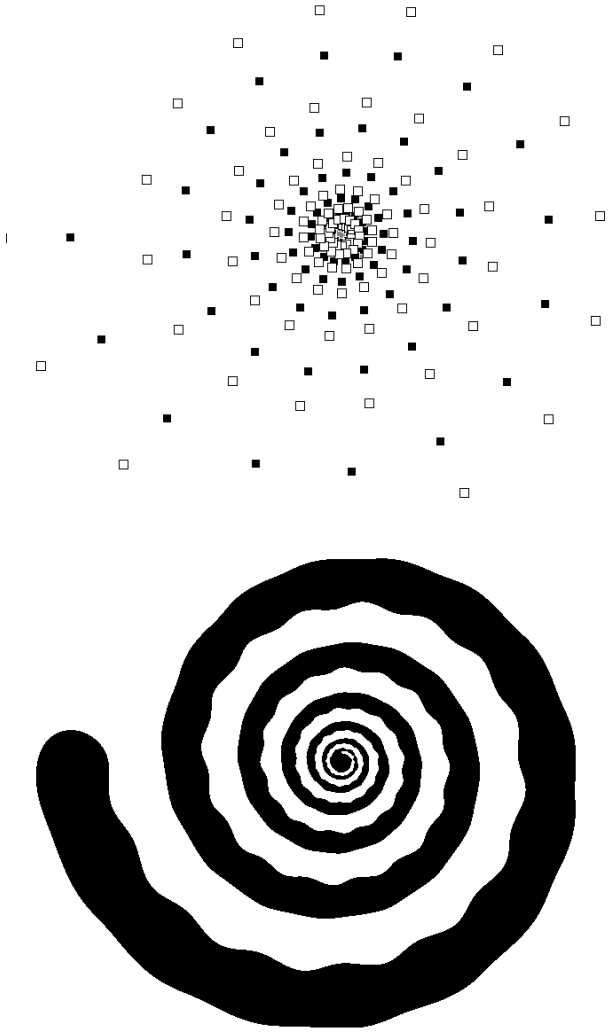

Figure 6: *The triangular kernel can separate two populations, even if it requires various scales. Training set is shown on the top, and classification with the triangular kernel is shown below.*

the estimation of an optimal scaling parameter. Such an estimation is usually based on cross-validation and is computationally intensive, since it requires to run several times the complete training process. Experiments demonstrate the very good performances of this kernel on real data, compared to those of the usual Gaussian kernel, even when the scale parameter of the later is optimized.

Our forthcoming works will study this specific property and more generally the good generalization performance of the triangular kernel.

References

[1] C. Berg. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer Verlag, 1984.

[2] B. E. Boser, I. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. *in Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 5:144–152, 1992.

[3] N. Christiani and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[4] N. Cristianini, C. Campbell, and J. Shawe-Taylor. Dynamically adapting kernels in support vector machines. In *Proceedings of NIPS*, volume 11, 1998.

[5] http://www.uk.research.att.com/facedatabase.html.

[6] http://yann.lecun.com/exdb/mnist/.

[7] H. Sahbi, D. Geman, and N. Boujemaa. Face detection using coarse-to-fine support vector classifiers. In *Proceedings of ICIP*, 2002.

[8] B. Scholkopf. The kernel trick for distances. In *Proceedings of NIPS*, pages 301–307, 2000.

[9] B. Scholkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[10] B. Scholkopf, K. K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transcations on Signal Processing*, 45:11:2758–2765, 1997.

[11] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1998.

[12] P. Viola and M. Jones. Robust real-time object detection. Technical Report 1, Compaq Cambridge Research Lab, 2001.

[13] G. Wahba. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. Technical Report 984, University of Wisconsin, Madison, 1997.