Random Projections and α-Shape to Support the Kernel Design (Student Abstract)

Daniel Moreira Cestari,[1] **Rodrigo Fernandes de Mello**[1*]

[1]Department of Computer Science, University of São Paulo
400 Avenida Trabalhador São Carlense
São Carlos, São Paulo, Brazil 13566-590
daniel.cestari@usp.br, mello@icmc.usp.br

Abstract

We demonstrate that projecting data points into hyperplanes is good strategy for general-purpose kernel design. We used three different hyperplanes generation schemes, random, convex hull and α-shape, and evaluated the results on two synthetic and three well known image-based datasets. The results showed considerable improvement in the classification performance in almost all scenarios, corroborating the claim that such an approach can be used as a general-purpose kernel transformation. Also, we discuss some connection with Convolutional Neural Networks and how such an approach could be used to understand such networks better.

Random projection (RP) guarantees minimal distortions in terms of pairwise distances when projecting points into a low dimensional Euclidean space, making it a powerful technique for dimensionality reduction (Xie, Li, and Xue 2017). Instead of applying RP in such typical scenario, in this paper we project data points onto hyperplanes in order to approximate a proper decision boundary, thus supporting the design of kernels from data.

The Statistical Learning Theory (SLT) provides tighter guarantees on the generalization bound when using a suitable bias (de Mello and Ponti 2018), which we intend to achieve when designing kernels to approximate decision boundaries. Nonetheless, such general-purpose design is still an open problem (Rojo-Álvarez et al. 2018) that is typically addressed by using trial-and-error approaches (Shawe-Taylor and Cristianini 2004; Scholkopf and Smola 2002), besides remaining one of the greatest challenges to be faced by the ML area. If we succeed in designing such a suitable kernel for specific problems, stronger generalization guarantees are expected, thus making learning more robust to general-purpose nonlinear classifiers.

We realized that hyperplanes derived from linearized decision boundaries could be used to project data points and estimate kernels. Such projections are seen as the explicit transformations a kernel would provide; hence, they represent new features to support the representation of classes. As consequence, in the absence of class overlap, a linear classifier can correctly classify all samples.

The concept of how RP is connected with the hyperplanes estimating the decision boundary is illustrated in Figure 1-(a) for the Banana dataset. During the process of plotting all decision boundary hyperplanes, we realized such a boundary could be obtained from random hyperplanes. In fact, it turned out to be a much simpler problem when using random hyperplanes to project data points onto. Leading to practical applications since it is based on the highly parallelizable inner-product operation.

Our approach consists of adding the sign projection of each hyperplane as a new attribute to the original input space. All those new features are then used to uniquely determine the class of each sample when there is no class overlap, and the decision boundary is approximated by a subset of such hyperplanes. The transformation illustrated on the right of Figure 1-(a) represents the following expression:

$$\Phi([x_i]) \mapsto [x_i, \mathrm{sgn}(\langle n_1, x_i\rangle), \dots, \mathrm{sgn}(\langle n_k, x_i\rangle)], \qquad (1)$$

where x_i is an input space point, n_i represents the normal vector of each hyperplane, and *sgn* is the sign function.

In addition to random hyperplanes, we also used the strategy of convex hull and its generalization, a.k.a. α-shape, as alternative methods to produce hyperplanes. Given the way they are generated, they turn out to be more adherent to the class instances of data samples. The closure formed to represent each class allows the generation schemes of hyperplanes to cope with class overlapping. Each hyperplane normal vector produces different projections per class, thus being capable of separating instances under different labels.

Our hypothesis is that the decision boundary can be composed of a subset of all produced hyperplanes. Because of the limited magnitude of the sign projection, we conjecture they do not jeopardize the classification task due to the curse of dimensionality, even if several hyperplanes are not useful in a particular separability scenario. Furthermore, there is a clear connection of our approach with Deep Learning, most specifically with Convolution Neural Networks (CNN).

In order to assess our approach, we used two bidimensional and well-known toy datasets, they show nonlinear data behaviors and compelling qualities that real-world datasets usually have, in both we also added several levels of Gaussian noise. Figure 1-(b) illustrates those two toy datasets, the first comprises the Banana set and the second is the Concentric Circles. In order to complement our study, we decided to assess other three typical classifications bench-

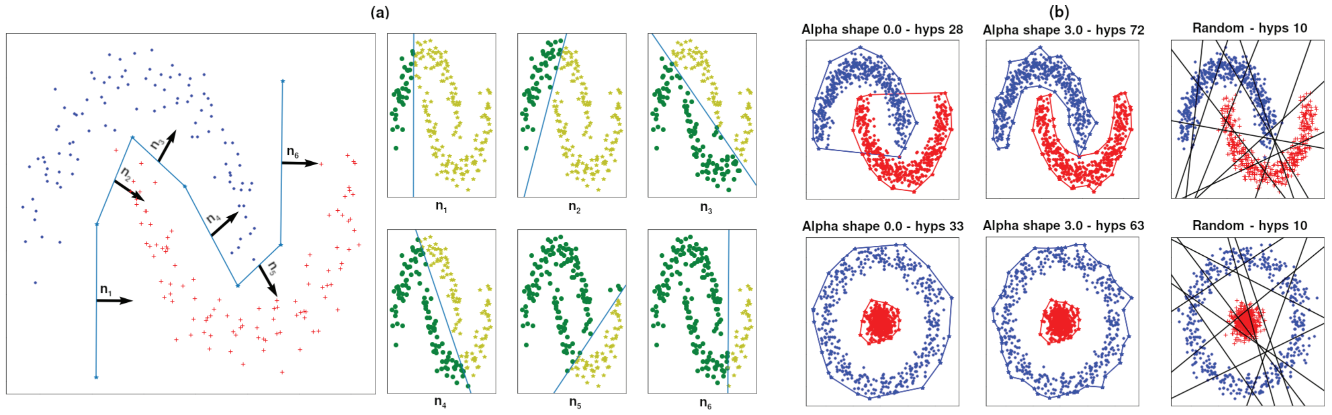

Figure 1: (a) - On the left, linear estimate of the decision boundary on the Banana toy dataset. The arrows define the normal vector of each hyperplane. On the right, the sign of the projection of the data point on each hyperplane: yellow points are aligned with the normal vector, and green ones are opposite to the normal vector of each hyperplane. (b) - Illustration of the generation methods of hyperplanes. Each column represents a different process to generate hyperplanes, the first one uses $\alpha = 0$, thus producing the convex hull, the second uses $\alpha = 3$, and the third randomly generates ten hyperplanes. At the first row, procedures are applied on the Banana dataset and at the second on the Concentric circles.

marks adopted in the deep learning literature (e.g., MNIST, CIFAR-10, and STL-10).

First, we produce hyperplanes on the original input space, then data points are projected into such linear boundaries taking only their signs. Therefore, signs are seen as new features added to the original dataset. Finally, using several classification algorithms, like SVM, we evaluate the class separability performance. We firstly assessed the randomly generated hyperplanes; then, the closure produced by the convex hull around each class, and at last, the α-shape.

Figure 1-(b) depicts the two schemes of generating hyperplanes on the toy datasets. We conclude that when one of the classes encloses the other or when the boundary forms a curve, the α-shape is the most suitable option. The in-sample noise makes the α-shape prone to produce more hyperplanes even when there is a clear class separability.

The proposed approaches (i.e., random, convex hull or α-shape) provided good performances in the general case at a low computational cost. We intend to further investigate which hyperplane generation procedure is more suitable, and what is the adequate number of hyperplanes. In that sense, we have already started the study of theoretical lower limits on the number of hyperplanes.

The overall results support the hypothesis that the projection approaches can significantly improve the classification performance. The synthetic scenario allowed us to assess the robustness and low parameter sensitivity (robustness) for both linear classifiers used (Perceptron and SVM), even in the presence of significant class overlapping, and they always presented the best or close to the best outcome. Although the SVM did not present better performances than the Perceptron on the image-based datasets, it confirmed stable results over the number of hyperplanes, even producing fewer support vectors in case of MNIST.

The connection between random hyperplanes and CNN (masks randomly initialized) can be used to reason about CNN inner working and good performances reported in the literature. Since, at times, the results on the convex hull and α-shape were similar to the random hyperplanes, it might be the case that they are also connected to CNN, we still plan to closer examine possible theoretical foundations to explain such connections. Finally, we reiterate the usefulness of such a hyperplane projection approach as a general-purpose to support the kernel design for classification tasks.

Acknowledgments

We would like to thank the Brazilian funding agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Google Research Awards for Latin America (LARA) for financial support. This research was carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

de Mello, R. F., and Ponti, M. A. 2018. *Machine Learning: A Practical Approach on the Statistical Learning Theory*. Springer.

Rojo-Álvarez, J. L.; Martínez-Ramón, M.; Marí, J. M.; and Camps-Valls, G. 2018. *Digital signal processing with Kernel methods*. Wiley Online Library.

Scholkopf, B., and Smola, A. J. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shawe-Taylor, J., and Cristianini, N. 2004. *Kernel methods for pattern analysis*. Cambridge university press.

Xie, H.; Li, J.; and Xue, H. 2017. A survey of dimensionality reduction techniques based on random projection. *arXiv preprint arXiv:1706.04371*.