Explaining a result to the end-user: a geometric approach for classification problems

I. Alvarez, S. Martin

Explaining a Result to the End-User: A Geometric Approach for Classification Problems

Isabelle Alvarez[1,2] and Sophie Martin[2]

[1] LIP6, UPMC, Paris, France,
isabelle.alvarez@lip6.fr
[2] Cemagref, LISC, Aubière, France

Abstract. This paper addresses the issue of the explanation of the result given to the end-user by a classifier, when it is used as a decision support system. We consider machine learning classifiers, which provide a class for new cases, but also deterministic classifiers that are built to solve a particular problem (like in viability or control problems). The end-user relies mainly on global information (like error rates) to assess the quality of the result given by the system. Even class membership probability, if available, describes only the statistical viewpoint, it doesn't take into account the context of a particular case. In the case of numerical state space, we propose to use the decision boundary of the classifier (which always exists, even implicitly), to describe the situation of a particular case: The distance of a case to the decision boundary measures the robustness of the decision to a change in the input data. Other geometric concepts can present a precise picture of the situation to the end-user. This geometric study is applied to different types of classifiers.

1 Introduction

Many real applications of Decision Support Systems are based on classification systems, and a great number of fields are concerned. (See [1] for examples of applications with decision trees). Numerous types of classifiers are developed: Rule-based systems (see [2] for references); Statistical and machine learning classifiers, with a great number of mathematical methods and classification algorithms [3], [4]; And deterministic classifiers, which are developed to solve a particular problem.

Explanation in rule-based systems has a long history. It was first based on the study of the trace of reasoning, to answer to how and why questions (see for example [5], and [6]). Works on the trace of reasoning eventually directed towards reconstructive explanation [7]. In fact it can be shown that in general, the trace of reasoning cannot be a good support for explanation [8].

Statistical and machine learning classifiers, when used as decision systems, provide as an outcome the class label of a new input case, possibly with some estimate of the class membership probability. Generally some information about the performance of the classifier as a model is also available (error rates or a confusion matrix). Works have also been done concerning data visualization [9].

The end-user of a deterministic model relies in the best case on some sensitivity analysis at the model level, as does more rarely the end-user of a statistical or machine learning classifier. The influence of parameter changes is studied at the model level (with experimental design and response surface analysis for example). Performance rate of the model itself carries no information about a particular case. But even conditional probability estimates, which carry information about the probability of the case to belong to the predicted class, don't say much about the link between a particular case and the predicted class [10]. Two cases which have the same class label and the same class membership probability can be very different in other respects. The more striking example is the respective position of the cases to the decision boundary (the boundary of the inverse image of the different classes in the input space). One case can be very close to the decision boundary, which means that a small change of its attribute values can change the decision. The other one can be far from the decision boundary, which means that a sizable perturbation is necessary to change the decision. This type of information concerns the context of the decision and not its probability.

In the case of numerical state space, when it is possible to define a metric, we propose a geometric method in order to produce a contextual description of the result given to the end-user, using part of the information encompassed in the classification system but which is generally not used. We study the relative position of the decision boundary to assess the robustness of the decision: If a case is far from the decision boundary, then a considerable change in its attribute value will be necessary to change the decision, and vice versa.

The paper is organized as follow: Section 2 presents some examples of the drawbacks of the trace of reasoning as explanation support, and also examples of situations where probabilistic estimate fails to represent the context of the decision to the end-user. Section 3 presents the geometric study, and discusses advantages and limits of the geometric method. Section 4 presents a complete example of a deterministic classifier for the lake eutophication problem.

2 Limits of logical and probabilistic viewpoints

2.1 Trace of reasoning

The logical viewpoint consists generally in a justification of the result through the trace of reasoning. The trace is a by-product of several types of classification systems: Knowledge-based system, rule-based system, decision tree, argumentation, etc. It is an easy way to propose explanation. But the limits of the trace as explanation support were underlined long ago (see [11] for a criticism of first generation of expert systems). Many works were done in order to bypass these problems, but, among other criticisms, there are technical arguments against the use of the trace of reasoning to generate explanation of a result [8]. The main problem is that rules in rule-based systems, or tests in decision trees, are in fact shortcuts from the case to the decision. Shortcuts are very useful to compute the decision: it's quick (the decision is known as soon as sufficient conditions are fulfilled) and efficient (the same test is used to classify large areas of the

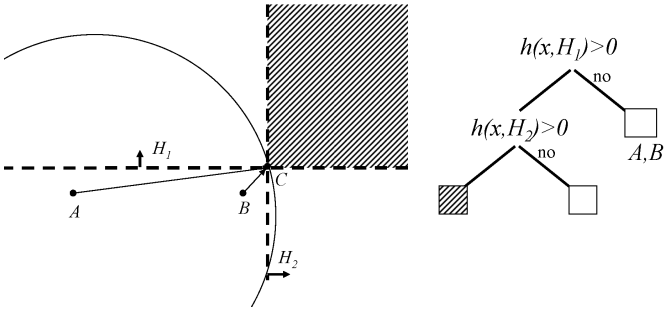

Fig. 1. Point A and B are classified by a single test $h(x, H_1) \leq 0$, where h is the algebraic distance to hyperplane H_1. Nevertheless, in a neighborhood of A, this is not a necessary condition for a point to belong to the class of A. In a neighborhood of B, the sign of the test $h(x, H_2)$ is also necessary to describe correctly the classification. The projection $C = p(A) = p(B)$ lies on both hyperplane H_1 and H_2. Both tests are necessary to describe the situation of point A (and B): In an open ball centered on A (or B), ($h(y, H_1) > 0$ and $h(y, H_2) > 0$) is the necessary and sufficient condition for a point y to belong to an other class area than A (or B respectively).

input space). But in return there is no useful contextual information about a case in the trace of reasoning. In particular, since the tests in the trace are just sufficient conditions, they could be modified without any change in the decision. Conversely, the trace can miss tests that are necessary to describe a change of decision in a neighborhood of a case, as it is the case for point B in figure 1. The logical viewpoint, when it is based on sufficient conditions only, is therefore little useful to describe the link between a case and its class.

2.2 Probability estimate

The probabilistic viewpoint in classification problem considers that the result (the decision) associated to a case is best described with the class membership conditional probability estimate. Obviously this information is very useful to assess the validity of the result. Choosing a decision with probability 0.95 is rather different from choosing the same decision with probability 0.55, or worse with highest probability 0.4 in a three-class problem. However, the probabilistic viewpoint lets aside important contextual information about the case. Figure 2 shows two cases whose class membership probability is the same and nevertheless the contextual situation is rather different. In particular, the robustness of each case to possible perturbation is very different: The distance of A from the decision boundary is much smaller than the distance of B to the same boundary (the distance from B to p(B) is very large compared to the distance from A to p(A)). From what we know of the human perception of probabilities [12], obviously the end-user would not consider the decision for case 'A' or for case 'B' in the same way, despite the probabilistic outcome. Figure 2 shows how contextual information can be interesting even in the probabilistic framework.

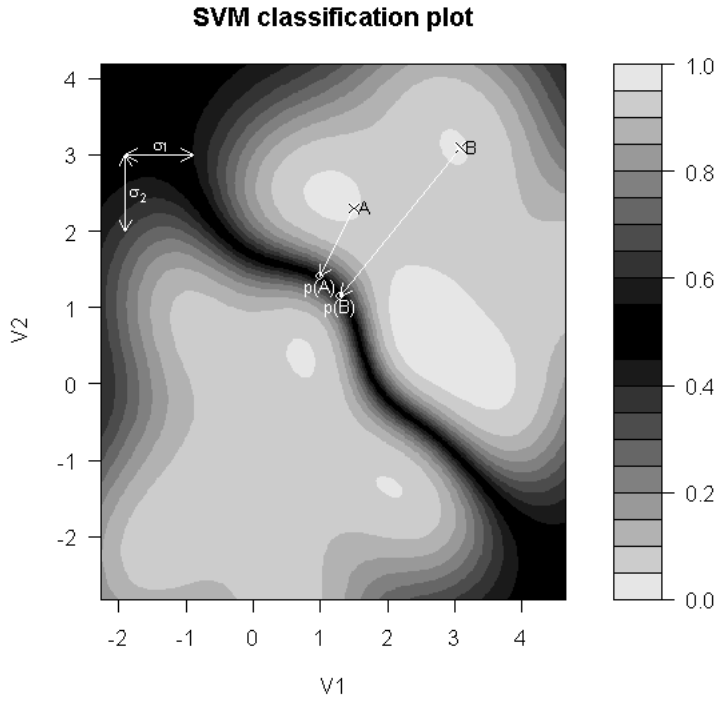

Fig. 2. Class membership probability from a SVM classifier (R package kernlab [13] with default parameters) built from examples drawn from two Gaussian distributions centered on (0,0) and (2,2) respectively. Points A and B have the same class membership probability (0.96), although A is much closer to the decision boundary than B. (p(A) and p(B) are their respective projection onto the decision boundary).

3 A geometric approach to contextual information

The explanation process can be seen as a particular problem solving task, with its own goals and tasks, as in [7]. In this paper we concentrate on one aspect only, the robustness of the decision when the case may change. We propose to evaluate this property of the case and to give this information to the end-user as a mean to assess the outcome of the decision support system. After the building stage, which depends on the type of classifier (supervised learning, modelling, etc.), the end-user submits new cases to the classifier which predicts their class label. The classifier works like a decision function, it associates a class label $c(x)$ to each vector x of the input space. We consider here classification systems operating on numerical data, such that it possible to define a metric on the input space E (see [8] for a discussion about the choice of the metric). Since a decision function performs a partition of the space, we define (classically) the robustness s of the decision at point x as the largest distance d such that

the class of each point in the open ball $B(x, d)$ is the same as x class: $s(x) = \max\{d \geq 0;\ \forall y \in B(x, d),\ c(y) = c(x)\}$. We define a sensitive move $m(x)$ at x as the smallest vectors v such that the robustness of the decision at $x+v$ is zero: $m(x) = argmin_{\delta \in E}\{\|\delta\|\ ,\ s(x + \delta) = 0\}$

In a colloquial, non-technical sense, the decision at point x is very sensitive when it cannot resist perturbations. Formally here it means that its robustness is near zero. Basically, if the robustness at x is $s(x)$, it means that the decision given by the classifier will be the same even if a perturbation is applied to x, as long as the size of the perturbation is smaller than $s(x)$. For example, in figure 2, the robustness of case A is $\|p(A) - A\|$. It is approximatively the value of the standard error. On the other hand, the robustness of case B, $\|p(B) - B\|$ is very large compared to the value of the standard error.

Robustness against perturbation gives complementary information to any statistical information given by the classifier (either the class with the highest probability estimate, or a vector of class membership probability estimates). Actually, a decision can be almost sure (a class with a probability estimate equals to or near one), and nevertheless very sensitive. This is the case for instance for determining the state of a dynamical system near bifurcation point, or the final attractor near the boundary of a basin of attraction; This is also the case for decision depending on a threshold, like means-tested benefit, choice of a device depending on weight or age, etc. In all these situations, it is important to inform the end-user about the sensitivity of the decision given by the classifier. With this additional information, the end-user can adapt his decision strategy (which can depend on decision cost, risk perception, etc., see [12].) On the same idea, sensitive moves carry relevant information to the end-user. Each sensitive move represents a set of smallest moves that are necessary to change the decision. The end-user can use this information to attempt to change the decision if it is not satisfying (for instance, in diagnosis or corrective maintenance application); He can also use it to appreciate the risk of a change in the decision, when there is some uncertainty on the value of the attributes. Since every classifier induces a partition of the input space that defines a decision boundary Γ as the union of the boundaries of the different areas corresponding to the different classes, the distance to the decision boundary is the robustness by construction. This distance is reached for some points of the boundary, which defines the sensitive moves. The geometric study of sensitivity analysis is then based on the computation of this projection onto the decision boundary.

When the decision boundary is described by an analytical formula, or by a set of constraints, it is possible to compute its exact value for all points of the input space, with efficient algorithms (see [8], [14]). It is the case for decision trees, and the geometric study gives the list of the tests in the tree that are relevant to describe the situation of a case, contrary to the trace of classification. For instance, for case A or B in figure 1, the trace of classification is reduced to a single test $h(A, H_1) \leq 0$ ($h(B, H_1) \leq 0$ respectively.) But the projection $p(A)$ ($p(B)$) lies on the intersection of the hyperplanes which define the decision boundary. Both tests $h(A, H_1) \leq 0$ and $h(A, H_2) \leq 0$ have to be changed in

order to change the decision. This list of tests, together with the distance to the decision boundary, are the relevant elements of explanation of the decision.

When the decision boundary is described with an implicit formula or in extension, it is necessary to approximate the distance and the projection with numerical algorithms, as in mathematical morphology. We have adapted an optimal distance algorithm from [15] in order to compute an approximation of the distance and of the projection, in the general case. As an example, we have computed the distance to the decision boundary and the projection for the classifier used in Figure 2. Decision for case B is very robust: A perturbation has to be more than twice the size of the standard error in order to make the decision change. In fact this is also the case for points in the upper right corner of Figure 2. Although their class membership probability drops towards 0.6 (because of the sparsity of the data in this area and the use of a Gaussian kernel in the SVM), they are very far from the decision boundary: The decision is very robust to perturbation.

4 Application to a deterministic classifier: A viability kernel for the eutrophication lake problem

We consider in this section an example of classifier for which the decision boundary is not available analytically. It is an illustrative application of a method developed to model usage conflicts in environmental engineering. Lake eutrophication is the sudden shift from clear-water (oligotrophic) to turbid-water (eutrophic) state. Oligotrophic state has obviously a tremendous higher economic value of ecosystem services. The issue is to determine whether the lake can remain in an oligotrophic state or if it is doomed to become eutrophic in finite time. The problem can be modelled by the dynamic between phosphorus concentration P and phosphorus inputs L (excess phosphorus is imported by farms). The main constraints are that agriculture requires a minimum value for L, and oligotrophic state requires a maximum value for P. Possible action on this socio ecosystem are regulation laws (constraints on $\frac{dL}{dt}$). In this model, using the concepts and methods of viability theory [16], the viability kernel is the subset of the (L, P)-plane that gathers all states (L, P) such that there exists at least one regulation law that allows the oligotrophic state to be maintained [17].

The information of the distance to the viability kernel boundary is valuable because of measure uncertainty or exogenous disturbances (which may cause a sudden increase in phosphorus concentration for instance). Figure 3 shows the viability kernel with the level curves of the distance to its boundary. The geometric study shows when a state is dangerously close to the decision boundary. For example, point B stands very close to the boundary compare to A: it is less robust to perturbation.

The geometric study also gives global information concerning the problem: Point M is the center of the maximal maximal ball in the viability kernel: It is the farthest point from the boundary with $d(M, \Gamma) \approx 0.36$. This distance can be compared with (half) the size of the constraint set ($0.9/2 = 0.45$ along L). So

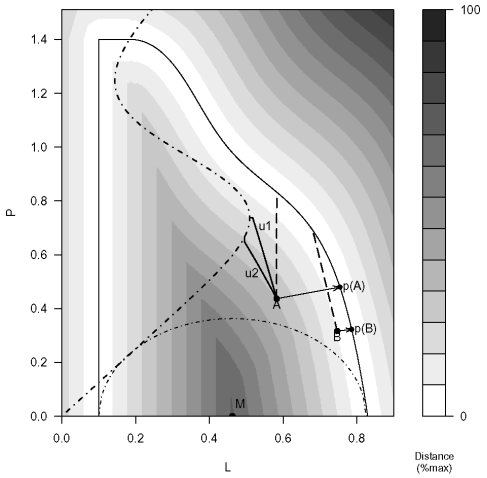

Fig. 3. Distance map for the viability kernel boundary. The dot dashed line is the set of equilibriums. $p(A)$ shows the direction and size of the most sensitive perturbation at point A (B respectively). Point M is the center of the maximal maximal ball. (Long dashed) trajectories that leave the viability kernel are not resilient.

this viability kernel has a reasonable size, it should be easy to find action policy that guarantees an oligotrophic state. In ecology, the resilience measures the capacity of a dynamical system to maintain given properties. Here the distance to the viability kernel boundary gives a geometric definition of resilience inside the viability kernel. Since the distance map gives the resilience at each point, it is possible to use this information to define resilience indicators at the level of a trajectory. Several definitions can be proposed, depending on the risk perception of the manager. For example, the most risk-adverse indicator is the min over the trajectory.

Definition 1. ***Resilience of a trajectory.***
Let $u : t \longmapsto u(t)$ be a trajectory in the viability kernel V. We note Γ_V its boundary. The resilience value of u is $r(u) = \min_t\{d(u(t), \Gamma_V)\}$.

With this definition, the trajectory corresponding to a constant nil control starting from A in Figure 3 has a resilience value of zero: it leaves the viability kernel in finite time. Both trajectories u_1 and u_2 starting from A with constant (negative) control have a strictly positive resilience value. The resilience of u_2 is greater than the resilience of u_1. This geometric indicator describes the situation. It can also be used to define particular action policy, for instance by maximizing the resilience.

5 Conclusion

In this paper, we have proposed an approach based on geometry, which objective is to provide the end-user with contextual information about the output of a clas-

sifier when it is used as a decision system. Sensitivity analysis gives information that probability estimates and error rates cannot reach. It gives a description of the situation that the trace of reasoning (if available) can generally not produce. When data are numerical, in the worst case (when the projection onto the decision boundary is unknown), an optimal algorithm coming from morphological mathematics computes the distance to the decision boundary and the projection. The projection shows the sensitive move, the smallest perturbation that makes the decision change, or, in a complementary viewpoint, it gives the maximal level of uncertainty that preserves the decision. The main limit of the method is that the input cases have to belong to a metric space.

References

1. Murthy, S.: Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery **2**(4) (1998) 345–389
2. Liao, S.H.: Knowledge management technologies and applications–literature review from 1995 to 2002. Expert Systems with Applications **25**(2) (2003) 155 – 164
3. Kotsiantis, S.B.: Supervised Machine Learning: A Review of Classification Techniques. Informatica **31** (2007) 249–268
4. Mitchell, T.: Machine Learning. McGraw Hill (1997)
5. Swartout, W.: XPLAIN: a system for creating and explaining expert consulting programs. Artificial Intelligence **21** (1983) 285–325
6. Chandrasekaran, B., Tanner, M., Josephson, J.: Explaining control strategies in problem solving. IEEE Expert **4**(1) (1989) 9–24
7. Paris, C., Wick, M., Thompson, W.: The line of reasoning versus the line of explanation. In: AAAI'88 Workshop on Explanation. (1988) 4–8
8. Alvarez, I.: Explaining the result of a decision tree to the end-user. In: Proc. of the 16th European Conference on Artificial Intelligence, IOS Press (2004) 411–415
9. Do, T., Poulet, F.: Enhancing svm with visualization. In: Proc. of Discovery Science, Springer-Verlag (2004) 183–194
10. Kukar, M., Kononenko, I.: Eliable classifications with machine learning. In: Proc. ECML–02, Springer (2002) 219–231
11. Hasling, D., Clancey, W., Rennels, G.: Strategic explanations for a diagnostic consultation system. Int J. Man-Machine Studies **20** (1984) 3–19
12. Plous, S.: The Psychology of Judgment and Decision Making. McGraw-Hill (1993)
13. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab - an s4 package for kernel methods in r. Journal of Statistical Software **11**(9) (2004) 1–20
14. Bauschke, H., Borwein, J.: On projection algorithms for solving convex feasibility problems. SIAM Review **38**(3) (1996) 367–426
15. Meijster, A., Roerdink, J., Hesselink, W.: A General Algorithm for Computing Distance Transforms in Linear Time. Morphology and Its Applications to Image and Signal Processing (2000) 331–340
16. Aubin, J.: Viability Theory. Birkhauser, Basel (1991)
17. Martin, S.: The cost of restoration as a way of defining resilience: a viability approach applied to a model of lake eutrophication. Ecology and Society **9**(2) (2004)