The Vapnik-Chervonenkis dimension of convex n-gon classifiers

Gábor Takács

HU ISSN 1418-7108: HEJ Manuscript no.: ANM-071203-A

Abstract

In statistical learning theory, the Vapnik-Chervonenkis dimension is an important property of classifier families. With the help of this combinatoral concept it is possible to bound the error probability of a classifier, based on its performance on the training set. Convex polygon classifiers are $\mathcal{R}^2 \mapsto \{+1, -1\}$ mappings that partition the plane into 2 distinct regions such that one of the regions is a convex polygon. In this paper, the Vapnik-Chervonenkis dimension of convex n-gon classifiers is determined. Note that the label of the inner (convex) region is unrestricted which makes the problem substantially different from the well known restricted case.

1 Introduction

In this article *classifiers* are $\mathbb{R}^d \mapsto \{+1, -1\}$ mappings. The input vector and the assigned output value are usually called the *observation* and the *class label*. *Convex n-hedron classifiers* are functions that can be expressed in one of the following forms:

$$g(\mathbf{x}) = \text{sgn}(\min_{1\le i\le n} \mathbf{w}_i^T\mathbf{x} + b_i),$$

$$g(\mathbf{x}) = \text{sgn}(\max_{1\le i\le n} \mathbf{w}_i^T\mathbf{x} + b_i),$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\text{sgn}(0) \stackrel{\text{def}}{=} 1$. The function class generated by only the first / second form is denoted by MIN(d, n) / MAX(d, n). The union of MIN(d, n) and MAX(d, n) is denoted by MINMAX(d, n). In the special case $d = 2$ convex n-hedron classifiers are called *convex n-gon classifiers*.

In statistical learning theory [1], the Vapnik-Chervonenkis (VC) dimension is an important property of classifier families. We say that a set of classifiers $\mathcal{G}$ *shatters* a finite set of points, if the points can be arbitrarily labeled by the members of $\mathcal{G}$. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{G}$ (denoted by $h(\mathcal{G})$) is the maximum number of points that can be shattered by $\mathcal{G}$. (If $\mathcal{G}$ can shatter arbitrarily many points, then $h(\mathcal{G}) = \infty$.) This combinatoral concept is very useful in the field of classification, because it appears in distribution-free error bounds [1]. Given a classifier g, there is no general connection between its error probability $R(g)$ and its error rate $R_m(g)$ measured on the m-element training set. However if we know a priori that $g \in \mathcal{G}$ and $h(\mathcal{G}) < \infty$, then with probability $1 - \delta$

$$R(g) \le R_m(g) + \sqrt{8\frac{h(\mathcal{G})\ln(2em/h(\mathcal{G})) + \ln(2/\delta)}{m}}.$$

It is easy to show that $\text{h}(\text{MIN}(2, n)) = \text{h}(\text{MAX}(2, n)) = 2n + 1$ [2]. This paper is about determining the VC dimension of MINMAX(2, n), which is a substantially different problem. Obviously, $\text{h}(\text{MINMAX}(2, n)) \ge 2n + 1$, because $\text{MINMAX}(2, n) \supseteq \text{MIN}(2, n)$. By Assouad's lemma [3] we know that for any two function classes $\mathcal{F}$ and $\mathcal{G}$ with finite VC dimension, $\text{h}(\mathcal{F} \cup \mathcal{G}) \le \text{h}(\mathcal{F}) + \text{h}(\mathcal{G}) + 1$. Applying this to MIN(2, n) an MAX(2, n) we get that $\text{h}(\text{MINMAX}(2, n)) \le 4n + 3$. In this paper we will prove that the truth is near the lower bound. More precisely our statement is the following:

Theorem 1.

$$\text{h}(\text{MINMAX}(2, n)) = \begin{cases} 3 & \textit{if } n = 1, \\ 2n + 2 & \textit{otherwise.} \end{cases}$$

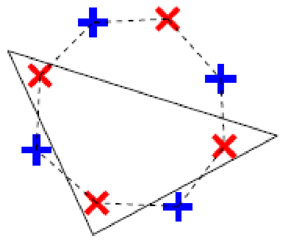

Figure 1: A point set in convex position. The red and a blue signs cannot be separated by a triangle, because for this we should intersect all edges of a convex 8-gon with 3 lines.

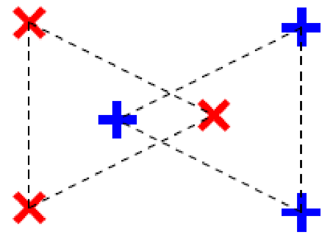

Figure 2: A point set in tangled position. Independent of the value of n, the red and the blue signs can never be separated by a convex n-gon.

2 Concepts for the proof

Definition 1. A planar point set $\mathcal{P}$ is said to be in *convex position*, if its elements are the vertices of a convex polygon.

In other words $\mathcal{P}$ does not have two distinct subsets $\mathcal{Q}_1$ and $\mathcal{Q}_2$, such that the convex hull of $\mathcal{Q}_1$ contains a point from $\mathcal{Q}_2$. The following simple facts can help in the proofs to reduce the infinite case to a finite one:

- $\mathcal{P}$ is in convex position, if and only if every 4-element subset of $\mathcal{P}$ is in convex position.
- A 2-element subset of $\mathcal{P}$ is called an *edge*, if the line segment connecting the two points is an edge of the convex hull of $\mathcal{P}$. $\mathcal{P}$ is in convex position, if and only if every 5-element subset of $\mathcal{P}$ containing an edge is in convex position.

Note that MINMAX(2, n) cannot shatter $2n+2$ convexly positioned points, because for the alternating labeling we should intersect all edges of a convex $2n+2$-gon with n lines (Fig. 1). This is also true for MIN(2, n), moreover it implies that $h(\text{MIN}(2,n)) \leq 2n+1$, since MIN(2, n) can shatter only convexly positioned point sets. The main difference between the two function classes is that MINMAX(2, n) is able to shatter a non-convexly positioned point sets too.

Definition 2. A planar point set $\mathcal{P}$ is said to be in *tangled position*, if it has two distinct subsets $\mathcal{Q}_1$ and $\mathcal{Q}_2$, such that the convex hull of $\mathcal{Q}_1$ contains a point from $\mathcal{Q}_2$ and the convex hull of $\mathcal{Q}_2$ contains a point from $\mathcal{Q}_1$.

$\mathcal{Q}_1$ and $\mathcal{Q}_2$ are called the *tangled subsets*. If $\mathcal{P}$ is not in tangled position, then it is said to be *tangle-free*. Note that for any n, a tangled set of points cannot be shattered by MINMAX(2, n), because it is impossible to separate $\mathcal{Q}_1$ from $\mathcal{Q}_2$ (Fig. 2).

3 The proof

First of all, let us recall the statement of the theorem:

$$\text{h}(\text{MINMAX}(2,n)) = \begin{cases} 3 & \text{if } n = 1, \\ 2n+2 & \text{otherwise.} \end{cases}$$

The case $n = 1$ is trivial, therefore we consider only the case $n \geq 2$. It is easy to see that $\text{h}(\text{MINMAX}(2,n)) \geq 2n+2$. Just place $2n+1$ points along a circle, in the vertices of a regular $(2n+1)$-gon and put an additional point in the center. Consider an arbitrary labeling of these $2n+2$ points. We will refer to the points that have the same label as the center as red points while to the others as blue ones. There can be at most n blue sequences along the circle. If the longest blue sequence is at most n long, then each blue sequence can be separated from the red points by 1 line. If the length of the longest blue sequence is more than n, then that blue sequence can be separated from the red points by 2 lines, and each remaining one by 1 line. If $n \geq 2$, then the number of the remaining blue sequences is not greater than $n-2$, therefore n lines are enough.

Proving the upper bound $\mathrm{h}(\mathrm{MINMAX}(2,n)) \leq 2n+2$ is a bit more difficult. We should show that no $2n+3$ points can be shattered by MINMAX(2, n). It suffices to consider point sets in general position (no 3 points are co-linear), because if there is a point set that can be shattered by MINMAX(2, n), then there also exists a generally positioned point set of the same size that can be shattered by MINMAX(2, n). (The second point set can be constructed from the first by infinitesimal perturbations.)

Assume that MINMAX(2, n) shatters a generally positioned point set $\mathcal{P}$. So far we know two necessary conditions for this:

- $\mathcal{P}$ contains no $2n+2$ points that are in convex position.
- $\mathcal{P}$ is tangle-free.

In the rest of the paper we will prove that if $n \geq 2$ and $|\mathcal{P}| \geq 2n+3$, then these requirements are contradictionary, therefore no $2n+3$ points can be shattered by MINMAX(2, n).

Theorem 2. *Let $\mathcal{P}$ be a planar point set in general position. If $\mathcal{P}$ is tangle-free and $|\mathcal{P}| \neq 6$, then $\mathcal{P}$ contains $|\mathcal{P}|-1$ points that are in convex position.*

Remark. General position is required, because we do not want to bother with degenerate polygons lying on the boundary of convex and concave. The theorem would remain valid, if we omitted this restriction.

Proof. Denote the convex hull of $\mathcal{P}$ by $\mathrm{conv}(\mathcal{P})$. If $\mathrm{conv}(\mathcal{P})$ is a point or a line segment, then the statement is trivial. The other cases are not so easy, because we can put arbitrarily many points into $\mathrm{conv}(\mathcal{P})$ such that the requirements of the theorem are fulfilled. By the property of being a vertex of $\mathrm{conv}(\mathcal{P})$ or not, the elements of $\mathcal{P}$ can be classified as *outside* or *inside* points.

At first consider the case when $\mathrm{conv}(\mathcal{P})$ is a triangle. Denote the 3 outside points by A, B and C. If $|\mathcal{P}| \leq 5$, then the statement of the theorem can be easily verified. Therefore we can assume that $|\mathcal{P}| \geq 7$, so we have at least 4 inside points.

Now select two arbitrary inside points and denote them by D and E. The line DE intersects two edges of the triangle ABC. Without the loss of generality we can assume that the line DE intersects the edge AB in the direction of D and intersects the edge AC in the direction of E. Draw the following line segments into the triangle ABC:

- Segment DE, extended to the edges AB and AC,
- Segment BD, extended to the edge AC,
- Segment CE, extended to the edge AB,
- The extension of segment AD in the direction of D,
- The extension of segment AE in the direction of E,
- Segment BE,
- Segment CD.

These line segments partition the triangle ABC into 14 distinct regions $(\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{14})$.[1]

Number them according to Fig. 3. Now try to put a third inside point F into the triangle ABC without introducing a tangle.

Lemma 1. *If $F \notin \mathcal{R}_3 \cup \mathcal{R}_5 \cup \mathcal{R}_{11} \cup \mathcal{R}_{13}$, then $\mathcal{P}$ is in tangled position.*

Proof.

- If $F \in \mathcal{R}_1 \cup \mathcal{R}_2$, then $\mathcal{Q}_1 = \{A, C, D\}$ and $\mathcal{Q}_2 = \{B, E, F\}$ are the tangled subsets.
- If $F \in \mathcal{R}_1 \cup \mathcal{R}_4$, then $\mathcal{Q}_1 = \{A, B, E\}$ and $\mathcal{Q}_2 = \{C, D, F\}$.
- If $F \in \mathcal{R}_6 \cup \mathcal{R}_7 \cup \mathcal{R}_8$, then $\mathcal{Q}_1 = \{B, D, E\}$ and $\mathcal{Q}_2 = \{A, C, F\}$.

[1] Boundary points belong to the region with the smaller index.

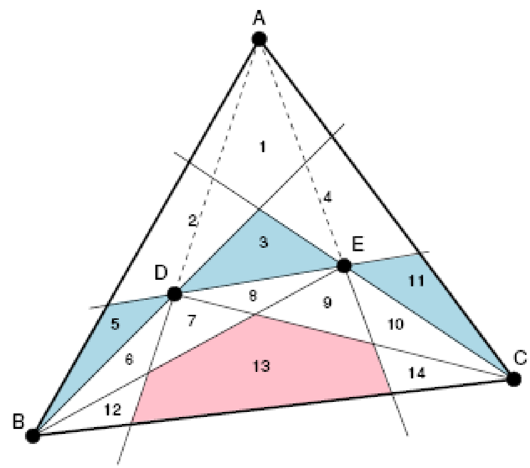

Figure 3:

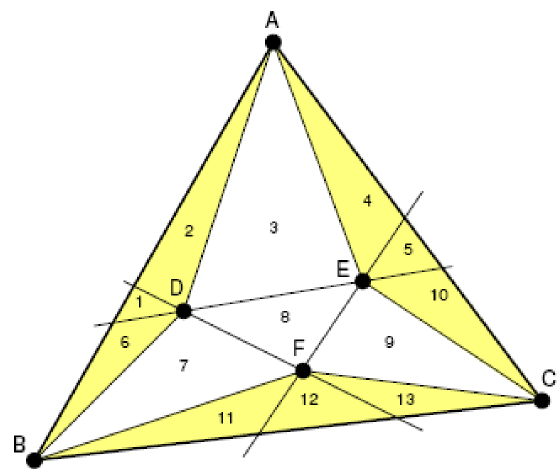

Figure 4:

- If $F \in \mathcal{R}_8 \cup \mathcal{R}_9 \cup \mathcal{R}_{10}$, then $\mathcal{Q}_1 = \{C, D, E\}$ and $\mathcal{Q}_2 = \{A, B, F\}$.
- If $F \in \mathcal{R}_6 \cup \mathcal{R}_{12}$, then $\mathcal{Q}_1 = \{B, C, D\}$ and $\mathcal{Q}_2 = \{A, E, F\}$.
- If $F \in \mathcal{R}_{10} \cup \mathcal{R}_{14}$, then $\mathcal{Q}_1 = \{B, C, E\}$ and $\mathcal{Q}_2 = \{A, D, F\}$.

□

Lemma 2. *If $F \in \mathcal{R}_{13}$, then $\mathcal{P}$ is in tangled position.*

Proof. If $F \in \mathcal{R}_{13}$, then lines DE, DF and EF partition the triangle ABC into 13 distinct regions $(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{13})$. Number them according to Fig. 4. Now try to place a 4th inside point G without introducing a tangle.

- If $G \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_4 \cup \mathcal{S}_5 \cup \mathcal{S}_6 \cup \mathcal{S}_{10} \cup \mathcal{S}_{11} \cup \mathcal{S}_{12} \cup \mathcal{S}_{13}$, then $\mathcal{P}$ is in tangled position by Lemma 1.
- If $G \in \mathcal{S}_3 \cup \mathcal{S}_8$, then $\mathcal{Q}_1 = \{A, D, E, F\}$ and $\mathcal{Q}_2 = \{B, C, G\}$ are the tangled subsets.
- If $G \in \mathcal{S}_7 \cup \mathcal{S}_8$, then $\mathcal{Q}_1 = \{B, D, E, F\}$ and $\mathcal{Q}_2 = \{A, C, G\}$.
- If $G \in \mathcal{S}_8 \cup \mathcal{S}_9$, then $\mathcal{Q}_1 = \{C, D, E, F\}$ and $\mathcal{Q}_2 = \{A, B, G\}$.

□

Remark. If F is a non-boundary point of $\mathcal{R}_{13}$, then $\{A,B,C,D,E,F\}$ is tangle-free, but has no 5-element subset in convex position. This is why the restriction $|\mathcal{P}| \neq 6$ had to be made. However, by Lemma 2 this arrangement is an irrelevant branch that cannot be continued.

The following fact is a simple consequence of Lemma 1 and Lemma 2:

Corollary 1. *If a set of two outside and three inside points is not in convex position, then then $\mathcal{P}$ is in tangled position.*

Now we are ready to finish the special case, when $\mathrm{conv}(\mathcal{P})$ is a triangle.

Lemma 3. *Let $\mathcal{P}$ be a planar point set in general position. If $\mathcal{P}$ is tangle-free, $|\mathcal{P}| \neq 6$ and conv($\mathcal{P}$) is a triangle, then we can select $|\mathcal{P}|-1$ points from $\mathcal{P}$ that are in convex position.*

Proof. Let us analyze the situation after placing m inside points. Denote the union of $\{B,C\}$ and the first m inside points with $\mathcal{T}_m$. We know that $\mathcal{T}_2$ is in convex position. We will show that if $m \geq 2$, then the convex position of $\mathcal{T}_m$ implies the convex position of $\mathcal{T}_{m+1}$. To verify this assume indirectly that $\mathcal{T}_m$ is in convex position but $\mathcal{T}_{m+1}$ is not. Since $\{B,C\}$ is always an edge of $\mathrm{conv}(\mathcal{T}_{m+1})$ and $m \geq 2$, this means that $\mathcal{T}_{m+1}$ has a 5-element subset that contains $\{B,C\}$ and is not in convex position. Then by Corollary 1, $\mathcal{P}$ is in tangled position, which is a contradiction. Thus the convex position of $\mathcal{T}_{m+1}$ follows from the convex position of $\mathcal{T}_m$. As a consequence, the set $\mathcal{T}_{|\mathcal{P}|-3} = \mathcal{P} \setminus \{A\}$ is also in convex position. □

Remark. If we prohibit to place the third inside point into $\mathcal{R}_{13}$, then the condition $|\mathcal{P}| \neq 6$ can be omitted.

At second, consider the case when $\mathrm{conv}(\mathcal{P})$ is a quadrangle. Denote the 4 outside points with A, B, C and D. If $|\mathcal{P}| \leq 5$, then the statement is trivial, therefore we can assume that we have at least two inside points. Select two arbitrary inside points and denote them by E and F. The line EF intersects two adjacent edges of the quadrangle $ABCD$, because otherwise $\mathcal{P}$ would be in tangled position. Without the loss of generality we can assume that the line EF intersects the edge AB in the direction of E and intersects the edge AC in the direction of F. Now try to place a third inside point G into the quadrangle $ABCD$.

Lemma 4. *If G is inside the pentagon $BCDEF$, then $\mathcal{P}$ is in tangled position.*

Proof. There are two possible cases:

1. The extension of AD in the direction of D and the extension of AE in the direction E intersect different edges of the quadrangle $ABCD$.
2. The extension of AD in the direction of D and the extension of AE in the direction E intersect the same edge of the quadrangle $ABCD$. We can assume without the loss of generality that the intersected edge is BD.

In the first case, the line segments DE and DF partition the pentagon $BCDEF$ into 3 distinct regions ($\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$), as it can be seen in Fig. 5. If $G \in \mathcal{R}_1 \cup \mathcal{R}_2$, then $\{B,D,E,F\}$ and $\{A,C,G\}$ are the tangled subsets. If $G \in \mathcal{R}_2 \cup \mathcal{R}_3$, then $\{C,D,E,F\}$ and $\{A,B,G\}$ are the tangled subsets.

In the second case, DE and the extension of AF in the direction of F partitions the pentagon $BCDEF$ into 4 distinct regions ($\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$), as it can be seen in Fig. 6. If $G \in \mathcal{S}_1 \cup \mathcal{S}_2$, then $\{B,D,E,F\}$ and $\{A,C,G\}$ are the tangled subsets. If $G \in \mathcal{S}_2 \cup \mathcal{S}_3$, then $\{C,D,E,F\}$ and $\{A,B,G\}$ are the tangled subsets. If $G \in \mathcal{S}_4$, then $\{B,C,D,F\}$ and $\{A,E,G\}$ are the tangled subsets. □

By Lemma 4, inside points up from the third can be placed only into the region $ABC \setminus BCEF$ without introducing a tangle. This and Lemma 3 (applied to $\mathcal{P} \setminus \{D\}$) implies that $\mathcal{P} \setminus \{A,D\}$ is in convex position. The restriction $|\mathcal{P} \setminus \{D\}| \neq 6$ can be now omitted, because the quadrangle $BCEF$ is a forbidden area. If $\mathcal{P} \setminus \{A,D\}$ is in convex position, then $\mathcal{P} \setminus \{A\}$ is too. This completes the proof of the special case when $\mathrm{conv}(\mathcal{P})$ is a quadrangle.

At third consider the case when $\mathrm{conv}(\mathcal{P})$ is a pentagon. Denote the 5 outside points by A, B, C, D and E. Using the same reasoning as before we can assume that we have at least two inside points. Pick two arbitrary inside points F and G. The line FG intersects two adjacent edges of the pentagon $ABCDE$, because otherwise $\mathcal{P}$ would be in tangled position. Without the loss of generality assume that the line FG intersects the edge AB in the direction of F, intersects the edge AC in the direction of G,

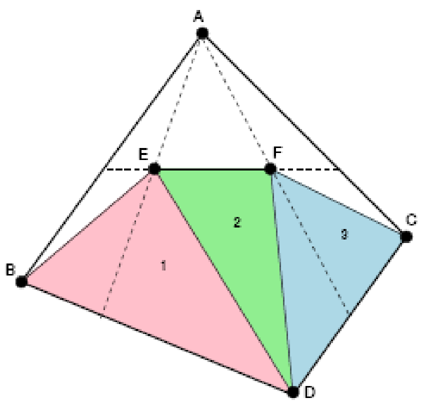

Figure 5:

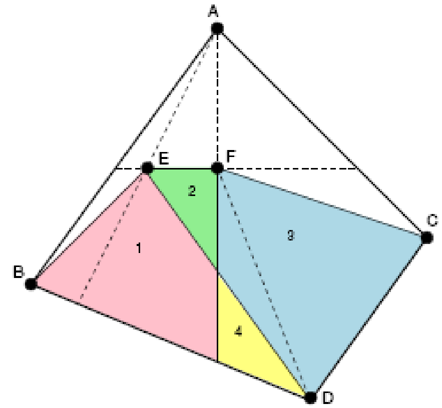

Figure 6:

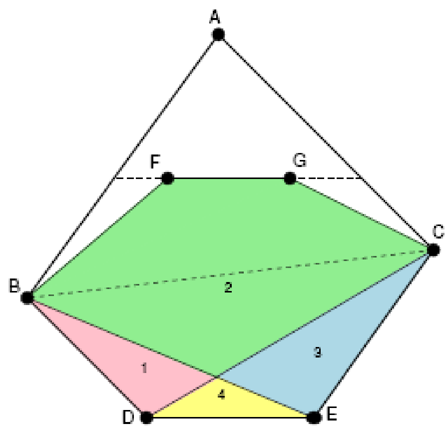

Figure 7:

moreover BD and CE are edges of the pentagon $ABCDE$. Now try to place a third inside point H into the pentagon $ABCDE$.

Lemma 5. *If H is inside the hexagon $BCDEFG$, then $\mathcal{P}$ is in tangled position.*

Proof. Line segments BE and CD partition the hexagon $BCDEFG$ into 4 distinct regions $(\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \mathcal{R}_4)$, as it can be seen in Fig. 7. If $H \in \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$, then $\mathcal{P}$ is in tangled position by Lemma 4. If $H \in \mathcal{R}_4$, then $\mathcal{P}$ is tangled too, because there exists a line connecting two inside points that intersects non-adjacent edges of $\mathrm{conv}(\mathcal{P})$. For example the line FH cannot intersect adjacent edges of $\mathrm{conv}(\mathcal{P})$. □

By Lemma 5, inside points up from the third can be placed only into the region $ABC \setminus BCFG$ without introducing a tangle. This and Lemma 3 (applied to $\mathcal{P} \setminus \{D, E\}$) implies that $\mathcal{P} \setminus \{A, D, E\}$ and this wise $\mathcal{P} \setminus \{A\}$ is in convex position.

Finally consider the case when $\mathrm{conv}(\mathcal{P})$ is a k-gon ($k \geq 6$). Denote the outside points by $A_1, A_2, \ldots, A_k$ and the inside points by $B_1, B_2, \ldots, B_m$. The line B_1B_2 intersects again two adjacent edges of $\mathrm{conv}(\mathcal{P})$. Without the loss of generality assume that the line B_1B_2 intersects the edges A_1A_k and A_1A_k. No inside point can be located in the $(k+1)$-gon $A_2A_3 \ldots A_kB_1B_2$, because otherwise $\mathcal{P}$ would be in tangled position by Lemma 5. But then it follows as before that $\{A_2, A_k\} \cup \{B_1, B_2, \ldots, B_m\} = \mathcal{P} \setminus A_1$ is in convex position. □

References

[1] O. Bousquet, S. Boucheron, G. Lugosi (2004). Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 3176:169–207.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965.

[3] P. Assouad. (1983). Densité et dimension. *Annales de l'Institut Fourier*, 33:233–282.