Machine Learning: Foundations **Fall Semester, 2012**

Lecture 8: Dec 16, 2012

Lecturer: Yishay Mansour *Scribe: Ben Klein, Tal Kaminker, Eli Daian*[1]

8.1 The PAC Model - Review

In the PAC Model we assume there exists a distribution D on the examples X that the learner receives, i.e., when choosing an instance from the sample it is drawn according to D. We assume that D is:

1. Fixed throughout the learning process.
2. Unknown to the learner.
3. The instances are chosen independently.

The target concept is specified as a computable function $c^\star$, thus our instances are of the form $< x, c^\star(x) >$. Our goal is to find a function h which approximates $c^\star$ with respect to D, in the following sense. Let

$$error(h) = \ Prob_D \left[c^\star(x) \neq h(x) \right].$$

We would like to ensure that $error(h)$ is below a certain threshold ε, which is given as a parameter to the algorithm. This parameter is a measure of the accuracy of the learning process.

As a measure of our confidence in the outcome of the learning process, we add another parameter δ. We require that the following hold:

$$Prob \left[error(h) < \varepsilon \right] \geq 1 - \delta.$$

The PAC algorithm has two inputs: the accuracy parameter ε and the confidence parameter δ. It also has access to instances using $EX(D, c^\star)$,

[1]Based on scribes written by Itamar Eskin, Yonatan Dinai and Assaf Mushinsky (November 28, 2010).

which generates a random example, using the distribution D, and labelled by $c^\star$.

We say that an algorithm A *learns* a family of concepts $\mathcal{C}$ if for **any** $c^\star \in \mathcal{C}$ and **any** distribution D on the instances in $\mathcal{X}$, A outputs a function h, such that the probability that $error(h) < \varepsilon$ is at least $1 - \delta$.

A PAC algorithm is *efficient* if its running time is polynomial in $\frac{1}{\epsilon}$, $\ln \frac{1}{\delta}$, the input size and the size of the target concept $c^\star$.

8.2 THE VC-DIMENSION

8.2.1 Motivation

Let us consider the following question: How many random examples does a learning algorithm need to draw before it has sufficient information to learn an unknown target concept chosen from the concept class C? For the case of a finite concept class C, we proved a lower bound on the number of examples required for PAC learning:

$$m \geq \frac{1}{\epsilon} \ln \frac{|\mathcal{C}|}{\delta}..$$

We would like to be able to handle infinite concept classes, perhaps even not enumerable. We already saw an example:

- Axis-aligned rectangles.

For this concept class we showed that the number of examples sufficient for PAC learning is $O\left(\frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$.

In many cases, C is defined by a significant structure, and we would like to formally quantify how this structure helps our learning algorithms. We will introduce the definition of VC–dimension and show the connection between the VC–dimension and learning. The concept of the VC–dimension, will provide us a substitute to $\ln |C|$, for infinite concept classes.

8.2.2 Definitions

We start with a few definitions. Assume C is a concept class defined over the instance space X. Let $c \in C$ identified with a set $c \subseteq X$ such that $c = \{x \in X \mid c(x) = 1\}$.

Definition For each class C over X and for any $S \subseteq X$:

$$\Pi_C(S) = \{c \cap S | c \in C\}$$

Equivalently, if $S = \{x_1, \ldots, x_m\}$ then we can think of $\Pi_C(S)$ as the set of vectors $\Pi_C(S) \subseteq \{0,1\}^m$ defined by $\Pi_C(S) = \{< c(x_1), \ldots, c(x_m) >: c \in C\}$.

This is the projection of the concept class C on the input S, namely $\Pi_C(S)$ is all the possible functions that C induces on S. We are interested in how many different functions C induces on S. In effect we are reducing the concept class C into the concept class $C|S$, where $S = \{x_1, ..., x_m\}$. The concept class $C|S$ is finite with at most 2^m different concepts, thus $|\Pi_C(S)| \leq 2^m$.

Definition A concept class C *shatters* S if $2^{|S|} = |\Pi_C(S)|$.

In other words a class shatters a set of inputs if every possible function on S can be represented by some $c \in C$.

Now we are ready to define the notion of VC–dimension.

Definition *VCdim (Vapnik-Chervonenkis dimension)* of C is the maximum size of a set shattered by C:

$$VCdim(C) = max\{d : \exists S : |S| = d \ \ and \ \ \Pi_C(S) = \{0,1\}^d\}.$$

If a maximum value does not exist then $VCdim(C) = \infty$.

8.2.3 Some examples of geometric concepts

Let us consider a few examples of simple concept classes and calculate their VC dimension. In order to show that the VC dimension of a class is at least d, we must simply find some shattered set of size d. In order to show that the VC dimension is at most d, we must show that no set of size $d + 1$ is shattered.

C_1: Half-lines

The concepts are c_α for $\alpha \in [0,1]$, $X = [0,1]$, where:

$$c_\alpha(x) = \begin{cases} 0 & x < \alpha \\ 1 & x \geq \alpha \end{cases}$$

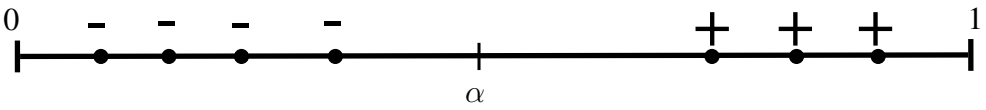

Figure 8.1: C_1

Note that although the number of concepts is uncountable the concept class is learnable. We shall prove that $VCdim(C_1) = 1$.

First, we will show that $VCdim(C_1) \geq 1$. Let $x = \frac{1}{2}$. Then we can show two concepts such that $|\Pi_C(\{\frac{1}{2}\})| = 2$:

$$c_{\frac{3}{4}}(\frac{1}{2}) = 0$$

$$c_{\frac{1}{4}}(\frac{1}{2}) = 1$$

thus the VC–dimension is at least 1.

We will now show that $VCdim(C_1) < 2$, by showing that for any set of size two, there exists an assignment which is not in the concept class. If $S = \{x, y\}$ where $y > x$, the assignment that lets x be '+' and y be '−', is impossible. Thus, $VCdim(C_1) < 2$, and we derive that $VCdim(C_1) = 1$.

C_2: Linear halfspaces in the plane

Consider a real line in the plane. For $w = (\alpha_1, \alpha_2, \theta), x \in \Re^2$, let

$$c_w(x) = 1 \iff \alpha_1 x_1 + \alpha_2 x_2 \geq \theta$$

All the positive points are above or on the line, and all the negative points are below the line. We shall prove that $VCdim(C_2) = 3$.

For this concept class, any three points that are not collinear can be shattered. Figure 8.2(a) shows how one assignment out of the possible 8

assignments can be satisfied by a halfspace. To see that no set of four points can be shattered, we consider two cases. In the first case (shown in Figure 8.2(b)), all four points lie on the convex hull defined by the four points. In this case, if we label one "diagonal" pair positive and the other "diagonal" pair negative as shown in Figure 8.2(b), no halfspace satisfies this assignment. In the second case (shown in Figure 8.2(c)), three of the four points define the convex hull of the four points, and if we label the interior point negative and the hull points positive, again no halfspace can satisfy the labeling. Thus the VC–dimension here is three. In general, for halfspaces in $\Re^d$. the VC–dimension is $d + 1$, which we will show latter in this lecture.

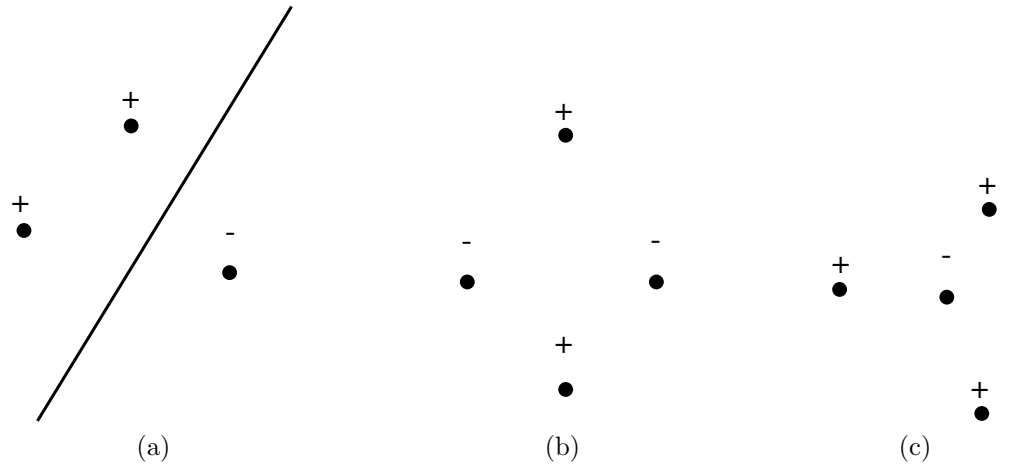

Figure 8.2: (a) example of 3 non-collinear points and a halfspace that satisfies them; (b) and (c) are examples of impossible assignments.

C_3: Axis-aligned rectangles in the plane

Consider the concept class where positive examples are points inside an axis-aligned rectangle, and negative examples are points outside the rectangle.

For this class, we can shatter the four points shown in Figure 8.3(a). However, not all sets of four points can be shattered, as indicated by the illegal assignment shown in Figure 8.3(b). Still, the existence of a single shattered set of size four is suffiecient to lower bound the VC dimension. Now for any set of five points in the plane, there must be some point that is neither the extreme left, right, top or bottom point of the five (see Figure 8.3(c)). If we label this non-extermal point negative and the remaining

four extermal points positive, no rectangle can satisfy the assignment. Thus the VC dimension is four.

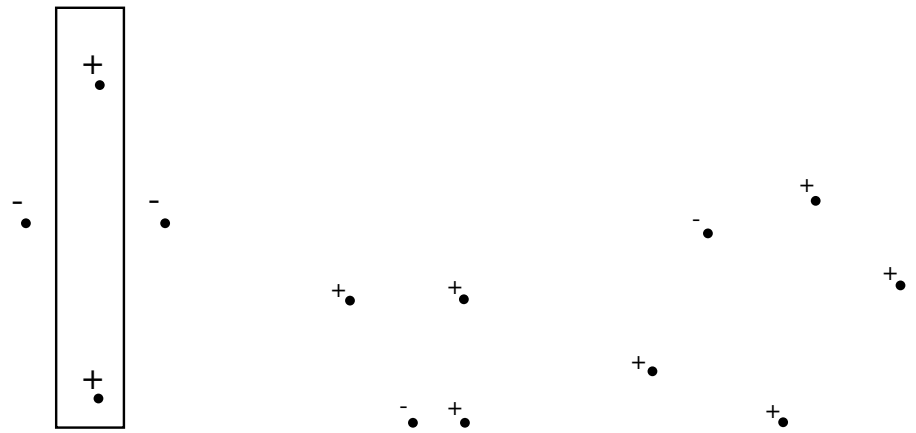

Figure 8.3: (a) 4 points that can be shattered (example for one of the possible assignments); (b) and (c) are examples of impossible assignments.

C_4: A finite union of intervals

For any set of points we could cover the positive points by choosing the intervals small enough.

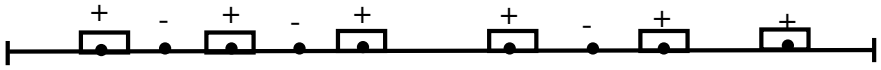

Figure 8.4: Finite union of intervals

Thus, $VCdim(C_4) = \infty$.

C_5: Convex polygons in the plane

Points inside the convex polygon are positive and outside are negative. Again, we have no bound on the number of edges, and we want to show $VCdim(C_5) = \infty$; i.e. for every d there is a set whose size is d that can be shattered by convex polygons.

Let S be a set of d points on the circle perimeter. Figure 8.5 shows that for every labelling of the points in S, there exists $c^\star \in C$ that is consistent

with the labelling. The concept $c^\star$ connects the positive points. The polygon includes all the positive examples and none of the negative ones. Thus, for any d points on the perimeter of the unit circle, all the 2^d classifications are possible. Therefore, $VCdim(C) = \infty$.

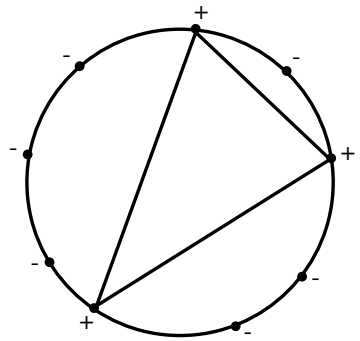

Figure 8.5: A convex polygon satisfying a labeling of points on a circle

8.2.4 Sample Size Lower Bounds

We would like to show that if a concept class has a finite VC-dimension d, then there is a function m dependent on ϵ,δ and d, such that if we sample less than $m(\epsilon, \delta, d)$ points, any PAC learning algorithm would fail.

Theorem 8.1 *If a concept class C has a VC-dimension $d+1$, then:*

$$m(\epsilon, \delta, d+1) \geq \frac{d}{16\epsilon} = \Omega\left(\frac{d}{\epsilon}\right)$$

Proof: For contradiction assume this is possible. Let $T = \{z_0, z_1, ..., z_d\}$ s.t. C shatters T (such a group must exist, since $VCdim(C) = d+1$). Now, construct a distribution D in the following manner:

$$D(x) = \begin{cases} 1 - 8\epsilon & x = z_0 \\ \frac{8\epsilon}{d} & x = z_i, 1 \geq i \leq d \\ 0 & otherwise \end{cases}$$

Choose $c^\star$ randomly:

$$c^\star(x) = \begin{cases} 1 & x = z_0; \\ 0 \text{ or } 1 \text{ (with probability } \frac{1}{2}) & x = z_i; \\ \star & otherwise \end{cases}$$

Note that there exists a $c^\star \in C$ that agrees on T, since C shatters T. We claim that if we sample less than $\frac{d}{2}$ points out of $\{z_1, ..., z_d\}$ then the error is at least 2ε. Let $RARE$ be the set of points $\{z_1, ..., z_d\}$, and $UNSEEN \subseteq RARE$ the points in $RARE$ which have not been samples. Then:

$$\Pr[ERROR] \leq \frac{1}{2} \cdot |UNSEEN| \cdot \frac{8\epsilon}{d} = \frac{4\varepsilon}{d} \cdot |UNSEEN|$$

The expected number of points in $RARE$ is $m \cdot 8\varepsilon < \frac{d}{16\varepsilon} \cdot 8\varepsilon = \frac{d}{2}$
With probability of at least $\frac{1}{2}$ we will sample at most $\frac{d}{2}$ points (recall that the expected value for the binomial distribution equals its median). In such a case, $|UNSEEN| \geq \frac{d}{2}$. This implies that with probability at least $\frac{1}{2}$ we have error at least 2ε, a contradiction. □

8.2.5 Some more examples: Boolean Functions

C_6: Parity

$X = \{0,1\}^n$. The concept class is

$$\chi_S(x) = \bigoplus_{i \in S} x_i$$

where $S \subset \{1, ..., n\}$. The lower bound: $VCdim(C_6) \geq n$.

Let $e_i =< 0...010...0 >$ be unit vectors, where '1' appears in the i-th place. There are n such vectors. For any bits assignment $b_1, ..., b_n$ for the vectors $e_1, ..., e_n$ we choose the set

$$S = \{i : b_i = 1\}$$

We get

$$\chi_S(e_j) = \begin{cases} 1 & \text{j} \in S \\ 0 & \text{j} \notin S \end{cases}$$

Thus, we conclude $VCdim(C_6) \geq n$.
The upper bound: $VCdim(C_6) \leq n$.
We present two simple proofs for the upper bound:

1. There are 2^n parity functions. Thus $VCdim(C_6) \leq \log_2 |C_6| = \log_2 2^n = n$.

2. Given $n+1$ vectors, there is a vector that is the linear combination of the others:
$$e_j = e_1 \oplus ... \oplus e_k$$
Therefore, the values of $e_1, ..., e_k$ fix the value of e_j. So the assignment
$$b_1 = 0, ..., b_k = 0,\ b_j = 1$$
is impossible.

C_7: OR of n literals

$X = \{0,1\}^n,\ S \subset \{1, ..., n\},;\ \bar{S} \subset \{1, ..., n\}$. The concept class is:

$$C_S(x) = (\vee_{i \in S} x_i) \vee (\vee_{j \in \bar{S}} \bar{x_j})$$

The lower bound: $VCdim(C_7) \geq n$.

Let $e_i =< 0...010...0 >$ where '1' appears in the i-th place. There are n such vectors. For any bits assignment $b_1, ..., b_n$ for the vectors $e_1, ..., e_n$ we choose the sets
$$S = \{i : b_i = 1\}$$
so the target concept is
$$C_S(x) = \vee_{i \in S} x_i$$
Thus, we conclude $VCdim(C_7) \geq n$. Next, we establish the upper bound.

Claim 8.2 $VCdim(C_7) \leq n$

Proof: We shall first see a connection with the Online Learning Model.

Lemma 8.3 *If a class C has an online algorithm A that does at most n mistakes then $VCdim(C) \leq n$*

Proof: By contradiction, we'll assume that $VCdim(C) \geq n+1$. We'll choose $n+1$ points that can get all the possible assignments. When running the online algorithm A: For every x_i, the response to the the prediction of A on x_i will be mistake. Therefore, the algorithm A will make $n+1$ mistakes. Contradiction to the fact that A does at most n mistakes. □

Suppose we have $n+1$ vectors. In one of the previous lectures we saw the ELIM algorithm that maintains a literals list L, that is initialized to the

set of all literals, and each assignment of 0 to vector removes all positive literals of the vector from L. We proved that in the first mistake, we eliminate exactly n from L, and that in any other mistake, we eliminate from L at least one literal. Since we started with a list L that contains $2n$ literals, the number of mistakes is at most $n+1$. Therefore, according to the lemma that we just proved, the VC-dim is at most $n+1$.

We will show that there exists an order of vectors, such that the second vector can remove at least **two** literals. This will reduce the number of errors to n. Note, that unlike the mistake bound model, we can select here any order of the points we like, to derive the upper bound.

Lemma 8.4 *Given a set of 3 (or more) vectors, there exists two of them that differ by at least two bits.*

Proof: Consider $z_1, z_2, z_3 \in \{0,1\}^n$. If z_1 and z_2 are different in more than one bit, we are done. Otherwise $z_1 \oplus z_2 = e_j$ for some $j \in [1,n]$. Similarly, if z_2 and z_3 differ in more than one bit we are done. Otherwise $z_2 \oplus z_3 = e_i$. It can not be the case that $i = j$, since then $z_1 = z_3$. Therefore $z_1 \oplus z_3 = z_1 \oplus z_2 \oplus z_2 \oplus z_3 = e_j \oplus e_i$, namely, z_1 and z_3 differ in two bits. □

Let a, b two such vectors. We can perform ther two first ELIM stages on this two vectors. So, b will remove at least two literals from L (in addition to n literals, that was removed by a), because b has two bits, in which it differs from a.

All the next vectors remove from L at least one literal. Therefore, after at most n vectors, the list L is empty. Thus,

$$VCdim(C_7) \leq n$$

□

C_8 - Hyper Plane

Let ℓ_w be a hyper plane which devides R^n into 2 sets of points:
ℓ_w^+ - points above or on the hyper plane ℓ_w.
ℓ_w^- - points below the hyper plane ℓ_w.

Formally; for $w = (\alpha_1, ..., \alpha_n, \theta) \in \Re^{n+1}$

$$l_w = \{x \in \Re^n | \sum_{i=1}^{n} \alpha_i x_i = \theta\}$$

We define the classification by C_w as,

$$C_w(x) = 1 \iff \sum_{i=1}^{n} \alpha_i x_i \geq \theta$$

We will prove the following bound.

Theorem 8.5

$$VCdim(C_8) = n + 1$$

First we will show that there are at least $n+1$ points that can be shattered by C_7.

Claim 8.6

$$VCdim(C_8) \geq n + 1$$

Proof: Consider $E = \{\overrightarrow{0}, \overrightarrow{e_1}, ..., \overrightarrow{e_n}\}$ of size $n + 1$ in order to show that C_8 shatters it.
Any classification of the vectors of E can be described by the following subsets:

$S \subset E$ - positively labeled vectors in E
$E \setminus S$ - negatively labeled vectors in E

For each S we define a hyper-plane W_s,

$$W_s = (\alpha_1^s, ..., \alpha_n^s, \theta^s),$$

where

$$\theta^s = \begin{cases} -\frac{1}{2} & \overrightarrow{0} \in S \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

and

$$\alpha_i^s = \begin{cases} 1 & \overrightarrow{e_i} \in S \\ -1 & \text{otherwise} \end{cases}$$

W_s is the hyper-plane which classifies every vector in S as '+' and vectors in $E \setminus S$ as '−'. Since

$$C_{W_s}(\overrightarrow{e_i}) = 1 \iff \alpha_i^s \geq \theta^s \iff \overrightarrow{e_i} \in S$$
$$C_{W_s}(\overrightarrow{0}) = 1 \iff 0 > \theta^s \iff \overrightarrow{0} \in S$$

□

We showed that there exists a set of size $n+1$ which C_8 shatters, hence $VCdim(C_8) \geq n+1$.
We will now show that $VCdim(C_8) = n+1$.
Before further examination can be done, some general definitions and Radon theorem will be shown.

Definition A subset A is *convex* if $\forall x_1, x_2 \in A$ the line connecting x_1 to x_2 is in A. Formally:

$$\forall \lambda \in [0,1]. \lambda x_1 + (1-\lambda) x_2 \in A$$

Definition The *Convex Hull* of S is the smallest convex set which contains all the points of S. We denote it as $conv(S)$.

We are now ready to state Radon Theorem.

Theorem 8.7 *(RADON Theorem) Let E be a set of $d+2$ points in $\Re^d$. There is a non empty subset S of E such that*

$$conv(S) \cap conv(E \setminus S) \neq \phi$$

Proof: Let:

$$E = \{x_0, ..., x_{d+1}\}$$

where $x_i \in R^d$.
Since E contains $d+2$ vectors, we can solve for the following $d+1$ equations and find $(\alpha_0, ..., \alpha_{d+1}) \neq \overrightarrow{0}$, such that,

$$\sum_{i=0}^{d+1} \alpha_i x_i = 0,$$

and

$$\sum_{i=0}^{d+1} \alpha_i = 0.$$

Thus, we have a set of $d+1$ linear equations over $d+2$ variables $\{\alpha_i\}_{i=0}^{d+1}$. There exist a non-zero vector $< \alpha_0, ..., \alpha_{d+1} >$ satisfying the above equations, because every $d+1$ points (vectors) are linear dependent.

Assume that $\alpha_0, ..., \alpha_p$ are positive, and $\alpha_{p+1}, ..., \alpha_{d+1}$ are negative (zeros can go in to either group).
We define:

- $\alpha = \sum_{i=0}^{p} \alpha_i > 0$
- $\beta_i = \frac{\alpha_i}{\alpha} > 0 \qquad 0 \leq i \leq p$
- $\gamma_i = \frac{-\alpha_i}{\alpha} > 0 \qquad p+1 \leq i \leq d+1$

We have that,

$$\sum_{i=0}^{d+1} \alpha_i x_i = 0 \;\Rightarrow\; \sum_{i=0}^{p} \beta_i x_i = \sum_{i=p+1}^{d+1} \gamma_i x_i$$

Notice that $\sum_{i=0}^{p} \beta_i = \sum_{i=p+1}^{d+1} \gamma_i = 1$.
By definition of convexity,

$$\sum_{i=0}^{p} \beta_i x_i \in conv(x_0, ..., x_p),$$

and

$$\sum_{i=p+1}^{d+1} \gamma_i x_i \in conv(x_{p+1}, ..., x_{d+1}).$$

Hence, there is a point that belongs to the intersection of

$$conv(x_0, ..., x_p) \cap conv(x_0, ..., x_p) \neq \emptyset$$

□

Claim 8.8

$$VCdim(C_8) < n + 2$$

Proof: Proof by contradiction. Assume $E = \{x_1, ..., x_{n+2}\}$ could be shattered. By RADON theorem there is a non-empty subset S of E such that $conv(S) \cap conv(E \setminus S) \neq \phi$.
Assume we have hyper-plane C_w which classifies S as '+' and $E \setminus S$ as '−', and since hyper-planes are convex,

$$S \subset \ell_w^+ \Rightarrow conv(S) \subset \ell_w^+,$$

and

$$E \setminus S \subset \ell_w^- \Rightarrow conv(E \setminus S) \subset \ell_w^-.$$

Combining the two, we have that,

$$conv(S) \cap conv(E \setminus S) \subset \ell_w^+ \cap \ell_w^- = \emptyset$$

which is a contradiction to the choice of S. Therefore we contradicted the assumption that such a hyper plane C_w exists. □

Combining Claim 7.6 and Claim 7.8, we can now conclude that $VCdim(C_8) = n + 1$.