29 Multiclass Learnability

In Chapter 17 we have introduced the problem of multiclass categorization, in which the goal is to learn a predictor $h : \mathcal{X} \to [k]$. In this chapter we address PAC learnability of multiclass predictors with respect to the 0-1 loss. As in Chapter 6, the main goal of this chapter is to:

- Characterize which classes of multiclass hypotheses are learnable in the (multiclass) PAC model.
- Quantify the sample complexity of such hypothesis classes.

In view of the fundamental theorem of learning theory (Theorem 6.8), it is natural to seek a generalization of the VC dimension to multiclass hypothesis classes. In Section 29.1 we show such a generalization, called the *Natarajan dimension*, and state a generalization of the fundamental theorem based on the Natarajan dimension. Then, we demonstrate how to calculate the Natarajan dimension of several important hypothesis classes.

Recall that the main message of the fundamental theorem of learning theory is that a hypothesis class of binary classifiers is learnable (with respect to the 0-1 loss) if and only if it has the uniform convergence property, and then it is learnable by any ERM learner. In Chapter 13, Exercise 2, we have shown that this equivalence breaks down for a certain convex learning problem. The last section of this chapter is devoted to showing that the equivalence between learnability and uniform convergence breaks down even in multiclass problems with the 0-1 loss, which are very similar to binary classification. Indeed, we construct a hypothesis class which is learnable by a specific ERM learner, but for which other ERM learners might fail and the uniform convergence property does not hold.

29.1 The Natarajan Dimension

In this section we define the Natarajan dimension, which is a generalization of the VC dimension to classes of multiclass predictors. Throughout this section, let $\mathcal{H}$ be a hypothesis class of multiclass predictors; namely, each $h \in \mathcal{H}$ is a function from $\mathcal{X}$ to $[k]$.

Understanding Machine Learning, © 2014 by Shai Shalev-Shwartz and Shai Ben-David
Published 2014 by Cambridge University Press.
Personal use only. Not for distribution. Do not post.
Please link to http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

To define the Natarajan dimension, we first generalize the definition of shattering.

DEFINITION 29.1 (Shattering (Multiclass Version)) We say that a set $C \subset \mathcal{X}$ is shattered by $\mathcal{H}$ if there exist two functions $f_0, f_1 : C \to [k]$ such that

- For every $x \in C$, $f_0(x) \neq f_1(x)$.
- For every $B \subset C$, there exists a function $h \in \mathcal{H}$ such that

$$\forall x \in B, h(x) = f_0(x) \text{ and } \forall x \in C \setminus B, h(x) = f_1(x).$$

DEFINITION 29.2 (Natarajan Dimension) The Natarajan dimension of $\mathcal{H}$, denoted $\text{Ndim}(\mathcal{H})$, is the maximal size of a shattered set $C \subset \mathcal{X}$.

It is not hard to see that in the case that there are exactly two classes, $\text{Ndim}(\mathcal{H}) = \text{VCdim}(\mathcal{H})$. Therefore, the Natarajan dimension generalizes the VC dimension. We next show that the Natarajan dimension allows us to generalize the fundamental theorem of statistical learning from binary classification to multiclass classification.

29.2 The Multiclass Fundamental Theorem

THEOREM 29.3 (The Multiclass Fundamental Theorem) *There exist absolute constants $C_1, C_2 > 0$ such that the following holds. For every hypothesis class $\mathcal{H}$ of functions from $\mathcal{X}$ to $[k]$, such that the Natarajan dimension of $\mathcal{H}$ is d, we have*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m^{\text{UC}}_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(k) + \log(1/\delta)}{\epsilon^2}.$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(k) + \log(1/\delta)}{\epsilon^2}.$$

3. *$\mathcal{H}$ is PAC learnable (assuming realizability) with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log\left(\frac{kd}{\epsilon}\right) + \log(1/\delta)}{\epsilon}.$$

29.2.1 On the Proof of Theorem 29.3

The lower bounds in Theorem 29.3 can be deduced by a reduction from the binary fundamental theorem (see Exercise 5).

The upper bounds in Theorem 29.3 can be proved along the same lines of the proof of the fundamental theorem for binary classification, given in Chapter 28 (see Exercise 4). The sole ingredient of that proof that should be modified in a nonstraightforward manner is Sauer's lemma. It applies only to binary classes and therefore must be replaced. An appropriate substitute is Natarajan's lemma:

LEMMA 29.4 (Natarajan) $|\mathcal{H}| \leq |\mathcal{X}|^{\mathrm{Ndim}(\mathcal{H})} \cdot k^{2\mathrm{Ndim}(\mathcal{H})}$.

The proof of Natarajan's lemma shares the same spirit of the proof of Sauer's lemma and is left as an exercise (see Exercise 3).

29.3 Calculating the Natarajan Dimension

In this section we show how to calculate (or estimate) the Natarajan dimension of several popular classes, some of which were studied in Chapter 17. As these calculations indicate, the Natarajan dimension is often proportional to the number of parameters required to define a hypothesis.

29.3.1 One-versus-All Based Classes

In Chapter 17 we have seen two reductions of multiclass categorization to binary classification: One-versus-All and All-Pairs. In this section we calculate the Natarajan dimension of the One-versus-All method.

Recall that in One-versus-All we train, for each label, a binary classifier that distinguishes between that label and the rest of the labels. This naturally suggests considering multiclass hypothesis classes of the following form. Let $\mathcal{H}_{\mathrm{bin}} \subset \{0,1\}^{\mathcal{X}}$ be a binary hypothesis class. For every $\bar{h} = (h_1, \ldots, h_k) \in (\mathcal{H}_{\mathrm{bin}})^k$ define $T(\bar{h}) : \mathcal{X} \to [k]$ by

$$T(\bar{h})(x) = \operatorname*{argmax}_{i \in [k]} h_i(x).$$

If there are two labels that maximize $h_i(x)$, we choose the smaller one. Also, let

$$\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}} = \{T(\bar{h}) \ : \ \bar{h} \in (\mathcal{H}_{\mathrm{bin}})^k\}.$$

What "should" be the Natarajan dimension of $\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}}$? Intuitively, to specify a hypothesis in $\mathcal{H}_{\mathrm{bin}}$ we need $d = \mathrm{VCdim}(\mathcal{H}_{\mathrm{bin}})$ parameters. To specify a hypothesis in $\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}}$, we need to specify k hypotheses in $\mathcal{H}_{\mathrm{bin}}$. Therefore, kd parameters should suffice. The following lemma establishes this intuition.

LEMMA 29.5 *If* $d = \mathrm{VCdim}(\mathcal{H}_{\mathrm{bin}})$ *then*

$$\mathrm{Ndim}(\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}}) \leq 3kd \log(kd).$$

Proof Let $C \subset \mathcal{X}$ be a shattered set. By the definition of shattering (for multiclass hypotheses)

$$\left|\left(\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}}\right)_C\right| \ \geq \ 2^{|C|}.$$

On the other hand, each hypothesis in $\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}}$ is determined by using k hypotheses from $\mathcal{H}_{\mathrm{bin}}$. Therefore,

$$\left|\left(\mathcal{H}_{\mathrm{bin}}^{\mathrm{OvA,k}}\right)_C\right| \ \leq \ |\,(\mathcal{H}_{\mathrm{bin}})_C\,|^k.$$

By Sauer's lemma, $|(\mathcal{H}_{\text{bin}})_C| \leq |C|^d$. We conclude that

$$2^{|C|} \;\leq\; \left|\left(\mathcal{H}_{\text{bin}}^{\text{OvA,k}}\right)_C\right| \;\leq\; |C|^{dk}.$$

The proof follows by taking the logarithm and applying Lemma A.1. □

How tight is Lemma 29.5? It is not hard to see that for some classes, $\text{Ndim}(\mathcal{H}_{\text{bin}}^{\text{OvA,k}})$ can be much smaller than dk (see Exercise 1). However there are several natural binary classes, $\mathcal{H}_{\text{bin}}$ (e.g., halfspaces), for which $\text{Ndim}(\mathcal{H}_{\text{bin}}^{\text{OvA,k}}) = \Omega(dk)$ (see Exercise 6).

29.3.2 General Multiclass-to-Binary Reductions

The same reasoning used to establish Lemma 29.5 can be used to upper bound the Natarajan dimension of more general multiclass-to-binary reductions. These reductions train several binary classifiers on the data. Then, given a new instance, they predict its label by using some rule that takes into account the labels predicted by the binary classifiers. These reductions include One-versus-All and All-Pairs.

Suppose that such a method trains l binary classifiers from a binary class $\mathcal{H}_{\text{bin}}$, and $r : \{0,1\}^l \to [k]$ is the rule that determines the (multiclass) label according to the predictions of the binary classifiers. The hypothesis class corresponding to this method can be defined as follows. For every $\bar{h} = (h_1, \ldots, h_l) \in (\mathcal{H}_{\text{bin}})^l$ define $R(\bar{h}) : \mathcal{X} \to [k]$ by

$$R(\bar{h})(x) = r(h_1(x), \ldots, h_l(x)).$$

Finally, let

$$\mathcal{H}_{\text{bin}}^r = \{R(\bar{h}) \;:\; \bar{h} \in (\mathcal{H}_{\text{bin}})^l\}.$$

Similarly to Lemma 29.5 it can be proven that:

LEMMA 29.6 *If* $d = \text{VCdim}(\mathcal{H}_{\text{bin}})$ *then*

$$\text{Ndim}(\mathcal{H}_{\text{bin}}^r) \;\leq\; 3\,l\,d\,\log(l\,d)\,.$$

The proof is left as Exercise 2.

29.3.3 Linear Multiclass Predictors

Next, we consider the class of linear multiclass predictors (see Section 17.2). Let $\Psi : \mathcal{X} \times [k] \to \mathbb{R}^d$ be some class-sensitive feature mapping and let

$$\mathcal{H}_\Psi = \left\{ x \mapsto \operatorname*{argmax}_{i \in [k]} \langle \mathbf{w}, \Psi(x,i) \rangle \;:\; \mathbf{w} \in \mathbb{R}^d \right\}. \tag{29.1}$$

Each hypothesis in $\mathcal{H}_\Psi$ is determined by d parameters, namely, a vector $\mathbf{w} \in \mathbb{R}^d$. Therefore, we would expect that the Natarajan dimension would be upper bounded by d. Indeed:

THEOREM 29.7 $\text{Ndim}(\mathcal{H}_\Psi) \leq d$.

Proof Let $C \subset \mathcal{X}$ be a shattered set, and let $f_0, f_1 : C \to [k]$ be the two functions that witness the shattering. We need to show that $|C| \leq d$. For every $x \in C$ let $\rho(x) = \Psi(x, f_0(x)) - \Psi(x, f_1(x))$. We claim that the set $\rho(C) \stackrel{\text{def}}{=} \{\rho(x) : x \in C\}$ consists of $|C|$ elements (i.e., ρ is one to one) and is shattered by the binary hypothesis class of homogeneous linear separators on $\mathbb{R}^d$,

$$\mathcal{H} = \{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}.$$

Since $\text{VCdim}(\mathcal{H}) = d$, it will follow that $|C| = |\rho(C)| \leq d$, as required.

To establish our claim it is enough to show that $|\mathcal{H}_{\rho(C)}| = 2^{|C|}$. Indeed, given a subset $B \subset C$, by the definition of shattering, there exists $h_B \in \mathcal{H}_\Psi$ for which

$$\forall x \in B, h_B(x) = f_0(x) \quad \text{and} \quad \forall x \in C \setminus B, h_B(x) = f_1(x).$$

Let $\mathbf{w}_B \in \mathbb{R}^d$ be a vector that defines h_B. We have that, for every $x \in B$,

$$\langle \mathbf{w}, \Psi(x, f_0(x)) \rangle > \langle \mathbf{w}, \Psi(x, f_1(x)) \rangle \;\Rightarrow\; \langle \mathbf{w}, \rho(x) \rangle > 0.$$

Similarly, for every $x \in C \setminus B$,

$$\langle \mathbf{w}, \rho(x) \rangle < 0.$$

It follows that the hypothesis $g_B \in \mathcal{H}$ defined by the same $\mathbf{w} \in \mathbb{R}^d$ label the points in $\rho(B)$ by 1 and the points in $\rho(C \setminus B)$ by 0. Since this holds for every $B \subseteq C$ we obtain that $|C| = |\rho(C)|$ and $|\mathcal{H}_{\rho(C)}| = 2^{|C|}$, which concludes our proof. □

The theorem is tight in the sense that there are mappings Ψ for which $\text{Ndim}(\mathcal{H}_\Psi) = \Omega(d)$. For example, this is true for the multivector construction (see Section 17.2 and the Bibliographic Remarks at the end of this chapter). We therefore conclude:

COROLLARY 29.8 *Let $\mathcal{X} = \mathbb{R}^n$ and let $\Psi : \mathcal{X} \times [k] \to \mathbb{R}^{nk}$ be the class sensitive feature mapping for the multi-vector construction:*

$$\Psi(\mathbf{x}, y) = [\ \underbrace{0, \ldots, 0}_{\in \mathbb{R}^{(y-1)n}}, \underbrace{x_1, \ldots, x_n}_{\in \mathbb{R}^n}, \underbrace{0, \ldots, 0}_{\in \mathbb{R}^{(k-y)n}}\ ].$$

Let $\mathcal{H}_\Psi$ be as defined in Equation (29.1). Then, the Natarajan dimension of $\mathcal{H}_\Psi$ satisfies

$$(k-1)(n-1) \;\leq\; \text{Ndim}(\mathcal{H}_\Psi) \;\leq\; kn.$$

29.4 On Good and Bad ERMs

In this section we present an example of a hypothesis class with the property that not all ERMs for the class are equally successful. Furthermore, if we allow an infinite number of labels, we will also obtain an example of a class that is

learnable by some ERM, but other ERMs will fail to learn it. Clearly, this also implies that the class is learnable but it does not have the uniform convergence property. For simplicity, we consider only the realizable case.

The class we consider is defined as follows. The instance space $\mathcal{X}$ will be any finite or countable set. Let $P_f(\mathcal{X})$ be the collection of all finite and cofinite subsets of $\mathcal{X}$ (that is, for each $A \in P_f(\mathcal{X})$, either A or $\mathcal{X} \setminus A$ must be finite). Instead of $[k]$, the label set is $\mathcal{Y} = P_f(\mathcal{X}) \cup \{*\}$, where $*$ is some special label. For every $A \in P_f(\mathcal{X})$ define $h_A : \mathcal{X} \to \mathcal{Y}$ by

$$h_A(x) = \begin{cases} A & x \in A \\ * & x \notin A \end{cases}$$

Finally, the hypothesis class we take is

$$\mathcal{H} = \{h_A \ : \ A \in P_f(\mathcal{X})\}.$$

Let $\mathcal{A}$ be some ERM algorithm for $\mathcal{H}$. Assume that $\mathcal{A}$ operates on a sample labeled by $h_A \in \mathcal{H}$. Since h_A is the *only* hypothesis in $\mathcal{H}$ that might return the label A, if $\mathcal{A}$ observes the label A, it "knows" that the learned hypothesis is h_A, and, as an ERM, must return it (note that in this case the error of the returned hypothesis is 0). Therefore, to specify an ERM, we should only specify the hypothesis it returns upon receiving a sample of the form

$$S = \{(x_1, *), \ldots, (x_m, *)\}.$$

We consider two ERMs: The first, $\mathcal{A}_{good}$, is defined by

$$\mathcal{A}_{good}(S) = h_\emptyset;$$

that is, it outputs the hypothesis which predicts '*' for every $x \in \mathcal{X}$. The second ERM, $\mathcal{A}_{bad}$, is defined by

$$\mathcal{A}_{bad}(S) = h_{\{x_1,\ldots x_m\}^c}.$$

The following claim shows that the sample complexity of $\mathcal{A}_{bad}$ is about $|\mathcal{X}|$-times larger than the sample complexity of $\mathcal{A}_{good}$. This establishes a gap between different ERMs. If $\mathcal{X}$ is infinite, we even obtain a learnable class that is not learnable by every ERM.

CLAIM 29.9

1. *Let $\epsilon, \delta > 0$, $\mathcal{D}$ a distribution over $\mathcal{X}$ and $h_A \in \mathcal{H}$. Let S be an i.i.d. sample consisting of $m \geq \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$ examples, sampled according to $\mathcal{D}$ and labeled by h_A. Then, with probability of at least $1 - \delta$, the hypothesis returned by $\mathcal{A}_{good}$ will have an error of at most ϵ.*
2. *There exists a constant $a > 0$ such that for every $0 < \epsilon < a$ there exists a distribution $\mathcal{D}$ over $\mathcal{X}$ and $h_A \in \mathcal{H}$ such that the following holds. The hypothesis returned by $\mathcal{A}_{bad}$ upon receiving a sample of size $m \leq \frac{|X|-1}{6\epsilon}$, sampled according to $\mathcal{D}$ and labeled by h_A, will have error $\geq \epsilon$ with probability $\geq e^{-\frac{1}{6}}$.*

Proof Let $\mathcal{D}$ be a distribution over $\mathcal{X}$ and suppose that the correct labeling is h_A. For any sample, $\mathcal{A}_{good}$ returns either $h_\emptyset$ or h_A. If it returns h_A then its true error is zero. Thus, it returns a hypothesis with error $\geq \epsilon$ only if all the m examples in the sample are from $\mathcal{X} \setminus A$ while the error of $h_\emptyset$, $L_{\mathcal{D}}(h_\emptyset) = \mathbb{P}_{\mathcal{D}}[A]$, is $\geq \epsilon$. Assume $m \geq \frac{1}{\epsilon}\log(\frac{1}{\delta})$; then the probability of the latter event is no more than $(1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta$. This establishes item 1.

Next we prove item 2. We restrict the proof to the case that $|\mathcal{X}| = d < \infty$. The proof for infinite $\mathcal{X}$ is similar. Suppose that $\mathcal{X} = \{x_0, \ldots, x_{d-1}\}$.

Let $a > 0$ be small enough such that $1 - 2\epsilon \geq e^{-4\epsilon}$ for every $\epsilon < a$ and fix some $\epsilon < a$. Define a distribution on $\mathcal{X}$ by setting $\mathbb{P}[x_0] = 1 - 2\epsilon$ and for all $1 \leq i \leq d-1$, $\mathbb{P}[x_i] = \frac{2\epsilon}{d-1}$. Suppose that the correct hypothesis is $h_\emptyset$ and let the sample size be m. Clearly, the hypothesis returned by $\mathcal{A}_{bad}$ will err on all the examples from $\mathcal{X}$ which are not in the sample. By Chernoff's bound, if $m \leq \frac{d-1}{6\epsilon}$, then with probability $\geq e^{-\frac{1}{6}}$, the sample will include no more than $\frac{d-1}{2}$ examples from $\mathcal{X}$. Thus the returned hypothesis will have error $\geq \epsilon$. □

The conclusion of the example presented is that in multiclass classification, the sample complexity of different ERMs may differ. Are there "good" ERMs for *every* hypothesis class? The following conjecture asserts that the answer is yes.

CONJECTURE 29.10 *The realizable sample complexity of every hypothesis class* $\mathcal{H} \subset [k]^{\mathcal{X}}$ *is*

$$m_{\mathcal{H}}(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{Ndim}(\mathcal{H})}{\epsilon}\right).$$

We emphasize that the $\tilde{O}$ *notation may hide only poly-log factors of* ϵ, δ*, and* $\mathrm{Ndim}(\mathcal{H})$*, but* no factor *of* k*.*

29.5 Bibliographic Remarks

The Natarajan dimension is due to Natarajan (1989). That paper also established the Natarajan lemma and the generalization of the fundamental theorem. Generalizations and sharper versions of the Natarajan lemma are studied in Haussler & Long (1995). Ben-David, Cesa-Bianchi, Haussler & Long (1995) defined a large family of notions of dimensions, all of which generalize the VC dimension and may be used to estimate the sample complexity of multiclass classification.

The calculation of the Natarajan dimension, presented here, together with calculation of other classes, can be found in Daniely et al. (2012). The example of good and bad ERMs, as well as conjecture 29.10, are from Daniely et al. (2011).

29.6 Exercises

1. Let $d, k > 0$. Show that there exists a binary hypothesis $\mathcal{H}_{\text{bin}}$ of VC dimension d such that $\text{Ndim}(\mathcal{H}_{\text{bin}}^{\text{OvA,k}}) = d$.
2. Prove Lemma 29.6.
3. Prove Natarajan's lemma.
   Hint: Fix some $x_0 \in \mathcal{X}$. For $i, j \in [k]$, denote by $\mathcal{H}_{ij}$ all the functions $f : \mathcal{X} \setminus \{x_0\} \to [k]$ that can be extended to a function in $\mathcal{H}$ both by defining $f(x_0) = i$ and by defining $f(x_0) = j$. Show that $|\mathcal{H}| \leq |\mathcal{H}_{\mathcal{X}\setminus\{x_0\}}| + \sum_{i\neq j} |\mathcal{H}_{ij}|$ and use induction.
4. Adapt the proof of the binary fundamental theorem and Natarajan's lemma to prove that, for some universal constant $C > 0$ and for every hypothesis class of Natarajan dimension d, the agnostic sample complexity of $\mathcal{H}$ is
$$m_{\mathcal{H}}(\epsilon, \delta) \leq C \frac{d \log\left(\frac{kd}{\epsilon}\right) + \log(1/\delta)}{\epsilon^2}.$$
5. Prove that, for some universal constant $C > 0$ and for every hypothesis class of Natarajan dimension d, the agnostic sample complexity of $\mathcal{H}$ is
$$m_{\mathcal{H}}(\epsilon, \delta) \geq C \frac{d + \log(1/\delta)}{\epsilon^2}.$$
   Hint: Deduce it from the binary fundamental theorem.
6. Let $\mathcal{H}$ be the binary hypothesis class of (nonhomogenous) halfspaces in $\mathbb{R}^d$. The goal of this exercise is to prove that $\text{Ndim}(\mathcal{H}^{\text{OvA,k}}) \geq (d-1) \cdot (k-1)$.
   1. Let $\mathcal{H}_{\text{discrete}}$ be the class of all functions $f : [k-1] \times [d-1] \to \{0, 1\}$ for which there exists some i_0 such that, for every $j \in [d-1]$
   $$\forall i < i_0, f(i, j) = 1 \text{ while } \forall i > i_0, f(i, j) = 0.$$
   Show that $\text{Ndim}(\mathcal{H}_{\text{discrete}}^{\text{OvA,k}}) = (d-1) \cdot (k-1)$.
   2. Show that $\mathcal{H}_{\text{discrete}}$ can be realized by $\mathcal{H}$. That is, show that there exists a mapping $\psi : [k-1] \times [d-1] \to \mathbb{R}^d$ such that
   $$\mathcal{H}_{\text{discrete}} \subset \{h \circ \psi \ : \ h \in \mathcal{H}\} .$$
   Hint: You can take $\psi(i, j)$ to be the vector whose jth coordinate is 1, whose last coordinate is i and the rest are zeros.
   3. Conclude that $\text{Ndim}(\mathcal{H}^{\text{OvA,k}}) \geq (d-1) \cdot (k-1)$.