State Space Realization Theorems For Data Mining

Robert L. Grossman* and Richard G. Larson
University of Illinois at Chicago

December 19, 2008

*Corresponding author: University of Illinois at Chicago, Department of Mathematics, Statistics & Computer Science (MC 249), 851 S. Morgan Street, Chicago IL 60607, grossman@uic.edu, 312 413 2176

Abstract

In this paper, we consider formal series associated with events, profiles derived from events, and statistical models that make predictions about events. We prove theorems about realizations for these formal series using the language and tools of Hopf algebras.

Keywords: realizations, formal series, learning sets, data mining, Myhill–Nerode Theorem, input-output maps, algebraic approaches to data mining

1 Introduction

Many data mining problems can be formulated in terms of events, profiles, models and predictions. As an example, consider the problem of predicting credit card fraud. In this application, there is a sequence of credit card transactions (called the learning set), each of which is associated with a credit card account and some of which have been labeled as fraudulent. The goal is to use the learning set to build a statistical model that predicts the likelihood that a credit card transaction is associated with a fraudulent account. Information about each credit card transaction is aggregated to produce a statistical profile (or state vector) about each credit card account. The profile consists of features. Applying the model to the profile produces a prediction about whether the account is likely to be fraudulent. Note that we can think of this example as a map from inputs (events) to outputs (predictions about whether the associated account is fraudulent). Given such an input-output map, we can ask whether there is a "realization" in which there is a state space of profiles (corresponding to accounts) in which each event updates the corresponding profile. We will see how to make this precise below.

Usually several different fraud models are developed and compared to one and another. Each fraud model is associated with a misclassification rate, which is the percent of fraudulent accounts that remain undetected. For many data mining applications, especially large-scale applications, we do not have a single learning set, but rather a collection of learning sets.

In this paper, we abstract this problem and use the language and tools of Hopf algebras to study it. To continue the example above, we abstract credit card transactions as *events*; state information about credit card accounts as *profiles*; credit card account numbers as *profile IDs or PIDs*; statistical models predicting the likelihood that a credit card account is fraudulent as *models*; a sequence of credit card transactions each of which is labeled either valid or fraudulent as *learning sets* of labeled events; and the accuracy rate of the credit card fraud model as the *classification rate* of the model.

We are interested in the following set up. Consider a collection $\mathcal{C}$, possibly infinite, of labeled learning sets w of events. For a labeled learning set w, we can build a model. Each model has a classification rate p_w. This information can be summarized in a formal series

$$p = \sum_{\mathcal{C}} p_w w$$

In this paper, we prove some theorems about these formal series using the language of Hopf algebras.

We now give the precise definitions we need. A *labeled event* is an event, together with a Profile Identifier (PID) and a label. Fix a set D of labeled events. We define a *labeled learning set of events* to be an element of $\mathcal{W}(D)$, the set of words $d_1 \cdots d_k$ of elements $d_i \in D$. If k is a field, then $H = k\mathcal{W}(D)$ is a k-algebra with basis $\mathcal{W}(D)$. In this paper we study formal series of the form

$$p = \sum_{w \in \mathcal{W}(D)} p_w w.$$

By a formal series, we mean a map

$$H \longrightarrow k,$$

associating to each element $w \in \mathcal{W}(D)$ the series coefficient p_w. The coefficient p_w is the classification (or misclassification) rate for the learning set of the events in w. Formal series occur in the formal theory of languages, automata theory, control theory, and a variety of other areas.

There is a more concrete realization of a model that we now describe. This requires a space X whose points $x \in X$ we interpret as profiles or states, which abstract the features used in a model. We can now define a model as a function from a space X of profiles that assigns a label (in k) to each element $x \in X$:

$$f : X \longrightarrow k.$$

Notice that given an initial profile $x_0 \in X$ associated with a PID, a sequence of events associated with a single PID will sweep out an orbit in X since each event will update the current profile in X associated with the PID. In the paper, we usually call the space X the *state space* and the initial profile the *initial state*.

Fix a formal series p. We investigate a standard question: given a formal series p built from the events D, is there a state space X, a (classification) model

$$f : X \longrightarrow k,$$

and a set of initial states that yield p. This is called a realization theorem. The state space captures the "essential" information in the data which is implicit in the series p. The formal definition is given below.

Realization theorems use a finiteness condition to imply the infinite object can be represented by a finite state space. One of the most familiar realization theorems is the Myhill–Nerode theorem. In this case, the infinite object is a formal series of words forming a language; the finiteness condition is the finiteness of a right invariant equivalence relation, and the state space is a finite automaton. In the case of data mining, the infinite object is a formal series of learning sets comprising a series of experiments, the finiteness condition is described by the finite dimensionality of a span of vectors, and the state space is $\mathbb{R}^n$.

The Myhill–Nerode theorem and more generally languages, formal series, automata, and finiteness conditions play a fundamental role in computer science. Our goal is to introduce analogous structures into data mining.

We now briefly recall the Myhill–Nerode theorem following [4, page 65]. Let the set D be an alphabet, $\mathcal{W}(D)$ be the set of words in D, and $L \subset \mathcal{W}(D)$ be a language. A language L defines an equivalence relation $\sim$ as follows: for $u, v \in \mathcal{W}(D)$, $u \sim v$ if and only if for all $w \in \mathcal{W}(D)$ either both or neither of uw and vw are in L. An equivalence relation $\sim$ is called *right invariant* with respect to concatenation in case $u \sim v$ implies $uw \sim vw$ for all $w \in \mathcal{W}(D)$.

Theorem 1.1 (Myhill–Nerode) *The following are equivalent:*

1. *L is the union of a finite number of equivalence classes generated by a right invariant equivalence relation.*
2. *The language $L \subset \mathcal{W}(D)$ is accepted by a finite automaton.*

We point out that in this case a language $L \subset \mathcal{W}(D)$ naturally defines a formal series. Fix a field k and the k-algebra $H = k\mathcal{W}(D)$. Given a language L, define the formal series p as follows:

$$p(w) = \begin{cases} 1 & \text{if } w \in L \\ 0 & \text{otherwise} \end{cases} .$$

Section 2 contains preliminary material. Section 3 constructs a finite state space X for the simple case of a formal series without profile identifiers or labels. Section 4 proves a theorem about parametrized classifiers and near to best realizations. Section 5 contains our main realization theorem.

One of the goals of this paper is to provide an algebraic foundation for some of the formal aspects of data mining. Other (non-algebraic) approaches can be found in [6], [7] and [1].

A short annoucement of the some of the results in this paper (without proofs) appeared in [3].

2 Preliminaries

Let D denote an *event space*. More precisely an element of D is a triple whose first element is a Profile IDentifier (PID) chosen from a finite set $\mathcal{I}$, whose second element is a label chosen from a finite set of labels $\mathcal{L}$, and whose third element is an element of S, a set of events associated with PIDs. In short, $D = \mathcal{I} \times \mathcal{L} \times S$, where $\mathcal{I}$ is the set of PIDs and $\mathcal{L}$ is the set of labels.

We use heavily the facts that $\mathcal{I}$ and $\mathcal{L}$ are finite sets.

We assume that S is a semigroup with unit 1 generated by $S_0 \subseteq S$. For example, S_0 might be a set of transactions and S might be sequences of transactions. Multiplication in S might be concatenation, or some operation related to the structure of the data represented by S.

A *labeled learning set* is an element of $\mathcal{W}(D)$, the set of words $w = d_1 \cdots d_k$ of events in D.

A *labeled learning sequence* is a sequence $\{w_1, w_2, \ldots\}$ of labeled learning sets; a corresponding *formal labeled learning series* is a formal series

$$\sum_w p_w w.$$

Let $H = k\mathcal{W}(D)$ denote the vector space with basis $\mathcal{W}(D)$, and kS denote the vector space with basis S. Then H is an algebra whose multiplication is induced by the semigroup structure of $\mathcal{W}(D)$, which is simply concatenation, and $U = kS$ is an algebra whose structure is induced by the semigroup structure of S.

Let $\overline{H}$ denote the space of formal labeled learning series. For $(i, \ell) \in \mathcal{I} \times \mathcal{L}$ define the map $\pi_{(i,\ell)} : \overline{H} \longrightarrow U^*$ as follows: first, define $\pi_{(i,\ell)}(p)(s) = p((i, \ell, s))$ for $p \in D$ and $s \in S$; then, extend $\pi_{(i,\ell)}$ to $\mathcal{W}(D)$ multiplicatively;

We have that $U = kS$ is a bialgebra, with coproduct given by $\Delta(s) = 1 \otimes s + s \otimes 1$ for $s \in S_0$, and with augmentation ϵ defined by $\epsilon(1) = 1$, $\epsilon(s) = 0$ for all non-identity elements $s \in S$. We will view S as acting on a state space. Since U is primitively generated, $U \cong U(P(U))$ (recall that $P(U) = \{\, x \in U \mid \Delta(x) = 1 \otimes x + x \otimes 1 \,\}$ is a Lie algebra, and that $U(L)$ is the universal enveloping algebra of the Lie algebra L [5]). We put a bialgebra structure on H by letting $\Delta((i, \ell, s)) = \sum_{(s)} (i, \ell, s_{(1)}) \otimes (i, \ell, s_{(2)})$ where $\Delta(s) = \sum_{(s)} s_{(1)} \otimes s_{(2)}$, and $\epsilon((i, \ell, s)) = \epsilon(s)$, for $i \in \mathcal{I}$, $\ell \in \mathcal{L}$, $s \in S$, and extending multiplicatively to $\mathcal{W}(D)$

A *simple formal learning series* is an element $p \in U^*$. We can think of a simple learning series p as an infinite series $\sum_{s \in S} c_s s$. Essentially, a simple formal learning series is a formal labeled learning series, but without the labels and PIDs.

3 Construction of the state space

We are concerned whether $p \in U^*$, or some finite set $\{p_\alpha\} \subset U^*$, arises from a finite dimensional state space X. The reason we work with a finite set of elements of U^* rather than with a single one is that this allows us to deal with individual profiles that get aggregated into the full dataset.

Since U is primitively generated, we know that $U \cong U(P(U))$.

Remark 3.1 If H is any bialgebra, we have a left H-module action of H on H^* defined by $h \rightharpoonup p(k) = p(kh)$ for $p \in H^*$, h, $k \in H$, and a right H-module action of H on H^* defined by $p \leftharpoonup h(k) = p(hk)$ for $p \in H^*$, h, $k \in H$.

The following definition is from [2].

Definition 3.2 *We say that the simple formal learning series $p \in U^*$ has* finite Lie rank *if* $\dim P(U) \rightharpoonup p$ *is finite.*

Let R be a commutative algebra with augmentation ϵ, and let $f \in R$. We say that $p \in U^*$ is differentially produced by the pair (R, f) *if*

1. *there is right U-module algebra structure $\cdot$ on R;*
2. *$p(u) = \epsilon(f \cdot u)$ for $u \in U$.*

A basic theorem on the existence of a state space is the following, which is a generalization of Theorem 1.1 in [2]. In this theorem, the state space is a vector space with basis $\{x_1, \ldots, x_n\}$.

Theorem 3.3 *Let $p_1, \ldots, p_r \in U^*$. Then the following are equivalent:*

1. *p_k has finite Lie rank for $k = 1, \ldots, r$;*
2. *there is an augmented algebra R for which* $\dim (\mathrm{Ker}\, \epsilon)/(\mathrm{Ker}\, \epsilon)^2$ *is finite, and for all k, there is $f_k \in R$ such that p_k is differentially produced by the pair (R, f_k);*
3. *there is a subalgebra R of U^* which is isomorphic to $k[[x_1, \ldots, x_n]]$, the algebra of formal power series in n variables, and for all k, there is $f_k \in R$ such that p_k is differentially produced by the pair (R, f_k).*

Proof: We first prove that part (1) of Theorem 3.3 implies part (3). Given $p_1, \ldots, p_r \in U^*$, we define three basic objects:

$$\begin{aligned} L &= \{ u \in P(U) \mid u \rightharpoonup p_k = 0, \text{ for all } k, \} \\ J &= UL \\ J^\perp &= \{ q \in U^* \mid q(j) = 0 \text{ for all } j \in J \}. \end{aligned}$$

Since $L \subseteq P(U)$, it follows that J is a coideal, that is, that $\Delta(J) \subseteq J \otimes U + U \otimes J$. Therefore $J^\perp \cong (U/J)^*$ is a subalgebra of U^*. We will show that $J^\perp$ is isomorphic to a formal power series algebra.

Lemma 3.4 *If* $\dim \sum_k P(U) \rightharpoonup p_k = n$*, then* $J^\perp$ *is a subalgebra of* U^* *satisfying*

$$J^\perp \cong k[[x_1, \ldots, x_n]].$$

Proof: Note that L is the kernel of the map

$$P(U) \longrightarrow \bigoplus_k P(U) \rightharpoonup p_k, \qquad u \mapsto \bigoplus_k u \rightharpoonup p_k.$$

and L has finite codimension n. Choose a basis $\{e_1, e_2, \ldots\}$ of $P(U)$ such that $\{e_{n+1}, e_{n+2}, \ldots\}$ is a basis of L. Note that if $\bar{e}_i$ is the image of e_i under the quotient map $P(U) \to P(U)/L$, then $\{\bar{e}_1, \ldots, \bar{e}_n\}$ is a basis for $P(U)/L$.

By the Poincaré-Birkhoff-Witt Theorem, U has a basis of the form

$$\{ e_{i_1}^{\alpha_{i_1}} \cdots e_{i_k}^{\alpha_{i_k}} \mid i_1 < \cdots < i_k \text{ and } 0 < \alpha_{i_r} \}.$$

Since the basis $\{e_i\}$ of $P(U)$ has been chosen so that $e_i \in L$ for $i > n$, it follows that the monomials $\{ e_1^{\alpha_1} \cdots e_n^{\alpha_n} \mid \alpha_k \geq 0 \}$ are a basis for a vector space complement to J. It follows that

$$\{ \bar{e}_1^{\alpha_1} \cdots \bar{e}_n^{\alpha_n} \mid \alpha_1, \ldots, \alpha_n \geq 0 \}$$

is a basis for U/J. It now follows that the elements

$$x_\alpha = \frac{x^\alpha}{\alpha!} = \frac{x_1^{\alpha_1} \cdots x_n^{\alpha_n}}{\alpha_1! \cdots \alpha_n!}$$

are in $J^\perp \subseteq U^*$, where $x_i \in U^*$ is defined by

$$x_i(e_{i_1}^{\alpha_{i_1}} \cdots e_{i_k}^{\alpha_{i_k}}) = \begin{cases} 1 & \text{if } e_{i_1}^{\alpha_{i_1}} \cdots e_{i_k}^{\alpha_{i_k}} = e_i, \\ 0 & \text{otherwise.} \end{cases}$$

The subalgebra $J^\perp$ consists precisely of the closure in U^* of the span of these elements. In other words,

$$J^\perp \cong k[[x_1, \ldots, x_n]],$$

completing the proof.

We will use the following facts from the proof of Lemma 3.4: suppose that $\{e_1, \ldots, e_n, \ldots\}$ is a basis for $P(U)$ such that $\{e_{n+1}, \ldots\}$ is a basis for L. Let $\{e^\alpha\}$ be the corresponding Poincaré-Birkhoff-Witt basis. Denote $J^\perp$ by R. Then $R \cong k[[x_1, \ldots, x_n]]$, and $x_1^{\alpha_1} \cdots x_n^{\alpha_n}/\alpha_1! \cdots \alpha_n!$ is the element of the dual (topological) basis of U^* to the Poincaré-Birkhoff-Witt basis $\{e^\alpha\}$ of U, corresponding to the basis element $e_1^{\alpha_1} \cdots e_n^{\alpha_n}$.

We now collect some properties of the ring of formal power series R which will be necessary for the proof of Theorem 3.3.

Lemma 3.5 *Assume $p \in U^*$ has finite Lie rank, and let $R \subseteq U^*$, $e^\alpha \in U$, and $x^\alpha \in R$ be as in Lemma 3.4. Define*

$$f = \sum_{\alpha=(\alpha_1,\ldots,\alpha_n)} c_\alpha x^\alpha \in R,$$

where $c_\alpha = \dfrac{p(e^\alpha)}{\alpha!}$. Then

1. *U measures R to itself via $\leftharpoonup$;*
2. *$p(u) = \epsilon(f \leftharpoonup u)$ for all $u \in U$.*

Proof: We begin with the proof of part (1). Since U measures U^* to itself and $R \subseteq U^*$, we need show only that $R \leftharpoonup U \subseteq R$. Take $r \in R$, $u \in U$ and $j \in J$. We have $(r \leftharpoonup u)(j) = r(uj)$. Since J is a left ideal, $uj \in J$, so $r(uj) = 0$, so $r \leftharpoonup u \in J^\perp = R$. This proves part (1).

We now prove part (2). Let $e^\alpha = e_{i_1}^{\alpha_{i_1}} \cdots e_{i_k}^{\alpha_{i_k}}$ be a Poincaré-Birkhoff-Witt basis element of U. Since $e^\alpha \in J$ unless $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, $p(e^\alpha) = 0$ unless $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$. Also $\epsilon(f \leftharpoonup e^\alpha) = f \leftharpoonup e^\alpha(1) = f(e^\alpha 1) = f(e^\alpha) = 0$ unless $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$. Now suppose $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$. We have in this case that $p(e^\alpha) = \alpha! c_\alpha = f(e^\alpha) = f \leftharpoonup e^\alpha(1) = \epsilon(f \leftharpoonup e^\alpha)$. Since $\{e^\alpha\}$ is a basis for U, this completes the proof of part (2) of the lemma.

Corollary 3.6 *Under the assumptions of Lemma 3.5, $f = p$.*

Lemmas 3.4 and 3.5 yield that part (1) implies part (3) in Theorem 3.3. It is immediate that part (3) implies part (2).

We now complete the proof of Theorem 3.3 by proving that part (2) implies part (1).

Let $x_1, \ldots, x_n \in \operatorname{Ker}\epsilon$ be chosen so that $\{\bar{x}_1, \ldots, \bar{x}_n\}$ is a basis for $(\operatorname{Ker}\epsilon)/(\operatorname{Ker}\epsilon)^2$. If $f \in R$ and $u \in U$, then

$$f \cdot u = q_0(u)1 + \sum_{i=1}^{n} q_i(u)x_i + g(u),$$

where $q_i \in U^*$ and $g(u) \in (\operatorname{Ker}\epsilon)^2$. Let $\ell \in P(U)$. Since U measures R to itself and $\Delta(\ell) = 1 \otimes \ell + \ell \otimes 1$, the map $f \mapsto f \cdot \ell$ is a derivation of R.

Now let $f_k \in R$ be the element such that

$$p_k(u) = \epsilon(f_k \cdot u).$$

Then

$$\begin{aligned} f_k \cdot ul &= (f_k \cdot u) \cdot \ell \\ &= q_{k,0}(u)1 \cdot l + \sum_{j=1}^{n} q_{k,j}(u)x_j \cdot l + g_k(u) \cdot l. \end{aligned}$$

Since the map $f \mapsto f \cdot \ell$ is a derivation, $1 \cdot \ell = 0$, and since $g_\alpha(u) \in (\operatorname{Ker}\epsilon)^2$, $g_\alpha(u) \cdot \ell \in \operatorname{Ker}\epsilon$. It follows that

$$\begin{aligned} \ell \rightharpoonup p_k(u) &= p_k(u\ell) \\ &= \epsilon(f_k \cdot u\ell) \\ &= \sum_{j=1}^{n} q_{k,j}(u)\epsilon(x_j \cdot \ell). \end{aligned}$$

Therefore $P(U) \rightharpoonup p_k \subseteq \sum_{j=1}^{n} kq_j$, so p_k has finite Lie rank. This completes the proof of Theorem 3.3

Definition 3.7 *A series $p \in H$ for which the set*

$$\{\, p_{(i,\ell)} = \pi_{(i,\ell)}(p) \mid \ell \in \mathcal{L},\ i \in \mathcal{I} \,\}$$

satisfies the conditions of Theorem 3.3 is called regular.

We have shown how to construct a state space X and a right U-module algebra R of observations for a regular series.

Although the R we have constructed is a power series algebra, for applications we will often use some other right U-module algebra of functions on X. We will assume that we have an action of S on X which induces the action of U on R, that is, that R is a U-module algebra.

4 Learning sets of profiles and realizations

Let $\mathcal{I}$ be a finite set of PIDs, and let $\mathcal{L}$ be a finite set of labels. Let H and U be the bialgebras described in Section 2, X be the corresponding state space as described in Section 3, and R be a right U-module algebra of functions from X to k.

Definition 4.1 *A* classifier *is a function* $f : X \longrightarrow \mathcal{L}$*. A* learning set of profiles *is a function* $\chi : \mathcal{I} \longrightarrow \mathcal{L} \times X$*, that is, a finite set* $\{(\ell_j, x_j)\}$*, where* $\ell_j \in \mathcal{L}$*,* $x_j \in X$*, and* $j \in \mathcal{I}$*.*

Note that a classifier is a model as defined in Section 1. We denote the set of classifiers by $\mathcal{F}$ and the set of learning sets of profiles by $\mathcal{C}$. $\mathcal{W}(D)$ acts on $\mathcal{C}$ as follows. If $d = (i, l, s) \in D$ and $\chi = \{(l_j, x_j)\}$, define $\chi \cdot d = \{(\ell_j, x_j) \cdot d\}$, where

$$(\ell_j, x_j) \cdot d = \begin{cases} (\ell, x_j \cdot s) & \text{if } i = j \\ (l_j, x_j) & \text{otherwise.} \end{cases}$$

That is, the event $d = (i, \ell, s)$ acts on the learning set of profiles $\chi = \{(l_j, x_j)\}$ by acting on the individual points (ℓ_j, x_j) as follows: if $j \neq i$ the point is unchanged; if $j = i$ the point x_j is moved to $x_j \cdot s$ and the label is changed to ℓ.

A *pairing* $\ll f, \chi \gg$ between classifiers and learning sets of profiles can be given as follows. Let $f : X \longrightarrow \mathcal{L}$ be a classifier, and $\chi = \{(\ell_i, x_i)\}$ be a learning set of profiles. Then

$$\ll f, \chi \gg = \frac{|\{\, i \in \mathcal{I} \mid f(x_i) = \ell_i \,\}|}{|\mathcal{I}|}. \tag{1}$$

Note that $0 \leq \ll f, \chi \gg \leq 1$. This pairing is a measure of how well the classifier f predicts the actual data represented by χ.

We define the notion of *realization* as follows.

Definition 4.2 *Let*

$$f : X \longrightarrow \mathcal{L}$$

be a classifier, let

$$\chi : \mathcal{I} \longrightarrow \mathcal{L} \times X$$

be a learning set of profiles, and let $\ll -, - \gg$ *be a pairing. We say that the triple* (X, f, χ) *is a* realization *of the series* $p \in \overline{H}$ *if*

$$p_h = \ll f, \chi \cdot h \gg.$$

Note that the classifier $\ll f, \chi \cdot h \gg$ defined in Equation (1) is bounded, in fact

$$0 \leq \ll f, \chi \cdot h \gg \leq 1.$$

Recall that $p = \sum_{h \in \mathcal{W}(D)} p_h h$ is the formal series of which we are studying realizations.

Lemma 4.3 *Fix a finite learning set* χ*, and fix* $A \subseteq \mathbb{R}^n$*. Suppose that there is a map* $M : A \longrightarrow \mathcal{F}$ *such that* $\ll M(a), \chi \cdot h \gg$ *is a bounded function of* $a \in A$*, and* $p \in \overline{H}$ *for which there is a state space* X *and a ring of functions* R *as described in section 3. Assume that* p_h*,* $h \in \mathcal{W}(D)$*, is bounded. Let*

$$\widetilde{M}(a) = \sup_{h \in \mathcal{W}(D)} |p_h - \ll M(a), \chi \cdot h \gg|,$$

Then for all $\epsilon > 0$ *there exists* $a_0 \in A$ *such that* $|\widetilde{M}(a_0) - \inf_{a \in A} \widetilde{M}(a)| < \epsilon$*.*

Note that the hypothesis on M includes models which are polynomials, tree classifiers, neural nets, and splines.

Proof:

Since everything in its definition is bounded, $\widetilde{M}(a)$ exists and is bounded. If $P : A \longrightarrow \mathbb{R}$ is any bounded function, then there is $a_0 \in A$ such that $P(a_0)$ is within ϵ of $\inf_{a \in A} P(a)$.

Note that for any realization $M(a)$ of p, we have that

$$\widetilde{M}(a) = \sup_{h \in \mathcal{W}(D)} |p_h - \ll M(a), \chi \cdot h \gg|$$

measures how well $M(a)$ realizes p. so that $\inf_{a \in A} \widetilde{M}(a)$ is the lower bound for the "goodness" of any realization. The lemma says that this lower bound can be approximated arbitrarily closely.

Theorem 4.4 *Let $p : k\mathcal{W}(D) \longrightarrow k$ be such that p_h is bounded, and let $M : A \longrightarrow \mathcal{F}$ be a parametrized classifier such that $\ll M(a), \chi \gg$ is a bounded function of a. Then for all $\epsilon > 0$ there is a realization $p_0 = M(a_0)$ of p such that the "goodness" of the realization afforded by p_0 is within ϵ of the lower bound, that is, $|\widetilde{M}(a_0) - \inf_{a \in A} \widetilde{M}(a)| < \epsilon$.*

PROOF:

Theorem 4.4 follows immediately from Corollary 4.3.

5 Parametrized realizations

In this section we consider an event space D, a realizable labeled learning series p, a state space X, and an algebra of functions R from the state space X to k.

Denote by $\mathcal{C}$ learning sets of profiles and denote by $\mathcal{F}$ the set of functions from X to the finite set of labels $\mathcal{L}$. Fix a vector space of parameters A, and a map

$$M : A \longrightarrow \mathcal{F}$$

giving a parametrized family of models.

In this section we study parametrized realizations of formal series $p \in \overline{H}$ of learning sets.

Compare Definition 5.1 to Definition 4.2 in which realizations are defined.

Definition 5.1 *A* parametrized realization *of a bounded function $p \in \overline{H}$ is:*

1. *A vector space of parameters A.*

2. *A parametrized family of models $M : A \longrightarrow \mathcal{F}$.*

*If A is a finite dimensional vector space, we say that the realization is A-*finite*.*

Theorem 5.2 below gives a finiteness condition on the action of A on $p \in \overline{H}$ which gives an A-finite realization.

For $f = M(a) \in \mathcal{S}$ and $\ell \in \mathcal{L}$, let f_ℓ be defined by

$$f_\ell(x) = \begin{cases} l & \text{if } f(x) = \ell, \\ \star & \text{otherwise,} \end{cases}$$

where $\star$ is unequal to any label $\ell \in \mathcal{L}$. Let $p_\ell(h)$ be defined by

$$p_\ell(h) = \ll f_\ell, \chi \cdot h \gg.$$

Theorem 5.2 *Let $p \in H^*$ be a formal sum of learning sets and $M : A \longrightarrow \mathcal{F}$ a family of labeled models parametrized by A. Assume:*

1. *there exists $f \in \operatorname{Im} M$ such that*

$$p(h) = \ll f, \chi \cdot h \gg,$$

2. *$\{ \beta \in A^* \mid \beta \rightharpoonup p_\ell = 0 \}$ is a subspace of A^* of finite codimension which is closed in the compact open topology for all $\ell \in \mathcal{L}$.*

Then there exists an A-finite realization of p.

Note that Theorem 4.4 gives the existence of a realization which approximates the desired one.

PROOF: We define three basic objects:

$$\begin{aligned} L_\ell &= \{ \beta \in A^* \mid \beta \rightharpoonup p_\ell = 0 \} \\ J_\ell &= k[A^*] L_\ell \\ J_\ell^\perp &= \{ q \in k[A^*]^* \mid q(j) = 0 \text{ for all } j \in J_\ell \}. \end{aligned}$$

We have that J_ℓ is a coideal in the Hopf algebra $k[A^*]$ generated by primitive elements in A^*, that is, that $\Delta(J_\ell) \subseteq J_\ell \otimes k[A^*] + k[A^*] \otimes J_\ell$. Therefore $J_\ell^\perp \cong (k[A^*]/J_\ell)^*$ is a subalgebra of $k[A^*]^*$. We will show that $J_\ell^\perp$ is isomorphic to a formal power series algebra in finitely many variables.

From hypothesis (2) we have that $L_\ell^\perp = (A^*/L_\ell)^*$ is finite dimensional subspace of A.

Lemma 5.3 *If $\dim L_\ell^\perp = n_\ell$, then $J_l^\perp$ is a subalgebra of $k[A^*]^*$ satisfying*

$$J_l^\perp \cong k[[a_1, \ldots, a_{n_\ell}]],$$

where $\{a_1, \ldots, a_{n_\ell}\}$ is a basis for $L_\ell^\perp$.

Proof: The subspace L_ℓ is a closed subspace of A^* of finite codimension n, so that $(A^*/L_\ell)^* = L_\ell^\perp$ is a finite dimensional subspace of A. Let $(A^*/L_\ell)^*$ have basis $\{a_1, \ldots, a_{n_\ell}\}$. Choose $\beta_{a_i} \in A^*$ with $\beta_{a_i}(a_j) = \delta_{ij}$. Now choose a basis $\mathcal{B}$ of A^* such that $\mathcal{B} \supseteq \{\beta_{a_1}, \ldots \beta_{a_n}\}$ and $\mathcal{B}' = \mathcal{B} \setminus \{\beta_{a_1}, \beta_{a_2}, \ldots, \beta_{a_{n_\ell}}\}$ is a basis of L_ℓ. We have that $k[A^*]$ has a basis

$$\{\, \beta_{i_1}^{\alpha_{i_1}} \cdots \beta_{i_k}^{\alpha_{i_k}} \mid \beta_k \in \mathcal{B},\ i_1 < \cdots < i_k,\ \text{and } 0 < \alpha_{i_r} \,\}.$$

By the choice of the basis of A^*, J_ℓ will have a basis of the form

$$\beta_{i_1}^{\alpha_{i_1}} \cdots \beta_{i_k}^{\alpha_{i_k}}$$

with at least one $\beta_k \in \mathcal{B}'$. It follows that

$$\{\, \beta_{a_1}^{\alpha_1} \cdots \beta_{a_{n_\ell}}^{\alpha_{n_\ell}} \mid \alpha_1, \ldots, \alpha_{n_\ell} \geq 0 \,\},$$

where we denote by β_{a_k} the image of that element in $k[A^*]/J_\ell$, is a basis for $k[A^*]/J_\ell$. It now follows that elements of the form

$$a^\alpha = a_{i_1}^{\alpha_{i_1}} \cdots a_{i_{n_\ell}}^{\alpha_{i_{n_\ell}}}$$

are in $J_l^\perp \subseteq U^*$. and that $J_l^\perp$ consists precisely of the closure in $k[A^*]^*$ of the span of such elements. In other words,

$$J_\ell^\perp \cong k[[a_1, \ldots, a_{n_\ell}]],$$

completing the proof.

By Lemma 5.3 each p_ℓ depends on a finite dimensional space of parameters $A_{(\ell)}$. Let A^0 be the finite dimensional subspace which is the spanned by the union of these finite dimensional subspaces. Since $p(h) = \sum_{\ell \in \mathcal{L}} p_\ell(h)$, $p(h)$ depends only on parameters in A^0.

Now $\ll f, \chi \cdot h \gg$ depends only on parameters in A^0. We may choose the other parameters which are linearly independent from A^0 arbitrarily. In other words we may choose f_0 so that it depends only on the parameters in A^0.

This completes the proof of Theorem 5.2.

References

[1] F. Cucker and S. Smale, *On the Mathematical Foundations of Learning,* Bull. Amer. Math. Soc. (N.S.), **39** (2002), 1–49.

[2] R. L. Grossman and R. G. Larson, *The realization of input-output maps using bialgebras,* Forum Math., **4** (1992), 109–121.

[3] R. L. Grossman and R. G. Larson, *An Algebraic Approach to Data Mining: Some Examples,* Proc. 2002 IEEE Int. Conf. Data Min. (ICDM 2002).

[4] John E. Hopcroft and Jeffrey D. Ullman, "Introduction to Automata Theory, Languages and Computation," Addison–Wesley, Reading, Massachusetts, 1979.

[5] Nathan Jacobson, "Lie Algebras," Dover Publications, New York, 1979.

[6] L. G. Valiant, *A Theory of the Learnable,* Comm. ACM, **27**, 1134–1142, 1984.

[7] V. Vapnik, "Statistical Learning Theory," John Wiley and Sons, 1998.