Probabilistic Classifiers and the Concepts they Recognize

Manfred Jaeger JAEGER@MPI-SB.MPG.DE
Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

Abstract

We investigate algebraic, logical, and geometric properties of concepts recognized by various classes of probabilistic classifiers. For this we introduce a natural hierarchy of probabilistic classifiers, the lowest level of which comprises the naive Bayesian classifiers. We show that the expressivity of classifiers on the different levels in the hierarchy is characterized algebraically by separability with polynomials of different degrees. A consequence of this result is that every linearly separable concept can be recognized by a naive Bayesian classifier. We contrast this result with negative results about the naive Bayesian classifier previously reported in the literature, and point out that these results only pertain to specific learning scenarios for naive Bayesian classifiers. We also present some logical and geometric characterizations of linearly separable concepts, thus providing additional intuitive insight into what concepts are recognizable by naive Bayesian classifiers.

1. Introduction

In spite of the very simplistic assumptions it is based on, the naive Bayesian classifier has proven very useful in practice. Many studies have been conducted to analyze this seemingly paradoxical success (Domingos & Pazzani, 1997; Rish et al., 2001; Garg & Roth, 2001). In spite of the wealth of material on the naive Bayesian classifier that is now available, there still seems to be some confusion as to some of its very fundamental properties. In particular, there appear to be some misunderstandings relating to the class of concepts it can recognize. While it is common knowledge that it can only recognize linearly separable concepts, there also appears to be a widespread belief, that not all linearly separable concepts can be recognized by a naive Bayesian classifier.

In large part the apparent confusion is due to the fact that it is not always unambiguously clear, what individual authors mean when they say that "the [naive] Bayesian classifier cannot learn some linearly separable concepts" (Domingos & Pazzani, 1997), or "not all linearly separable functions can be represented using this [naive Bayes] predictor" (Roth, 1998). Without a very careful analysis of the context in which such a statement is made, one would take it as negative answer to the following question:

> *(Q1)* Can every linearly separable concept be recognized by a naive Bayesian classifier, i.e. does there exist for every linearly separable concept C a naive Bayesian classifier that will assign to every example $\boldsymbol{x}$ class label $\oplus$ iff $\boldsymbol{x}$ belongs to C?

However, the negative statements one finds in the literature really are answers to the following question:

> *(Q2)* Is for every linearly separable concept A a naive Bayesian classifier learnable by maximum likelihood inference from an enumeration of A?

We require some formal definitions before, in section 3 we can recast *(Q2)* in precise terms. For the time being, it only needs to be realized that here the question is whether a classifier recognizing A will be learned using a specific learning principle, and a specific data set. The significance of asking *(Q2)* rather than *(Q1)* has sometimes been realized: Zhang, Ling and Zhao (2000) remark that the negative results of (Domingos & Pazzani, 1997) depend on their assumption of a uniform sampling distribution for the examples, and present empirical evidence that with data sampled from non-uniform distributions the class of learnable concepts changes. Conversely, Roth (1999) remarks that the positive results of (Domingos & Pazzani, 1997) on the learnability of pure disjunctive or conjunctive concepts, too, depend on a uniform sampling assumption.

To the best of our knowledge, so far no one has addressed the arguably much more pertinent question *(Q1)*. In this paper we will provide an affirmative answer to this question. Indeed, we will provide a much more general result: we introduce a hierarchy of probabilistic classifiers, the lowest level of which exactly corresponds to naive Bayesian clas-

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

sifiers, and show an exact correspondence between the levels of this hierarchy and separability by polynomials with different degrees. We thus obtain results that also are applicable to other classes of probabilistic classifiers, e.g. the class of *tree augmented naive Bayesian* classifiers (Friedman et al., 1997).

We also provide some results on logical and topological characterizations of concept classes recognizable with the probabilistic classifiers in our hierarchy. These results are much more mixed, and only lead to necessary, but not to sufficient conditions for recognizability.

2. Definitions and Main Result

Definition 2.1 Let $X_1, \ldots, X_n, C$ be a set of binary random variables. The random variables X_i are called features, and C is called the class variable. We denote the two possible values of the X_i with 0 and 1, and the two possible values of C with $\oplus$ and $\ominus$. We also use $\mathbb{F}^n$ to denote the space $\{0,1\}^n$ of all possible value assignments to the features (the instance space). A *Bayesian probabilistic classifier* for the class variable C given the features X_i is given by a probability distribution P on $\mathbb{F}^n \times \{\oplus, \ominus\}$. A *classical probabilistic classifier* is given by a pair $P_\oplus$ and $P_\ominus$ of probability distributions on $\mathbb{F}^n$.

A Bayesian classifier assigns to an *instance* $\boldsymbol{x} \in \mathbb{F}^n$ the class label $\oplus$ iff

$$P(C = \oplus \mid \boldsymbol{x}) \geq P(C = \ominus \mid \boldsymbol{x}), \tag{1}$$

whereas a classical classifier assigns class label $\oplus$ iff

$$P_\oplus(\boldsymbol{x}) \geq P_\ominus(\boldsymbol{x}) \tag{2}$$

Any subset of $\mathbb{F}^n$ is called a *concept*. The concept *recognized by a classifier* is the set of instances to which class label $\oplus$ is assigned.

We note that in the case of the classical classifier, the class variable C is not really considered a random variable, but a parameter governing the distribution of the features. The definitions of Bayesian and classical classifiers are very similar, and we will see in the sequel that for the types of classifiers we consider there is no substantial difference between these two variations in the definitions. In fact, the only difference is the following: a classical classifier cannot recognize the empty concept, as (2) must be satisfied for at least one $\boldsymbol{x}$. The Bayesian classifier, on the other hand, can recognize the empty concept by placing a high prior probability on the class label $\ominus$. This, of course, is not a fundamental difference, but only an artefact of the definitions given here.

Example 2.2 For $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{F}^n$ we denote with $1(\boldsymbol{x})$ the set $\{i \mid 1 \leq i \leq n,\ x_i = 1\}$. The *m-of-n* concept is the set $A^{m-n} := \{\boldsymbol{x} \in \mathbb{F}^n \mid |1(\boldsymbol{x})| \geq m\}$.

We construct a (classical) probabilistic classifier that recognizes A^{m-n} as follows (for the time being we ignore the question whether, and how, this classifier is learnable from data; we return to that question in section 3). For $i = 1, \ldots, n$ let

$$P_\oplus(X_i = 1) = m/n \quad P_\ominus(X_i = 1) = (m-1)/n,$$

and define the joint distributions as the products

$$P_\circ(\boldsymbol{x}) = \prod_{i=1}^{n} P_\circ(X_i = x_i) \quad (\circ = \oplus, \ominus) \tag{3}$$

This is a "naive classical" classifier that can be turned into an equivalent naive Bayesian version by assuming prior class probabilities $P(C = \oplus) = P(C = \ominus) = 1/2$.

It is readily verified that this classifier recognizes A^{m-n}. For this, we only have to observe that the distributions $P_\oplus$ and $P_\ominus$ are of the binomial form

$$P_\circ(\boldsymbol{x}) = \lambda_\circ^{|1(\boldsymbol{x})|}(1-\lambda_\circ)^{n-|1(\boldsymbol{x})|} \tag{4}$$

with parameters $\lambda_\oplus = m/n$ and $\lambda_\ominus = (m-1)/n$, respectively. For $\boldsymbol{x}$ with $|1(\boldsymbol{x})| = l$ we have that (4) is maximized by $\lambda_\circ = l/n$, and that (4) is monotonically increasing in $\lambda_\circ$ on $[0, l/n]$, and monotonically decreasing on $[l/n, 1]$. For $l = m$ it immediately follows that $P_\oplus(\boldsymbol{x}) > P_\ominus(\boldsymbol{x})$. If $l > m$ the same result follows from the monotonicity of (4) on $[0, l/n]$, and the fact that $\lambda_\ominus < \lambda_\oplus < l/n$. Similarly, for $l < m$, we obtain that $P_\oplus(\boldsymbol{x}) < P_\ominus(\boldsymbol{x})$ from the monotonicity of (4) on $[l/n, 1]$, and the fact that $l/n \leq \lambda_\ominus < \lambda_\oplus$.

The class of naive Bayesian classifiers is determined by condition (3) on their defining distributions. More generally, any restriction on the form of the admissible distributions $P_\circ(\cdot)$, respectively $P(\cdot \mid C = \circ)$, induces a class of probabilistic classifiers. We now define different classes of distributions that in this way induce a hierarchy of different classifiers. We define our classes of distributions in terms of *log-linear models*, which are the most widely used classes of distributions on categorical state spaces such as $\mathbb{F}^n$ (see (Agresti, 2002) or (Schafer, 1997) for in-depth discussion of log-linear models). In our special context, where the state space is generated by binary variables only, the usual definitions of log-linear models can be given the following special form.

Definition 2.3 Let $\mathcal{I} \subseteq 2^{\{1,\ldots,n\}}$ be a collection of subsets of $\{1, \ldots, n\}$ with $\emptyset \in \mathcal{I}$. The *log-linear* model defined by $\mathcal{I}$ is the set of all distributions P on $\mathbb{F}^n$ that are given by

parameters $\lambda^I \in \mathbb{R}$ ($I \in \mathcal{I}$) in the form

$$log P(\boldsymbol{x}) = \sum_{I \in \mathcal{I}} \sigma^I_{\boldsymbol{x}} \lambda^I, \quad \text{where } \sigma^I_{\boldsymbol{x}} := (-1)^{|I| - |I \cap 1(\boldsymbol{x})|}. \tag{5}$$

We define

$$\mathcal{I}^{n,k} := \{I \subseteq \{1, \ldots, n\} \mid |I| \leq k\}, \quad \mathcal{I}^{n,k}_+ := \mathcal{I}^{n,k} \setminus \emptyset,$$

and call the log-linear model defined by $\mathcal{I}^{n,k}$ the *order-k association model*.

A probabilistic classifier is called an *order-k association classifier* if $P_\circ(\cdot)$ (for classical classifiers), respectively $P(\cdot \mid C = \circ)$ (for Bayesian classifiers) are distributions in the order-k association model ($\circ = \oplus, \ominus$).

For any $\mathcal{I}$, one can construct distributions in the log-linear model defined by $\mathcal{I}$ by choosing arbitrary parameters $\lambda^I \in \mathbb{R}$ for $I \neq \emptyset$, and then setting $\lambda^\emptyset$ so that the probabilities sum to one. For the order-k association model this means that we must have:

$$\begin{aligned} 1 &= \sum_{\boldsymbol{x} \in \mathbb{F}^n} P(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(\sum_{I \in \mathcal{I}^{n,k}} \sigma^I_{\boldsymbol{x}} \lambda^I) \\ &= \sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(\lambda^\emptyset + \sum_{I \in \mathcal{I}^{n,k}_+} \sigma^I_{\boldsymbol{x}} \lambda^I) \\ &= exp(\lambda^\emptyset) \sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(\sum_{I \in \mathcal{I}^{n,k}_+} \sigma^I_{\boldsymbol{x}} \lambda^I) \end{aligned}$$

so that

$$\lambda^\emptyset = -log(\sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(\sum_{I \in \mathcal{I}^{n,k}_+} \sigma^I_{\boldsymbol{x}} \lambda^I)) \tag{6}$$

Example 2.4 The order-1 association model contains just the distributions according to which the X_i are independent. The correspondence between the parameters $p_i := P(X_i = 1)$ in the direct formulation of an independence model, and the parameters λ^I in (5) is:

$$\lambda^{\{i\}} = log(\sqrt{p_i/(1-p_i)}), \quad \lambda^\emptyset = log(\sqrt{\prod_{i=1}^n p_i(1-p_i)}).$$

Friedman, Geiger and Goldszmidt (1997) have introduced the class of *tree-augmented naive Bayesian classifiers*. This class is determined by allowing for $P(\cdot \mid \circ)$ distributions that are representable with a Bayesian network in which every node has at most one parent (moreover, the Bayesian network has to be the same for $\circ = \oplus$ and $\circ = \ominus$). As a result, the conditional probability of any instance $\boldsymbol{x}$ given the class label is given as product of terms, each of which depends on at most two of the feature variables. From this it follows that all these distributions are in the order-2 association model, and tree augmented naive Bayesian classifiers are order-2 association classifiers. We will see in section 4 that the converse does not hold.

We are now ready to turn to our main objective, the investigation of the expressive power of different classes of classifiers: what classes of concepts can be recognized with order-k association classifiers? Our main result gives a complete *algebraic* characterization by showing the equivalence of these classifiers with polynomial discriminant functions.

Definition 2.5 A concept A is order-k polynomially separable, if there exist $a^H \in \mathbb{R}$ ($H \in \mathcal{I}^{n,k}_+$) and $b \in \mathbb{R}$, such that for all $\boldsymbol{x} \in \mathbb{F}^n$

$$\boldsymbol{x} \in A \quad \Leftrightarrow \quad \sum_{H \in \mathcal{I}^{n,k}_+} a^H \prod_{i \in H} x_i \leq b. \tag{7}$$

Order-1 polynomial separability then is just linear separability.

Theorem 2.6 Let $\emptyset \neq A \subset \mathbb{F}^n$, $k \in \{1, \ldots, n\}$. The following are equivalent

(i) A is recognized by an order-k association Bayesian classifier.

(ii) A is order-k polynomially separable.

(iii) A is recognized by an order-k association classical classifier.

The proof is given in appendix A. The restriction to nonempty A is due to the abovementioned limitation of the classical classifier. As $\emptyset$ is both linearly separable and recognizable by a naive Bayesian classifier, we obtain:

Corollary 2.7 A concept $A \subseteq \mathbb{F}^n$ is recognizable by a naive Bayesian classifier iff it is linearly separable.

3. Learning the Classifiers

The result in the previous section contrasts with negative results previously reported for the naive Bayesian classifier. In this section we investigate these discrepancies.

As mentioned in the introduction, when investigating the expressivity of classifiers, one has to watch carefully what the actual questions are that are being answered with positive or negative results. In the preceding section we have dealt with question *Q1*. This question is completely independent from the problem of how a classifier may be learned from data.

Given a set $(\boldsymbol{x}_1, \circ_1), \ldots, (\boldsymbol{x}_N, \circ_N)$ ($\boldsymbol{x}_i \in \mathbb{F}^n, \circ_i \in \{\oplus, \ominus\}$) of classified examples one will usually learn the classifier that maximizes the likelihood of the examples, i.e. for the classical classifier one chooses the $P_\circ$ that maximize

$$\prod_{i:\circ_i=\circ} P_\circ(\boldsymbol{x}_i) \quad (\circ \in \{\oplus, \ominus\}),$$

and for the Bayesian classifier one chooses P that maximizes

$$\prod_i P(\boldsymbol{x}_i, \circ_i).$$

Note that this describes a pure maximum likelihood approach (also for the Bayesian classifier), and that additional learning techniques and principles (like using Bayesian priors on parameters, or applying techniques to prevent overfitting) may be used.

A data set may be either *noise-free*, i.e. $\boldsymbol{x}_i = \boldsymbol{x}_j$ implies $\circ_i = \circ_j$, or *noisy*, i.e. it contains examples with conflicting labels. We say that a data set is *a description of the concept* A if it is noise-free, and contains for every $\boldsymbol{x} \in \mathbb{F}^n$ a data-item $(\boldsymbol{x}, \circ)$ with $\circ = \oplus$ iff $\boldsymbol{x} \in A$. We say that a data set *is an enumeration of* A if in addition it does not have any two identical data items.

We can now rephrase the negative result of (Kohavi & John, 1997) and (Domingos & Pazzani, 1997) as follows: there exist $m, n \in \mathbb{N}$, such that the naive Bayesian classifier learned by maximum likelihood inference from an enumeration of $A^{m,n}$ will not recognize $A^{m,n}$. This result need not be seen as a serious limitation of the naive Bayesian classifier, as there is little reason to focus on data sets that are enumerations of a certain concept. Indeed, when the training examples are real-life data, it is all but certain not to be an enumeration of a concept (and is likely to be noisy as well).

The naive classical classifier $(P_\oplus, P_\ominus)$ for the m-of-n concept described in example 2.2 will be learned by maximum likelihood inference from any data set in which the relative frequencies of $X_i = 1$ is m/n in the examples labelled with $\oplus$, and $(m-1)/n$ in the examples labelled with $\ominus$. The simplest data set that has this property is the one that contains m copies of the example $((1, \ldots, 1), \oplus)$, $m - n$ copies of $((0, \ldots, 0), \oplus)$, $m - 1$ copies of $((1, \ldots, 1), \ominus)$, and $n - m + 1$ copies of $((0, \ldots, 0), \ominus)$. This is a noisy data set. Note that we would not learn the correct classifier from the data set obtained by removing the mislabelled data items $((1, \ldots, 1), \ominus)$ and $((0, \ldots, 0), \oplus)$.

Zhang et al. (2000) present empirical results that show that some m-of-n concepts for which a naive Bayesian classifier is not learned from an enumeration of A^{m-n}, are learnable from a suitable description of A^{m-n}. It is an open problem whether for every linearly separable concept A there exists a noise-free data set, such that the naive Bayesian classifier learned from the data set will recognize A. The following theorem answers the converse question. To formulate it we call a concept A *general in the ith component* if there exist instances $\boldsymbol{x}, \boldsymbol{x}' \in A$ and $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{F}^n \setminus A$ with $x_i = y_i = 0$ and $x'_i = y'_i = 1$.

Theorem 3.1 Let A be a concept such that is general in at least one component. Then there exists a description D of A such that the naive classical classifier learned by maximum likelihood inference from D does not recoginize A.

Theorem 3.1 also holds for the naive Bayesian classifier (it is actually rather more obvious in the Bayesian case, as there one can skew the learning results by providing many more data items with class label $\oplus$ (or $\ominus$), thereby learning a high prior probability for $C = \oplus$, which leads to a classifier that will assign class label $\oplus$ to $\boldsymbol{x}$ whenever the conditional probabilities $P(X_i = x_i \mid \oplus)$ in the data are nonzero for $i = 1, \ldots, n$).

It is an open question to what extent theorem 3.1 also holds for order-k polynomially separable sets and order-k association classifiers for $k \geq 2$. It does seem to be a fairly safe conjecture that this negative result holds for all $k < n$ (but note that it does not hold for $k = n$, because the order-n association model comprises all probability distributions, and therefore from a description of A a classifier will be learned with $P_\oplus(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \notin A$, and $P_\ominus(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in A$). Things become much easier when one turns to noisy data sets: here it is clear that every order-k association classifier $(P_\oplus, P_\ominus)$ can be learned by maximum likelihood inference from a noisy data set that consists of representative samples from $P_\oplus$ and $P_\ominus$. Thus, if a concept A is order-k polynomially separable, and hence by theorem 2.6 recognized by a classical order-k association classifier $(P_\oplus, P_\ominus)$, then $(P_\oplus, P_\ominus)$ can be learned from a suitable noisy data set. There is one qualification, however: for $k \geq 2$ the parameters of the maximum likelihood order-k association classifier do not have a closed-form representation in terms of empirical counts, as is the case for $k = 1$. Thus, numerical techniques here have to be used for learning (see (Agresti, 2002, section 8.7) for some applicable techniques).

4. Logical and Geometric Characterizations

Theorem 2.6 provides an algebraic characterization of concepts recognizable by certain classes of classifiers. While this already provides a good deal of insight, and can be helpful in identifying the type of classifier needed to recognize a given concept, it is also true that the algebraic characterization does not always provide a good intuitive understanding of the expressive power of the various classifiers. More intuitive insights can be obtained by logical

and geometric characterizations.

To give logical characterizations, we interpret a concept $\mathbb{F}^n$ as the set of truth assignments to propositional variables $X_1, \ldots, X_n$. A concept A can then be identified with boolean formulas that are satisfied by the truth assignments in A. The goal then is to identify classes of concepts (those recognizable by certain classes of classifiers) with syntactically defined classes of formulas.

For geometric characterizations we interpret the elements of $\mathbb{F}^n$ as the vertices of the n-dimensional unit cube. Concepts A can then be classified according to certain connectedness properties of sets of vertices. Connections between syntactic and geometric properties of concepts (or Boolean functions) have been extensively studied (Ekin et al., 1999). Here we connect this existing work with properties of identifiability by order-k association classifiers. The results in this section are not uniform for order-k association classifiers for different k. We therefore first concentrate on the $k = 1$ case, and then discuss possible generalizations to $k > 1$. Beginning with logical characterizations, we recall the following standard definition.

Definition 4.1 Let ϕ be a boolean formula in $X_1, \ldots, X_n$ in negation normal form (NNF), i.e. ϕ is constructed from positive and negative literals using conjunction and disjunction only. ϕ is called *unate* iff for all i ϕ only contains the literal X_i, or only the literal $\neg X_i$. A concept A is called unate if it can be described with a unate formula.

Theorem 4.2 If A is linearly separable, then A is unate.

Proof: We have $\boldsymbol{x} \in A$ iff $\sum_i a^i x_i \leq b$ for some $a^i, b \in \mathbb{R}$. The intuition for the construction of a unate ϕ is now quite simple: we just axiomatize that whenever too many X_i with positive coefficients are true, then also sufficiently many X_i with negative coefficients are true.

For the details, let $I^+ := \{i \in \{1, \ldots, n\} \mid a_i > 0\}$ and $I^- := \{i \in \{1, \ldots, n\} \mid a_i < 0\}$. Define

$$\mathcal{J} := \{I \subseteq I^+ \mid \sum_{i \in I} a^i > b\},$$

and for every $I \in J$:

$$\mathcal{H}(I) := \{H \subseteq I^- \mid \sum_{i \in I \cup H} a_i \leq b\}.$$

Now consider the formula

$$\phi(A) := \bigwedge_{I \in \mathcal{J}} (\wedge_{i \in I} X_i \rightarrow \bigvee_{H \in \mathcal{H}(I)} \wedge_{h \in H} X_h) \qquad (8)$$

It is easy to see that $\phi(A)$ is unate (by eliminating the implications one obtains a NNF in which X_i with $a_i > 0$ only appear in negative literals, and X_i with $a_i < 0$ in positive literals). Now consider $\boldsymbol{x} \in \mathbb{F}^n$. Let $I^+_{\boldsymbol{x}} := 1(\boldsymbol{x}) \cap I^+$, and $I^-_{\boldsymbol{x}} := 1(\boldsymbol{x}) \cap I^-$. If $\boldsymbol{x} \in A$, then for every $I \in \mathcal{J}$ with $I \subseteq I^+_{\boldsymbol{x}}$ (i.e. those I for which $\boldsymbol{x}$ satisfies the left side of the implication in (8) we have that $I^-_{\boldsymbol{x}} \in \mathcal{H}(I)$, and therefore $\boldsymbol{x}$ satisfies the right hand side of the implication. If $\boldsymbol{x} \notin A$ then $\boldsymbol{x}$ does not satisfy the conjunct of $\phi(A)$ for $I = 1(\boldsymbol{x})$. □

The converse of theorem 4.2 does not hold:

Example 4.3 Let $\phi := (X_1 \wedge X_2) \vee (X_3 \wedge X_4)$. For the concept A defined by ϕ we have $(1,1,0,0), (0,0,1,1) \in A$, $(1,0,1,0), (0,1,0,1) \notin A$. If A was linearly separable with coefficients $a_1, \ldots, a_4, b$, then this would mean $a_1 + a_2 \leq b$, $a_3 + a_4 \leq b$, $a_1 + a_3 > b$, and $a_2 + a_4 > b$, an obvious contradiction.

For $\psi := (X_1 \wedge X_2) \vee (X_3)$, on the other hand, we obtain that the concept defined by ψ is linearly separable with coefficients $a_1 = a_2 = -1, a_3 = b = -2$.

The preceding example indicates that it may not be possible to find a natural logical if-and-only-if characterization of linear separability, as such a characterization would have to distinguish the syntactically very similar formulas ϕ and ψ. We now turn to geometric properties of linearly separable concepts. The following definitions are adopted from (Ekin et al., 1999).

Definition 4.4 Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{F}^n$. A *path* from $\boldsymbol{x}$ to $\boldsymbol{x}'$ is a sequence $\boldsymbol{x} = \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k = \boldsymbol{x}'$ in $\mathbb{F}^n$ such that $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$ have different values in at most one coordinate. A concept A is called *geodetically* connected, if for all $\boldsymbol{x}, \boldsymbol{x}' \in A$ there exists a shortest path connecting $\boldsymbol{x}$ and $\boldsymbol{x}'$ such that all instances on the path belong to A.

The intuition behind geodetic connectedness is that from any point in A we can reach any other point in A by taking a shortest walk along the edges of the unit cube, and never leave A on this walk. Ekin et al. (1999) show that when A is defined by a unate formula, then A and $\mathbb{F}^n \setminus A$ are geodetically connected (but the converse does not hold). It follows that for linearly separable A both A and $\mathbb{F}^n \setminus A$ are geodetically connected. The connectedness property we have thus found for linearly separable concepts has a very intuitive meaning: it says that between any two vertices of the unit cube that lie on the same side of some hyperplane, one can find a shortest path along the edges of the unit cube that connects the two vertices and always stays on the same side of the hyperplane.

Theorem 4.2 and its geometric implications do not seem to allow a meaningful generalization for concept classes determined by order-k polynomial separability for $k \geq 2$. First, there does not seem to exist a useful generalization

of the class of unate propositional formulas that, in analogy to theorem 4.2, would provide a nontrivial upper bound on the richness of the class of order-k separable concepts. As for the connectedness properties of linearly separable sets, one might entertain the thought that these generalize in the sense that it is possible to bound the number of connected components an order-k polynomially separable concept can have. As the following example shows, no bound that is polynomial in k can be given.

Example 4.5 Consider the order-2 polynomials of the form

$$p(\boldsymbol{x}) = \sum_{i=1}^{n} a x_i + \sum_{i,j=1}^{n} x_i x_j$$

with $a \in \mathbb{R}$. For $\boldsymbol{x}$ with $\mid 1(\boldsymbol{x}) \mid = k$ we obtain $p(\boldsymbol{x}) = ak + k(k-1)/2$. This is minimized for $k = 1/2 - a$ with a value of $(1/2 - a)^2/2$. To obtain a polynomial that has a unique minimum for some given $k_0 \in \{1, \ldots, n\}$, we can choose $a = 1/2 - k_0$. Then

$$\sum_{i=1}^{n} (1/2 - k_0) x_i - \sum_{i,j=1}^{n} x_i x_j \leq -k_0^2/2$$

holds exactly when $\mid 1(\boldsymbol{x}) \mid = k_0$. Thus, for every $k_0 \in \{1, \ldots, n\}$ the concept $A^{k_0} := \{\boldsymbol{x} \mid\mid 1(\boldsymbol{x}) \mid = k_0\}$ is order-2 polynomially separable.

By setting $k_0 = \lfloor n/2 \rfloor$ we obtain an example for a concept that consists of $\binom{n}{\lfloor n/2 \rfloor}$ distinct connected components.

One can show that when $k_0 < n$, then A^{k_0} cannot be represented by a tree-augmented Bayesian network. This is even true when instead of limiting to one the number of parents any node can have, one imposes a limit of some $l \leq n-2$. One thus sees that order-2 association classifiers are significantly more expressive than the ordinary tree augmented naive Bayesian classifiers. With the following example, on the other hand, we illustrate the limitations of order-2 classifiers.

Example 4.6 Let $n = 3$, and consider the concept $A = \{\boldsymbol{x} = (x_1, x_2, x_3) \mid \mid 1(\boldsymbol{x}) \mid \in \{0, 2\}\}$. Assume that A is separable by an order-2 polynomial

$$\sum_{i=1}^{3} a^i x_i + \sum_{i,j=1}^{3} a^{i,j} x_i x_j \leq b.$$

We obtain the following inequalities for the parameters, where (ii) and (iii) hold for all $i, j \in \{1, 2, 3\}, i \neq j$.

$$\begin{array}{ll} (i) & 0 \leq b \\ (ii) & a^i > b \\ (iii) & a^i + a^j + a^{i,j} \leq b \\ (iv) & a^1 + a^2 + a^3 + a^{1,2} + a^{1,3} + a^{2,3} > b \end{array}$$

From (ii) and (iii) it follows that $a^j + a^{i,j} < 0$. As (i) and (ii) imply that $a^j > 0$, this means that $a^{i,j} < 0$. From (iii) and (iv) we derive that $a^3 + a^{1,3} + a^{2,3} > 0$, which contradicts our previous results that both $a^3 + a^{1,3} < 0$ and $a^{2,3} < 0$.

5. Conclusions

We have introduced the hierarchy of order-k association classifiers, and shown that it corresponds exactly to the hierarchy of order-k polynomially separable concepts. The most interesting consequences are for the case $k = 1$, where this result implies that the concepts recognizable by a naive Bayesian classifier are exactly the linearly separable sets, thus, apparently for the first time, proving the converse of a well-known result from (Duda & Hart, 1973). For $k = 1$ we also derived some useful logical and geometric properties of linearly separable concepts.

There are a number of interesting open questions: is a naive Bayesian classifier for a linearly separable concept always learnable from a noise-free data set? Are there meaningful logical or geometric properties one can derive for order-k polynomially separable concepts?

The order-k association classifiers here were introduced with a mainly formal motivation. However, they might also be considered for use in practice: an order-k association model is determined by $O(n^k)$ independent parameters. The hierarchy of order-k association classifiers thus provides a smooth spectrum of models of increasing complexity, allowing a fine-tuned tradeoff between expressivity on the one hand, and efficiency of learning and guarding against overfitting on the other hand.

Acknowledgement

I thank Ursula Garczarek for the stimulating discussions about classification we had during her visit to Saarbrücken. A first version of theorem 2.6 was derived in collaboration with her at that time.

A. Proof of Main Theorem

Proof: The theorem clearly holds for $A = \mathbb{F}^n$. In the sequel we therefore assume $A \neq \emptyset$ and $A \neq \mathbb{F}^n$. We first rewrite the representation (5) of $logP(\boldsymbol{x})$ in polynomial form (throughout we use the convention that a product over an empty index set is equal to 1):

$$logP(\boldsymbol{x}) = \sum_{J \in \mathcal{I}^{n,n}} \prod_{i \in J} x_i \prod_{i \notin J} (1 - x_i) \sum_{I \in \mathcal{I}^{n,k}} (-1)^{\mid I \mid - \mid I \cap 1(\boldsymbol{x}) \mid} \lambda^I =$$

$$\sum_{I\in\mathcal{I}^{n,k}}\sum_{J\in\mathcal{I}^{n,n}}(-1)^{|I|-|I\cap 1(\boldsymbol{x})|}\lambda^I\prod_{i\in J}x_i\prod_{i\notin J}(1-x_i)=$$
$$\sum_{I\in\mathcal{I}^{n,k}}\sum_{J\subseteq I}(-1)^{|I|-|J|}\lambda^I\prod_{i\in J}x_i\prod_{i\in I\setminus J}(1-x_i)=$$

With

$$\prod_{i\in J}x_i\prod_{i\in I\setminus J}(1-x_i)=\sum_{H:J\subseteq H\subseteq I}(-1)^{|H\cap I\setminus J|}\prod_{i\in H}x_i$$

this becomes

$$\sum_{\substack{J,H,I\in\mathcal{I}^{n,k}\\ J\subseteq H\subseteq I}}(-1)^{|I|-|J|+|H\cap I\setminus J|}\lambda^I\prod_{i\in H}x_i=$$
$$\sum_{H\in\mathcal{I}^{n,k}}\Big(\sum_{I\in\mathcal{I}^{n,k}:H\subseteq I}(\nu(H,I))\lambda^I\Big)\prod_{i\in H}x_i=$$
$$\sum_{I\in\mathcal{I}^{n,k}}(-1)^{|I|}\lambda^I+\sum_{H\in\mathcal{I}_+^{n,k}}\Big(\sum_{I\in\mathcal{I}^{n,k}:H\subseteq I}(\nu(H,I))\lambda^I\Big)\prod_{i\in H}x_i$$

where

$$\begin{aligned}\nu(H,I)&:=\sum_{J:J\subseteq H}(-1)^{|I|-|J|-|H\cap I\setminus J|}\\&=\sum_{J:J\subseteq H}(-1)^{|I|+|H|-2|J|}\\&=\sum_{J:J\subseteq H}(-1)^{|I|+|H|}\\&=(-1)^{|I|+|H|}2^{|H|}\end{aligned}$$

(i)⇒(ii) Let A be recognized by an order-k log-linear Bayesian classifier given by class probabilities $P(C=\circ)=c_\circ$, conditional distribution $P(\boldsymbol{X}\mid C=\oplus)$ of the form (5) with parameters λ^I, and conditional distribution $P(\boldsymbol{X}\mid C=\ominus)$ of the form (5) with parameters κ^I.

Then $\boldsymbol{x}\in A$ iff

$$\mathit{logP}(C=\oplus\mid\boldsymbol{x})\geq\mathit{logP}(C=\ominus\mid\boldsymbol{x})\qquad(9)$$

⇔

$$\mathit{logP}(\boldsymbol{x}\mid C=1)-\mathit{logP}(\boldsymbol{x}\mid C=0)+\mathit{logc}_\oplus-\mathit{logc}_\ominus\leq 0$$

⇔

$$\sum_{I\in\mathcal{I}^{n,k}}(-1)^{|I|}(\lambda^I-\kappa^I)+\sum_{H\in\mathcal{I}_+^{n,k}}\Big(\sum_{I\in\mathcal{I}^{n,k}:H\subseteq I}\nu(H,I)(\lambda^I-\kappa^I)\Big)\prod_{i\in H}x_i+\mathit{logc}_\oplus-\mathit{logc}_\ominus\leq 0\qquad(10)$$

⇔

$$\sum_{H\in\mathcal{I}_+^{n,k}}\Big(\sum_{I\in\mathcal{I}^{n,k}:H\subseteq I}\nu(H,I)(\lambda^I-\kappa^I)\Big)\prod_{i\in H}x_i\leq\sum_{I\in\mathcal{I}^{n,k}}(-1)^{|I|}(\kappa^I-\lambda^I)+\mathit{logc}_\ominus-\mathit{logc}_\oplus.\qquad(11)$$

(ii)⇒(iii)

Assume that $\boldsymbol{x}\in A$ iff

$$p(\boldsymbol{x}):=\sum_{H\in\mathcal{I}_+^{n,k}}a^H\prod_{i\in H}x_i\leq b.\qquad(12)$$

Let $\boldsymbol{u}:=\mathit{arg\,min}_{\boldsymbol{x}\in\mathbb{F}^n}p(\boldsymbol{x})$, $\boldsymbol{w}:=\mathit{arg\,max}_{\boldsymbol{x}\in\mathbb{F}^n}p(\boldsymbol{x})$. With $A\neq\emptyset,\mathbb{F}^n$ we have $p(\boldsymbol{u})\leq b<p(\boldsymbol{w})$. As one can always replace b with $b+\epsilon$ for some small positive ϵ without changing the concept defined, we may furthermore assume that $p(\boldsymbol{u})<b<p(\boldsymbol{w})$.

A is recognized by an order-k classical classifier, iff it can be defined by a condition of the form (11) omitting the term $\mathit{logc}_\ominus-\mathit{logc}_\oplus$. We thus have to show that we can choose parameters λ^I,κ^I, such that

$$a^H=\sum_{I\in\mathcal{I}^{n,k}:H\subseteq I}\nu(H,I)(\lambda^I-\kappa^I)\quad(H\in\mathcal{I}_+^{n,k})\qquad(13)$$

and

$$b=\sum_{I\in\mathcal{I}^{n,k}}(-1)^{|I|}(\kappa^I-\lambda^I).\qquad(14)$$

□

To define the a^H, let $H_0,H_1,\ldots,H_m$ be an enumeration of $\mathcal{I}_+^{n,k}$ such that $i<j\rightarrow H_i\not\subseteq H_j$. We can now choose parameters $\lambda^{H_i},\kappa^{H_i}$ inductively so that

$$(\lambda^{H_i}-\kappa^{H_i})=a^{H_i}-\sum_{I\in\mathcal{I}^{n,k}:H_i\subsetneq I}\nu(H_i,I)(\lambda^I-\kappa^I).$$

Obviously, there is one degree of freedom in choosing the pairs λ^I,κ^I, and we can always replace λ^I,κ^I with $\lambda^I+c^I,\kappa^I+c^I$ for arbitrary $c^I\in\mathbb{R}$.

As $\lambda^\emptyset,\kappa^\emptyset$ are determined via (6) by the $\lambda^I,\kappa^I\in\mathcal{I}_+^{n,k}$, the right-hand side of (14) is already fully determined by our choices for the $\lambda^I,\kappa^I\in\mathcal{I}_+^{n,k}$. It remains to show, therefore, that using the degree of freedom in the choice of λ^I,κ^I, these parameters can be set so as to satisfy (14).

Using our representation of a^H in terms of the chosen

λ^I, κ^I, we obtain:

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{H \in \mathcal{I}_+^{n,k}: H \subseteq 1(\boldsymbol{x})} a^H \\ &= \sum_{H \in \mathcal{I}_+^{n,k}: H \subseteq 1(\boldsymbol{x})} \sum_{I \in \mathcal{I}^{n,k}: H \subseteq I} \nu(H, I)(\lambda^I - \kappa^I) \\ &= \sum_{I \in \mathcal{I}_+^{n,k}} (\lambda^I - \kappa^I) \sum_{H \in \mathcal{I}_+^{n,k}: H \subseteq 1(\boldsymbol{x}) \cap I} \nu(H, I). \end{aligned}$$

With

$$\begin{aligned} &\sum_{H \in \mathcal{I}_+^{n,k}: H \subseteq 1(\boldsymbol{x}) \cap I} \nu(H, I) \\ &\quad = \sum_{H \in \mathcal{I}_+^{n,k}: H \subseteq 1(\boldsymbol{x}) \cap I} (-1)^{|I|+|H|} 2^{|H|} \\ &\quad = \sum_{h=1}^{|1(\boldsymbol{x}) \cap I|} \binom{|1(\boldsymbol{x}) \cap I|}{h} (-1)^{|I|} (-2)^h \\ &\quad = (-1)^{|I|}((1-2)^{|1(\boldsymbol{x}) \cap I|} - 1) \\ &\quad = (-1)^{|I|+|1(\boldsymbol{x}) \cap I|} - (-1)^{|I|} \end{aligned}$$

and using $\sigma_{\boldsymbol{x}}^I = (-1)^{|I|-|1(\boldsymbol{x}) \cap I|} = (-1)^{|I|+|1(\boldsymbol{x}) \cap I|}$, we obtain

$$p(\boldsymbol{x}) = \sum_{I \in \mathcal{I}_+^{n,k}} (\sigma_{\boldsymbol{x}}^I - (-1)^{|I|})(\lambda^I - \kappa^I). \qquad (15)$$

Plugging this into (14) and solving for $(\kappa^\emptyset - \lambda^\emptyset)$ we obtain

$$\kappa^\emptyset - \lambda^\emptyset = b - p(\boldsymbol{x}) + \sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I (\lambda^I - \kappa^I). \qquad (16)$$

The problem of satisfying (14) now becomes that of choosing the λ^I, κ^I such that (16) holds. For any $c \in \mathbb{R}$ and $\boldsymbol{z} \in \mathbb{F}^n$ let

$$c_{\boldsymbol{z}}^I := \sigma_{\boldsymbol{z}}^I \cdot c. \qquad (17)$$

Substituting $\lambda^I + c_{\boldsymbol{z}}^I, \kappa^I + c_{\boldsymbol{z}}^I$ for λ^I, κ^I we obtain from (6)

$$\begin{aligned} &\kappa^\emptyset - \lambda^\emptyset \\ &\quad = log \frac{\sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I (\lambda^I + c_{\boldsymbol{z}}^I))}{\sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I (\kappa^I + c_{\boldsymbol{z}}^I))} \\ &\quad = log \frac{\sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(c \sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I \sigma_{\boldsymbol{z}}^I) exp(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I \lambda^I)}{\sum_{\boldsymbol{x} \in \mathbb{F}^n} exp(c \sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I \sigma_{\boldsymbol{z}}^I) exp(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I \kappa^I)} \end{aligned}$$

As $\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{x}}^I \sigma_{\boldsymbol{z}}^I$ is uniquely maximized for $\boldsymbol{x} \in \mathbb{F}^n$ by $\boldsymbol{x} = \boldsymbol{z}$, we obtain that for $c \to \infty$

$$\kappa^\emptyset - \lambda^\emptyset \longrightarrow log \frac{exp(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{z}}^I \lambda^I)}{exp(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{z}}^I \kappa^I)} \qquad (18)$$

$$= \sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{z}}^I (\lambda^I - \kappa^I) \qquad (19)$$

For the possible values of b, the right-hand side of (16) lies in the interval

$$(\sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{u}}^I (\lambda^I - \kappa^I), \sum_{I \in \mathcal{I}_+^{n,k}} \sigma_{\boldsymbol{w}}^I (\lambda^I - \kappa^I))$$

From the continuity of $\kappa^\emptyset - \lambda^\emptyset$ as a function of the c^I, and with (19) for $\boldsymbol{z} = \boldsymbol{u}$ and $\boldsymbol{z} = \boldsymbol{w}$ it follows that (16) is satisfiable for all possible values of b.

(iii)⇒(i) is immediate.

References

Agresti, A. (2002). *Categorical data analysis*. Wiley.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Duda, R. O., & Hart, E. (1973). *Pattern classification and scene analysis*. Wiley.

Ekin, O., Hammer, P. L., & Kogan, A. (1999). On connected Boolean functions. *Discrete Applied Mathematics*, *96-97*, 337–362.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Garg, A., & Roth, D. (2001). Understanding probabilistic classifiers. *Proceedings of ECML-01*. Springer.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*, 273–324.

Rish, I., Hellerstein, J., & Thathachar, J. (2001). *An analysis of data charasteristics that affect naive Bayes performance*Technical Report RC21993). IBM T.J.Watson Research Center.

Roth, D. (1998). Learning to resolve natural language ambiguities: a unified approach. *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence* (pp. 806–813).

Roth, D. (1999). Learning in natural language. *Proceedings of IJCAI-99*.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Zhang, H., Ling, C. X., & Zhao, Z. (2000). The learnability of naive bayes. *Canadian AI 2000* (pp. 432–441). Springer.