Pattern Recognition 82 (2018) 31–37

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Complexity of concept classes induced by discrete Markov networks and Bayesian networks

Benchong Li[*], Youlong Yang

School of Mathematics and Statistics, Xidian University, Xi'an 710126, PR China

ARTICLE INFO

Article history:
Received 1 January 2017
Revised 5 March 2018
Accepted 26 April 2018
Available online 27 April 2018

Keywords:
Bayesian networks
Classification
Markov networks
Toric ideal
Vapnik–Chervonenkis dimension

ABSTRACT

Markov networks and Bayesian networks are two popular models for classification. Vapnik–Chervonenkis dimension and Euclidean dimension are two measures of complexity of a class of functions, which can be used to measure classification capability of classifiers. One can use Vapnik–Chervonenkis dimension of the class of functions associated with a classifier to construct an estimate of its generalization error. In this paper, we study Vapnik–Chervonenkis dimension and Euclidean dimension of concept classes induced by discrete Markov networks and Bayesian networks. We show that these two dimensional values of the concept class induced by a discrete Markov network are identical, and the value equals dimension of the toric ideal corresponding to this Markov network as long as the toric ideal is nontrivial. Based on this result, one can compute the dimensional value in terms of a computer algebra system directly. Furthermore, for a general Bayesian network, we show that dimension of the corresponding toric ideal offers an upper bound of Euclidean dimension. In addition, we illustrate how to use Vapnik–Chervonenkis dimension to estimate generalization error in binary classification.

1. Introduction

Markov networks and Bayesian networks, also known as undirected and directed acyclic graphical models respectively, are widely used in many fields, such as machine learning and bioinformatics [8,12,22,24,25]. We know that a statistical model is a family of probability distributions, it is a (part of) real algebraic variety from the viewpoint of algebraic geometry [26]. In particular, consider discrete data, a statistical model is the set of all solutions of some polynomials in the probability simplex [15,27]. There are two ways to describe Markov networks and Bayesian networks, either by conditional independence statements or by a factorization of probability distributions. This corresponds to the computational algebraic geometry principle that some varieties can be described using either defining equations or parametric equations [see §3.3 of 9]. We know that these two representations are inequivalent for Markov networks without positivity assumptions on probability distributions [see Proposition 3.8 and Example 3.10 in 22], while these two ways are equivalent for Bayesian networks [see Theorem 3.27 in 22].

Classification is one of the main problems in machine learning. To improve classification, there has been a growing body of work on Markov networks in the computer vision community [21], in social and affiliation networks [39] and in classifier learning [30]. Ghofrani et al. proposed a new probabilistic classifier based on decomposable models and applied it to two real-world internet traffic data sets [16]. Bayesian networks are another powerful tool for classification due to their simplicity and accuracy [4,6,11,13]. Naive Bayes classifiers, a special case of Bayesian network classifiers, are a popular classification tool for processing discrete data [4,34,35]. For more about discrete Bayesian network classifiers, please see a comprehensive survey published recently [3]. Several research groups combined kernel method and probabilistic models, for instance, Ben-David et al. [2], Altun et al. [1], Taskar et al. [31], Chechik et al. [5]. A major advantage of this technique is that one can construct a wide variety of more flexible classifiers.

The generalization performance of a learning method is extremely important in practice, because it provides a measure of the quality of the ultimately chosen model. One can use the Vapnik–Chervonenkis (VC) dimension to construct an estimate of generalization error [32], where the VC dimension is an approach of measuring the complexity of a class of functions by assessing how wiggly its members can be [19]. Another measure of complexity of a class of functions, called Euclidean dimension, is the minimum dimension of the Euclidean space equipped with the standard dot product, into which these functions can be embedded. Given a Markov network (Bayesian network) $\mathcal{N}$, one can get a concept class $\mathcal{C}_{\mathcal{N}}$ induced by it. Since VC dimension and Euclidean dimension

[*] Corresponding author.
E-mail addresses: libc580@nenu.edu.cn (B. Li), ylyang@mail.xidian.edu.cn (Y. Yang).

https://doi.org/10.1016/j.patcog.2018.04.026
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

are two important indexes to assess the classification ability of a Markov network (Bayesian network). Three natural questions arise: (1) Whether the two dimensional values can be obtained from the dimension of a model as a (smooth) sub-manifold of Euclidean space? (2) Are the two positive integer numbers always equal? (3) How to calculate these dimensional values? In this paper, we study the classification capability induced by a Markov network (Bayesian network) without considering the training data, the algorithm and the construction of the network graph. We aim to solve the three questions above.

Kearns and Schapire [20] studied a formal model of general concept learning, they focused on learnability and uniform convergence of probabilistic concept classes, and provided many efficient algorithms. Using tools in algebraic geometry, Geiger et al. [17] formulated necessary and sufficient conditions for an arbitrary discrete distribution to factor according to a general exponential model. Garcia et al. [14] studied the algebraic geometry of discrete Bayesian networks. Nakamura et al. [23] established the upper and lower bounds on Euclidean dimension for Bayesian networks on binary random variables, and determined the exact values of Euclidean dimension for some special classes of Bayesian networks. For Bayesian networks with k-valued vertices, Yang and Wu [37] studied fully connected Bayesian networks and Bayesian networks without V-structures. For each Bayesian network belongs to these two cases, they showed that the two dimensional values are identical and offered an explicit formula to compute this dimension. Varando et al. [33] used linear combinations of products of Lagrange basis polynomials to evaluate the expressive power of Bayesian network classifiers. In absence of V-structures, they showed that the polynomial representation provides complete characterization for the type of Bayesian network classifier.

Motivated by the work of Geiger et al. [17] and Garcia et al. [14], in this paper, for a general (non-complete) discrete Markov network, we show that the two dimensional values are identical and they equal dimension of the toric ideal corresponding to this Markov network. Since Bayesian networks without V-structures can be viewed as chordal graphs, we provide two new characterizations of partial results of Yang and Wu [37] and Varando et al. [33]. Moreover, we present upper bounds of Euclidean dimension (and thus upper bounds of VC dimension) for Bayesian network classifiers.

In Section 2, we introduce basic definitions and notations. Our main result is presented in Section 3. In Section 4, we show that VC dimension, Euclidean dimension and dimension of the toric ideal corresponding to a general discrete Markov network are the same number, and present upper bounds of Euclidean dimension for Bayesian network classifiers. We demonstrate the application of our results in estimating generalization error of binary classification. Section 6 is devoted to discussion.

2. Preliminaries

In this section, we give formal definitions for concepts appear in this article. Discrete Markov networks and Bayesian networks and related notations are introduced in Section 2.1 and Section 2.2, respectively. VC dimension and Euclidean dimension of a family of functions which follow notations in Nakamura et al. [23] are presented in Section 2.3. Basic concepts and results in algebraic geometry are reviewed in Section 2.4. Algebraic geometry characterization of discrete Markov networks is provided in Section 2.5. Algebraic geometry of Bayesian networks is summarized in Section 2.6.

2.1. Discrete Markov networks

Let $\overline{G} = (V, \overline{E})$ be an undirected graph (UG), where V is the vertex set and E is the set of undirected edges. We consider simple UGs. A Markov network $\mathcal{N} = (\overline{G}, \mathcal{P})$ consists of two components, where $\overline{G}$ is a UG and $\mathcal{P}$ is a class of probability distributions. Given a UG $\overline{G}$, a clique is a maximal complete subgraph of $\overline{G}$, and we use $\kappa_{\overline{G}}$ to denote the set of all cliques of $\overline{G}$. $\overline{G}$ is said to be a chordal graph if and only if every cycle of length 4 or more has a chord.

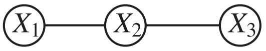

Fig. 1. A UG $\overline{G}_1$.

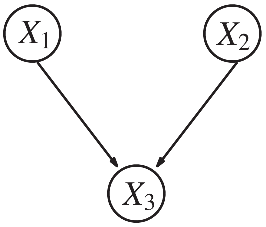

Fig. 2. A DAG $\overrightarrow{G}_2$.

Let $X_1, \ldots, X_n$ be discrete random variables where X_i take values in the finite set $[d_i] = \{1, 2, \ldots, d_i\}$, and $\mathcal{X} = [d_1] \times [d_2] \times \cdots [d_n]$, where $d_i \geq 2$. An undirected graphical model $\mathcal{P}$ is defined as the collection of distributions on $\mathcal{X}$ of the form

$$p_{x_1 \cdots x_n} \doteq P(x_1, \ldots, x_n) = P(x) \propto \prod_{\mathcal{K} \in \kappa_G} \psi_{\mathcal{K}}(x), \qquad (1)$$

where $x_i \in [d_i]$, $x = (x_1, \ldots, x_n)$ and $\psi_{\mathcal{K}}(x)$ is a potential function that depends on x only through the values of variables in $\mathcal{K}$. We focus on the collection of positive probability distributions in $\mathcal{P}$, denoted as $\mathcal{P}^+$.

Example 2.1. Note that $\overline{G}_1$ in Fig. 1 is a chordal graph. Suppose $X_i \in \{0,\ 1\}$, for $i = 1, 3$, and $X_2 \in \{0,\ 1,\ 2\}$. $\forall P \in \mathcal{N}_1 = (\overline{G}_1, \mathcal{P}_1)$, $p_{021} \propto \psi_{\{X_1,X_2\}}(0,2)\psi_{\{X_2,X_3\}}(2,1)$.

2.2. Discrete Bayesian networks

A directed acyclic graph (DAG) $\overrightarrow{G} = (V, \overrightarrow{E})$ is a directed graph with no directed cycles. Each vertex $X_i \in V$ represents a random variable and a directed edge $(X_i, X_j) \in \overrightarrow{E}$ represents the conditional dependence between X_i and X_j, where $V = \{X_1, \ldots, X_n\}$, $i, j \in \{1, \ldots, n\}$, and $i \neq j$. If $(X_i, X_j) \in \overrightarrow{E}$, X_i is called a parent of X_j. We use PA_i to denote the set of parents of the vertex X_i.

The class of probability distributions $\mathcal{P}$ induced by a DAG $\overrightarrow{G}$ consists of all distributions on $\mathcal{X}$ of the form

$$p_{x_1 \cdots x_n} \doteq P(x_1, \ldots, x_n) = P(x) = \prod_{i=1}^{n} P_i(x_i | x_{pa(i)}), \qquad (2)$$

where $x_i \in [d_i]$, $x_{pa(i)} \in \prod_{X_j \in PA_i} [d_j]$, and $P_i(x_i | x_{pa(i)})$ is the conditional probability of $X_i = x_i$ given $PA_i = x_{pa(i)}$. Given a DAG $\overrightarrow{G}$, $\mathcal{N} = (\overrightarrow{G}, \mathcal{P})$ is called a Bayesian network. As the case of a Markov network, we consider $\mathcal{P}^+$ here.

Example 2.2. Consider the DAG $\overrightarrow{G}_2$ in Fig. 2, we assume all variables are binary, that is, $X_i \in \{0,\ 1\}$. For $\overrightarrow{G}_2$, $PA_3 = \{X_1, X_2\}$ and $\forall P \in \mathcal{N}_2 = (\overrightarrow{G}_2, \mathcal{P}_2)$, $p_{011} = P_1(0)P_2(1)P_3(1|0,1)$.

2.3. Concept classes, VC dimension, and Euclidean dimension

A concept class $\mathcal{C}$ over domain $\mathcal{X}$ is a family of functions of the form $f : \mathcal{X} \to \{1, -1\}$. Each $f \in \mathcal{C}$ is called a concept, that is, a concept over a domain $\mathcal{X}$ is a Boolean function over $\mathcal{X}$. A finite set $S = \{s_1, s_2, \ldots, s_m \subseteq \mathcal{X}\}$ is said to be shattered by $\mathcal{C}$ if for every m-dimensional binary vector $b \in \{1, -1\}^m$, there exists some concept

$f \in \mathcal{C}$ such that $f(s_i) = b_i$ for $i = 1, 2, \ldots, m$. The VC dimension of $\mathcal{C}$ is given by

$$\mathrm{VCdim}(\mathcal{C}) = sup\{m| \text{ there is some } S \subseteq \mathcal{X} \text{ shattered by } \mathcal{C} \text{ and } |S| = m\}.$$

We use the sign function for mapping a real-valued function g to a ± 1-valued concept sign$\circ$g, where $\forall\ c \in \mathbb{R}$, $\mathrm{sign}(c) = 1$ if $c \geq 0$; otherwise, $\mathrm{sign}(c) = -1$.

Given a concept class $\mathcal{C}$ over domain $\mathcal{X}$ and a r-dimensional Euclidean space equipped with standard inner product, if there exist r-dimensional collections of vectors $(u_f)_{f\in\mathcal{C}}$, $(v_x)_{x\in\mathcal{X}}$ in $\mathbb{R}^r$ such that

$$\forall f \in \mathcal{C}, \forall x \in \mathcal{X}, f(x) = \mathrm{sign}(u_f^T v_x) \text{ holds,}$$

we say that the concept class $\mathcal{C}$ can be embedded into a r-dimensional Euclidean space, where u_f^T denotes the transpose of u_f. The smallest r such that $\mathcal{C}$ can be embedded into $\mathbb{R}^r$ is called the Euclidean dimension of $\mathcal{C}$, denoted as $\mathrm{Edim}(\mathcal{C})$. For a concept class $\mathcal{C}$, if there is no finite-dimensional Euclidean space into which $\mathcal{C}$ can be embedded, we say $\mathrm{Edim}(\mathcal{C})$ is infinite. For finite concept classes, it is easy to see that $\mathrm{Edim}(\mathcal{C}) \leq \min\{|\mathcal{C}|, |\mathcal{X}|\}$.

The concept class induced by the network $\mathcal{N} = (G, \mathcal{P}^+)$ is the collection of all ± 1-valued functions on $\mathcal{X}$ of the form $\mathrm{sign}(log(P(x)/Q(x)))$ for $P, Q \in \mathcal{P}^+$, and we use $\mathcal{C}_{\mathcal{N}}$ to denote this set. Note that, $\mathrm{sign}(log(P(x)/Q(x))) = 1$ if $P(x) \geq Q(x)$ and $\mathrm{sign}(log(P(x)/Q(x))) = -1$ otherwise.

2.4. Basic concepts and results in algebraic geometry

We review some basic concepts from algebraic geometry, one can see Cox et al. [9] for more details. All algebraic geometry terminology we use which is not defined in this paper can be found in [9].

We work in the polynomial ring $\mathbb{R}[y] = \mathbb{R}[y_1, y_2, \ldots, y_d]$. A subset $I \subseteq \mathbb{R}[y]$ is an ideal if it satisfies: (1) $0 \in I$, (2) if $f, g \in I$, then $f + g \in I$, and (3) if $f \in I$ and $h \in \mathbb{R}[y]$, then $hf \in I$. Given an ideal $I \subseteq \mathbb{R}[y]$, we can define a set

$$V(I) = \{y \in \mathbb{R}^n : f(y) = 0 \text{ for all } f \in I\},$$

called real variety of I. Hilbert's basis theorem shows that every $I \subseteq \mathbb{R}[y]$ contains a finite set $F = \{f_1, \ldots, f_m\}$, called an ideal basis of I, such that every $g \in I$ can be expressed as $g(y) = \sum_{i=1}^m h_i(y) f_i(y)$, where $\{h_1, \ldots, h_m\} \subset \mathbb{R}[y]$. The ideal generated by F is denoted by $\langle f_1, \ldots, f_m\rangle$.

In this paper, we consider ideals generated by binomials, namely, a set of polynomials each has precisely two terms. We use $ker_{\mathbb{Z}}(A)$ to denote the integer kernel of A. The toric ideal I_A associated with a $q \times d$ integer matrix A is generated by the binomials $y_1^{u_1} \cdots y_d^{u_d} - y_1^{v_1} \cdots y_d^{v_d}$, where $u = (u_1, \ldots, u_d)$, $v = (v_1, \ldots, v_d) \in \mathbb{N}^d$, and $u - v \in ker_{\mathbb{Z}}(A)$. A toric variety is a variety that corresponds to a toric ideal. One can see Sturmfels [29] for more about toric ideals.

2.5. Algebraic geometry of discrete Markov networks

We identify $\mathcal{X}$ with the set $\{1, 2, \ldots, d\}$ and define a probability distribution to be a vector $P = (p_1, p_2, \ldots, p_d) \in \mathbb{R}^d_{\geq 0}$ such that $p_1 + p_2 + \cdots + p_d = 1$, where $d = \prod_{X_i \in V} d_i$. For a UG $\overline{\mathrm{G}}$, by (1), it is natural to view the corresponding graphical model $\mathcal{P}$ as the image of the monomial mapping $\phi_{A(\overline{\mathrm{G}})}$ [17]:

$$\phi_{A(\overline{\mathrm{G}})} : \mathbb{R}^q_{\geq 0} \to \mathbb{R}^d_{\geq 0}, \quad (t_1, \ldots, t_q) \mapsto \left(\prod_j t_j^{a_{j1}}, \prod_j t_j^{a_{j2}}, \ldots, \prod_j t_j^{a_{jd}}\right),$$

where, $A(\overline{\mathrm{G}}) = (a_{jk})$ is a $q \times d$ matrix of nonnegative integers, $q = \Sigma_{\mathcal{K}\in\kappa_G} \prod_{X_l \in \mathcal{K}} d_l$, and $t^0 = 1$ for $t \geq 0$.

For a UG $\overline{\mathrm{G}}$, the columns of $A(\overline{\mathrm{G}})$ are indexed by $\prod_{X_i \in V}[d_i]$, and the rows of $A(\overline{\mathrm{G}})$ are indexed by pairs consisting of a clique $\mathcal{K}$ and an element of $\prod_{X_l \in \mathcal{K}} d_l$. Note that each entry of $A(\overline{\mathrm{G}})$ is 0 or 1, the entry is 1 if and only if the element in the row index is equal to the projection of the column index to the corresponding variables in the row index, and the number of1′s in each column of $A(\overline{\mathrm{G}})$ is the constant $|\kappa_{\overline{G}}|$.

2.6. Algebraic geometry of discrete Bayesian networks

For a DAG $\vec{G}$, the rows of $A(\vec{G})$ are indexed by pairs consisting of a variable X_i and an element of $\prod_{X_l \in X_i \cup PA_i}[d_l]$ (a conditional probability), $q = \Sigma_{X_i \in V} \prod_{X_l \in X_i \cup PA_i} d_l$, and each column sum of $A(\vec{G})$ is n.

Example 2.3. (Example 2.2 continue). For $\mathcal{N}_2$, $q = 12$. Then $A(\vec{G}_2) =$

$$\begin{array}{l} \\ t_1 = P_1(0) \\ t_2 = P_1(1) \\ t_3 = P_2(0) \\ t_4 = P_2(1) \\ t_5 = P_3(0|0,0) \\ t_6 = P_3(1|0,0) \\ t_7 = P_3(0|0,1) \\ t_8 = P_3(1|0,1) \\ t_9 = P_3(0|1,0) \\ t_{10} = P_3(1|1,0) \\ t_{11} = P_3(0|1,1) \\ t_{12} = P_3(1|1,1) \end{array}
\begin{array}{c} \begin{array}{cccccccc} p_{000} & p_{001} & p_{010} & p_{011} & p_{100} & p_{101} & p_{110} & p_{111} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

A probability distribution $P = (p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111})$ factors according to $\vec{G}_2$ only if it lies in the image of the associated monomial mapping

$$\phi_{A(\vec{G}_2)} : \mathbb{R}^{12}_{\geq 0} \to \mathbb{R}^8_{\geq 0}, \quad (t_1, \ldots, t_{12}) \mapsto (t_1t_3t_5, t_1t_3t_6, t_1t_4t_7, t_1t_4t_8, t_2t_3t_9, t_2t_3t_{10}, t_2t_4t_{11}, t_2t_4t_{12}).$$

For a Bayesian network $\mathcal{N} = (\vec{G}, \mathcal{P})$, the toric ideal $I_{A(\vec{G})}$ we defined here along the idea of Geiger et al. [17] is different from the prime ideal $\ker(\Phi)$ given in Garcia et al. [14], where Φ is the ring homomorphism $\mathbb{R}[y_1, y_2, \ldots, y_d] \to \mathbb{R}[t_1, t_2, \ldots, t_q]$ combined with natural constraints on $t_1, t_2, \ldots, t_q$, that is, the sum of conditional probabilities under any condition is 1. For example, $\mathcal{N}_2$ in Example 2.3, Garcia et al. [14] view t_i as $1 - t_{i+1}$ for $i = 1, 3, 5, 7, 9, 11$. It is obvious that $I_{A(\vec{G})} \subseteq \ker(\Phi)$.

Example 2.4. (Examples 2.1 and 2.2 continue). The results below are computed using the computer algebra system Singular [10].

$I_{A(\overline{G}_1)} = \langle y_1 - t_1t_7, y_2 - t_1t_8, y_3 - t_2t_9, y_4 - t_2t_{10}, y_5 - t_3t_{11}, y_6 - t_3t_{12}, y_7 - t_4t_7, y_8 - t_4t_8, y_9 - t_5t_9, y_{10} - t_5t_{10}, y_{11} - t_6t_{11}, y_{12} - t_6t_{12}\rangle \cap \mathbb{R}[y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}, y_{12}] = \langle y_1y_8 - y_2y_7, y_3y_{10} - y_4y_9, y_5y_{12} - y_6y_{11}\rangle$.

$I_{A(\vec{G}_2)} = \langle y_1 - t_1t_3t_5, y_2 - t_1t_3t_6, y_3 - t_1t_4t_7, y_4 - t_1t_4t_8, y_5 - t_2t_3t_9, y_6 - t_2t_3t_{10}, y_7 - t_2t_4t_{11}, y_8 - t_2t_4t_{12}\rangle \cap \mathbb{R}[y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8] = \{0\}$.

Remark 2.1. For us, the objective of interest is the complexity of concept classes induced by discrete Markov networks and Bayesian networks. So it is natural to explore the connections between concepts in Section 2.3 and algebraic geometry of two kinds of graphical models introduced in the last two subsections. We will pursue this theme in Section 3 and Section 4.

3. Main result

We introduce a definition of dimension of $I_{A(G)}$ and then present the main result in this paper.

Definition 3.1. Let $I^S_{A(G)}$ be the elimination ideal obtained from $I_{A(G)}$ by eliminating all unknowns in $\{y_1, \dots, y_d\}$ corresponding to elements in $\mathcal{X} \setminus S$. The biggest m such that $I^S_{A(G)}$ is the zero ideal is called the algebraic elimination dimension of $I_{A(G)}$, where $m = |S|$.

We denote the dimension of $I_{A(G)}$ by $\dim(I_{A(G)})$. In algebraic geometry, there are many ways to reformulate the concept of dimension of an ideal (or equivalently, dimension of a affine variety), and Definition 3.1 is one of the interesting approaches [9]. If $I^S_{A(G)}$ is the zero ideal, then its variety is full-dimensional. Lemma 4.2 in Sturmfels [29] states that $\dim(I_{A(G)}) = \text{rank}(A(G))$. The following theorem is our main result in the paper.

Theorem 3.1. *For every Markov network $\mathcal{N} = (\overline{G}, \mathcal{P})$ with $I_{A(\overline{G})} \neq \{0\}$, we have $dim(I_{A(\overline{G})}) = VCdim(\mathcal{C}_{\mathcal{N}}) = Edim(\mathcal{C}_{\mathcal{N}}) = rank(A(\overline{G}))$.*

Corollary 4.1 and Lemma 4.2 in Section 4 demonstrate that Theorem 3.1 holds. A special case of Theorem 3.1 is that $\overline{G}$ is a chordal graph. From the point of view representing conditional independence relationships among variables, each chordal graph can be viewed as a DAG [see Lemma 7.3 in 28], and every DAG without V-structures is identical to a chordal graph. The corollary below provides two new characterizations of the results of Yang and Wu [37] and Varando et al. [33].

Corollary 3.1. *Given a Bayesian network $\mathcal{N} = (\overrightarrow{G}, \mathcal{P})$ with $I_{A(\overrightarrow{G})} \neq \{0\}$ and $\overrightarrow{G}$ has no V-structure, we have $dim(I_{A(\overrightarrow{G})}) = VCdim(\mathcal{C}_{\mathcal{N}}) = Edim(\mathcal{C}_{\mathcal{N}}) = rank(A(\overrightarrow{G}))$.*

As an illustration, consider the following example.

Example 3.1. (Example 2.1 continue). $\overline{G}_1$ in Fig. 1 is a UG (a chordal graph), and $\text{rank}(A(\overline{G}_1)) = \text{VCdim}(\mathcal{C}_{\mathcal{N}_1}) = \text{Edim}(\mathcal{C}_{\mathcal{N}_1}) = \dim(I_{A(\overline{G}_1)}) = 9$.

Theorem 3.1 in Yang and Wu [37] shows that for a fully connected Bayesian network with each variable has k levels, $\text{VCdim}(\mathcal{C}_{\mathcal{N}}) = \text{Edim}(\mathcal{C}_{\mathcal{N}}) = k^n - 1$. If $I_{A(\overrightarrow{G})} = \{0\}$ and $\overrightarrow{G}$ has no V-structure, that is, $\overrightarrow{G}$ is a fully connected Bayesian network, then $\overrightarrow{G}$ can be viewed as a complete UG, we have $\text{rank}(A(\overrightarrow{G})) = \dim(I_{A(\overrightarrow{G})}) = d$, and $\text{VCdim}(\mathcal{C}_{\mathcal{N}}) = \text{Edim}(\mathcal{C}_{\mathcal{N}}) = d - 1$.

4. Proof of Theorem 3.1 and related results

In this section, we prove Theorem 3.1 and provide some related results.

4.1. Bounding VC dimension by dimension of the toric ideal corresponding to a discrete Markov network

In this subsection, we show that $\dim(I_{A(G)})$ yields a lower bound for $\text{VCdim}(\mathcal{C}_{\mathcal{N}})$ of discrete Markov networks.

For each unknown that corresponds to an element in S, we set it equal to an arbitrary positive real number, then we get a partial solution. What is interest is whether the partial solution can be extended to a complete solution. Based on the fact that each row of $A(G)$ which corresponds to an element in $\mathcal{X} \setminus S$ can be represented linearly by rows of $A(G)$ and the representation is unique, one can see that for a toric ideal corresponding to a discrete Markov (Bayesian) networks the answer to the question of real extension of a partial solution is yes in some situations.

Proposition 4.1. *Suppose $I_{A(G)} \neq \{0\}$ and $S \subset \mathcal{X}$ is one of the subsets with the biggest cardinality such that $I^S_{A(G)} = \{0\}$. Let each unknown corresponding to an element in S be an arbitrary positive real number, then this partial solution can be extended to exactly one complete solution.*

Fig. 3. A UG $\overline{G}_3$.

Note that the complete solution in Proposition 4.1 corresponds to a strictly positive probability distribution, that is, dividing each component of the complete solution by the sum of all components, which is called a normalizing constant. For discrete Markov networks, the well-known Hammersley–Clifford theorem states that a strictly positive probability distribution $(y_1, y_2, \dots, y_d) \in \mathcal{P}^+$ if and only if $(y_1, y_2, \dots, y_d)$ is a solution of all quadratic binomials corresponding to all the saturated conditional independence statements (a finite subset of $I_{A(\overline{G})}$) [17]. If we fix $m-1$ unknowns in $\{y_{(1)}, y_{(2)}, \dots, y_{(m)}\}$, that is, $y_{(j)} = y_{(j)0} > 0$ for $j \in \{1, 2, \dots, m\} \setminus \{i\}$, Proposition 4.1 shows that the normalizing constant $(\sum_{j=1, j\neq i}^{m} y_{(j)0} + y_{(i)} + \sum_{j=m+1}^{d} y_{(j)})$ is a function of $y_{(i)}$, where $i \in \{1, 2, \dots, m\}$.

Example 4.1. Consider $\overline{G}_3$ in Fig. 3, $I_{A(\overline{G}_3)} = \langle y_1y_4 - y_2y_3 \rangle$. Suppose $y_2 = 3$, $y_3 = 2$ and $y_4 = 1$, the partial solution (3,2,1) can be extended to the complete solution (6,3,2,1), and then $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{12}\right) \in \mathcal{P}_1^+$.

For a Markov network $\mathcal{N} = (\overline{G}, \mathcal{P})$ with $I_{A(\overline{G})} \neq \{0\}$, the Hammersley–Clifford theorem implies that each $y_{(j)}$ is either a linear function or a hyperbolic function of $y_{(i)}$ (that is, be of the form $y_{(j)} = ay_{(i)}$ or $y_{(j)} = \frac{b}{y_{(i)}}$, where $a > 0$, $b > 0$), then the normalizing constant function is either the form $g_i(y_{(i)}) = k_0 + k_1y_{(i)}$ or $g_i(y_{(i)}) = k_0 + k_1y_{(i)} + \frac{k_2}{y_{(i)}}$, where $j \in \{m+1, \dots, d\}$, $i \in \{1, 2, \dots, m\}$, $k_l > 0$ for $l = 0, 1, 2$. A basic observation is that there exists at least one $y_{(i)} \in \{y_{(1)}, y_{(2)}, \dots, y_{(m)}\}$ such that the normalizing constant has the form $g_i(y_{(i)}) = k_0 + k_1y_{(i)} + \frac{k_2}{y_{(i)}}$ when each unknown in $\{y_{(1)}, y_{(2)}, \dots, y_{(m)}\} \setminus \{y_{(i)}\}$ is specified a positive number.

Lemma 4.1. *Given a Markov network $\mathcal{N} = (\overline{G}, \mathcal{P})$ with $I_{A(\overline{G})} \neq \{0\}$, we have $dim(I_{A(\overline{G})}) \leq VCdim(\mathcal{C}_{\mathcal{N}})$.*

Proof. Suppose $\dim(I_{A(\overline{G})}) = m$ and $S = \{s_1, s_2, \dots, s_m\} \subset \mathcal{X}$ satisfies $I^S_{A(\overline{G})} = \{0\}$, we need to show that $\forall b \in \{1, -1\}^m$, there exist two positive distributions P, Q such that $\text{sign}(\frac{P(s_i)}{Q(s_i)}) = b_i$ for $i = 1, 2, \dots, m$. It suffices to consider the following two cases.

Case 1, $b = \pm(1, 1, \dots, 1)$. Without loss of generality, we assume the normalizing constant is a function of $y_{(1)}$ with the form $g_1(y_{(1)}) = k_0 + k_1y_{(1)} + \frac{k_2}{y_{(1)}}$ when $y_{(2)0}, \dots, y_{(m)0}$ are given. Note that k_0, k_1, k_2 are determined by $y_{(2)0}, \dots, y_{(m)0}$, and one can choose a $y_{(1)0} \in (0, \sqrt{\frac{k_2}{k_1}})$ and a small $\epsilon > 0$ such that $y_{(1)0} + \epsilon < \sqrt{\frac{k_2}{k_1}}$. Let P, Q be two distributions corresponding to $y_{(1)0}, y_{(2)0}, \dots, y_{(m)0}$ and $y_{(1)0} + \epsilon, y_{(2)0}, \dots, y_{(m)0}$, respectively. We know that each function $y = k_0 + k_1x + \frac{k_2}{x}$ is strictly decreasing on $(0, \sqrt{\frac{k_2}{k_1}})$, and strictly increasing on $(\sqrt{\frac{k_2}{k_1}}, +\infty)$, where $k_0 > 0$, $k_1 > 0$, $k_2 > 0$, $x \in (0, +\infty)$. Then, $g_1(y_{(1)0}) > g_1(y_{(1)0} + \epsilon) > 0$, we have $\text{sign}(\frac{P(s_i)}{Q(s_i)}) = -1$ and $\text{sign}(\frac{Q(s_i)}{P(s_i)}) = 1$ for $i = 1, 2, \dots, m$.

Case 2, $b = (1, 1, \dots, 1, -1, \dots, -1)$, that is, if $i \in \{1, 2, \dots, m_1\}$, $b_i = 1$; else $b_i = -1$, where $m_1 < m$ is a positive integer. Let P, Q be two distributions corresponding to $y_{(1)0}, y_{(2)0}, \dots, y_{(m_1)0}, y_{(m_1+1)0}, \dots, y_{(m)0}$ and

$y_{(1)0}, y_{(2)0}, \dots, y_{(m_1)0}$, $y_{(m_1+1)0} + \epsilon_{m_1+1}, \dots, y_{(m)0} + \epsilon_m$, respectively, where $\{\epsilon_{m_1+1}, \dots, \epsilon_m\}$ are $m - m_1$ small positive numbers, $y_{(1)0}$ is a large number belonging to the strictly increasing interval of the normalizing constant function of $(y_{(1)}, y_{(2)0}, \dots, y_{(m_1)0}, y_{(m_1+1)0} + \epsilon_{m_1+1}, \dots, y_{(m)0} + \epsilon_m, y_{(m+1)}, \dots, y_{(d)})$, that is, a function of $y_{(1)}$ denoted by $g_1^Q(y_{(1)})$.

If $g_1^P(y_{(1)0}) = g_1^Q(y_{(1)0})$, then $\text{sign}(\frac{P(s_i)}{Q(s_i)}) = b_i$ for $i = 1, 2, \dots, m$, where $g_1^P(y_{(1)})$ is the normalizing constant function of $(y_{(1)}, y_{(2)0}, \dots, y_{(m_1)0}, y_{(m_1+1)0}, \dots, y_{(m)0}, y_{(m+1)}, \dots, y_{(d)})$.

If $g_1^P(y_{(1)0}) < g_1^Q(y_{(1)0})$, there exists an $\epsilon > 0$ such that $g_1^P(y_{(1)0}) = g_1^Q(y_{(1)0} - \epsilon)$ as long as each element in $\{\epsilon_{m_1+1}, \dots, \epsilon_m\}$ is small enough and $y_{(1)0}$ is an interior of the strictly increasing interval of the normalizing constant function of $y_{(1)}$. Denoting the distribution corresponds to $y_{(1)0} - \epsilon, y_{(2)0}, \dots, y_{(m_1)0}, y_{(m_1+1)0} + \epsilon_{m_1+1}, \dots, y_{(m)0} + \epsilon_m$ as Q_1, then $\text{sign}(\frac{P(s_i)}{Q_1(s_i)}) = b_i$ for $i = 1, 2, \dots, m$.

If $g_1^P(y_{(1)0}) > g_1^Q(y_{(1)0})$, the form of strictly increasing interval of the normalizing function $g_m^Q(y_{(m)})$ is $(\sqrt{\frac{k_2}{k_1}}, +\infty)$ or $(0, +\infty)$, hence there is an $\epsilon > 0$ such that $g_1^P(y_{(1)0}) = g_m^P(y_{(m)0}) = g_m^Q(y_{(m)0} + \epsilon_m + \epsilon)$, where $k_1 > 0$, $k_2 > 0$ (In fact, $g_m^Q(y_{(m)})$ can be replaced by any $g_j^Q(y_{(j)})$, where $j \in \{m_1 + 1, \dots, m\}$). Denoting the distribution corresponds to $y_{(1)0}, y_{(2)0}, \dots, y_{(m_1)0}, y_{(m_1+1)0} + \epsilon_{m_1+1}, \dots, y_{(m)0} + \epsilon_m + \epsilon$ as Q_1, then $\text{sign}(\frac{P(s_i)}{Q_1(s_i)}) = b_i$ for $i = 1, 2, \dots, m$.

Case 1 and Case 2 together complete the proof of this conclusion, because all the $2^n - 2$ values (except the two values in Case 1) of b containing 1 and -1 can be the form of Case 2 through rearranging the order of s_i for $i = 1, 2, \dots, m$. □

Lemma 7 of Nakamura et al. [23] states that $\text{VCdim}(\mathcal{C}) \le \text{Edim}(\mathcal{C})$ holds for every concept class $\mathcal{C}$. Immediately, we have the following result.

Corollary 4.1. *For every Markov network $\mathcal{N} = (\overline{G}, \mathcal{P})$ satisfying $I_{A(\overline{G})} \neq \{0\}$, we have $dim(I_{A(\overline{G})}) \le VCdim(\mathcal{C}_{\mathcal{N}}) \le Edim(\mathcal{C}_{\mathcal{N}})$.*

Consider Markov network and the case that $I_{A(\overline{G})} = \{0\}$, $I_{A(\overline{G})} = \{0\}$ if and only if $\overline{G}$ is a complete graph. If $I_{A(\overline{G})} = \{0\}$, then $\dim(I_{A(\overline{G})}) = d$, and $\mathcal{P}^+$ is the collection of interior points of the probability simplex, we have $\text{VCdim}(\mathcal{C}_{\mathcal{N}}) = \text{Edim}(\mathcal{C}_{\mathcal{N}}) = d - 1$.

4.2. An upper bound of Euclidean dimension

In this subsection, using $\dim(I_{A(G)})$, we study an upper bound of Euclidean dimension of the concept class induced by a Markov (or Bayesian) network .

Lemma 4.2. *Given a Markov (or Bayesian) network $\mathcal{N} = (G, \mathcal{P})$, we have $\text{Edim}(\mathcal{C}_{\mathcal{N}}) \le \dim(I_{A(G)})$.*

Proof. $\text{Edim}(\mathcal{C}_{\mathcal{N}}) = m$ means that there are two matrices $U_{|\mathcal{C}_{\mathcal{N}}| \times m}$ and $V_{m \times d}$ such that

$$\begin{pmatrix} f_1(x_1) & f_1(x_2) & \cdots & f_1(x_d) \\ f_2(x_1) & f_2(x_2) & \cdots & f_2(x_d) \\ \vdots & \vdots & \ddots & \vdots \\ f_{|\mathcal{C}_{\mathcal{N}}|}(x_1) & f_{|\mathcal{C}_{\mathcal{N}}|}(x_2) & \cdots & f_{|\mathcal{C}_{\mathcal{N}}|}(x_d) \end{pmatrix} = (\text{sign}((UV)_{ij})), \quad (3)$$

and $\forall\, U_{|\mathcal{C}_{\mathcal{N}}| \times r}$, $V_{r \times d}$ satisfying $r \le m - 1$, (3) does not hold, where $f_i \in \mathcal{C}_{\mathcal{N}}$, and $1 \le i \le |\mathcal{C}_{\mathcal{N}}|$, $1 \le j \le d$.

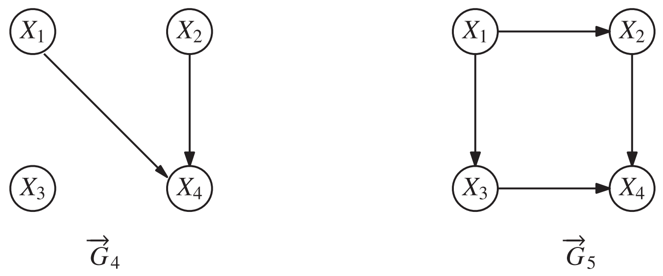

Fig. 4. Two DAGs with V-structures.

$\forall$ P_i, $Q_i \in \mathcal{P}^+$ $(1 \le i \le |\mathcal{C}_{\mathcal{N}}|)$ such that $f_i(x) = \text{sign}(log(P_i(x)/Q_i(x)))$, consider the $|\mathcal{C}_{\mathcal{N}}| \times d$ matrix

$$\begin{pmatrix} \log(\frac{P_1(x_1)}{Q_1(x_1)}) & \log(\frac{P_1(x_2)}{Q_1(x_2)}) & \cdots & \log(\frac{P_1(x_d)}{Q_1(x_d)}) \\ \log(\frac{P_2(x_1)}{Q_2(x_1)}) & \log(\frac{P_2(x_2)}{Q_2(x_2)}) & \cdots & \log(\frac{P_2(x_d)}{Q_2(x_d)}) \\ \vdots & \vdots & \ddots & \vdots \\ \log(\frac{P_{|\mathcal{C}_{\mathcal{N}}|}(x_1)}{Q_{|\mathcal{C}_{\mathcal{N}}|}(x_1)}) & \log(\frac{P_{|\mathcal{C}_{\mathcal{N}}|}(x_2)}{Q_{|\mathcal{C}_{\mathcal{N}}|}(x_2)}) & \cdots & \log(\frac{P_{|\mathcal{C}_{\mathcal{N}}|}(x_d)}{Q_{|\mathcal{C}_{\mathcal{N}}|}(x_d)}) \end{pmatrix}. \quad (4)$$

If $\dim(I_{A(G)}) \le m - 1$, without loss of generality, we assume $\dim(I_{A(G)}) = m - 1$. Then, $\forall\, y_{(1)}, \dots, y_{(m)}$, there is at least one binomial $f \in I_{A(G)} \cap \mathbb{R}[y_{(1)}, \dots, y_{(m)}]$ such that $f(P_i(x_1), \dots, P_i(x_d)) = 0$ and $f(Q_i(x_1), \dots, Q_i(x_d)) = 0$ for $1 \le i \le |\mathcal{C}_{\mathcal{N}}|$. Thus, we know that the rank of the matrix in (3) is less than $m - 1$. This fact contradicts $\forall\, U_{|\mathcal{C}_{\mathcal{N}}| \times r}$, $V_{r \times d}$ satisfying $r \le m - 1$, (3) does not hold. The conclusion is confirmed. □

Lemma 4.2 provides an upper bound of Euclidean dimension of concept classes induced by Markov networks and Bayesian networks. Although the upper bounds are larger than real ones for some Bayesian networks with V-structures, we note that the upper bound is sharp.

Example 4.2. Consider $\overrightarrow{G}_4$ and $\overrightarrow{G}_5$ in Fig. 4. Suppose that $|X_i| = 2$ for $i = 1, 2, 3, 4$. Using the computer algebra system Singular [10], we obtain that $\dim(I_{A(\overrightarrow{G}_4)}) = 9$, $\dim(I_{A(\overrightarrow{G}_4)}) = 11$. Yang and Wu [38] showed that $\text{VCdim}(\mathcal{C}_{\mathcal{N}_4}) = \text{Edim}(\mathcal{C}_{\mathcal{N}_4}) = 8$ and $\text{VCdim}(\mathcal{C}_{\mathcal{N}_5}) = \text{Edim}(\mathcal{C}_{\mathcal{N}_5}) = 11$.

Remark 4.1. For general Bayesian networks, Yang and Wu [37] provided lower bounds of VC dimension and upper bounds of Euclidean dimension. Examples presented in Yang and Wu [38] indicate that both the two bounds are sharp. Classifying the relationship between the upper bounds appeared in Yang and Wu [37] and the one posed in Lemma 4.2 is an interesting problem.

In fact, for a general Bayesian network, neither the prime ideal ker(Φ) presented in Garcia et al. [14], nor the toric ideal we defined in Section 2.6 is suitable to connect corresponding VC dimension and Euclidean dimension.

5. Application of VC dimension in estimating classification error

Hastie et al. [19] told us that the VC dimension can be used to construct an estimate of (extra-sample) prediction error, and there are different types of results. However, as they pointed out, the difficulty is how to calculate the VC dimension of a class of functions. In this section, we introduce a result in binary classification through an example which comes from Han et al. [18].

Example 5.1. Table 1 presents a training set of class-labeled tuples randomly selected from the AllElectronics customer database. The data tuples are described by discrete variables age, income, student, credit-rating. Age and income have three levels, student

Table 1
Class-labeled training tuples from the AllElectronics customer database.

RID	Age	Income	Student	Credit-rating	Class: buy-computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle-aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle-aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle-aged	Medium	No	Excellent	Yes
13	Middle-aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

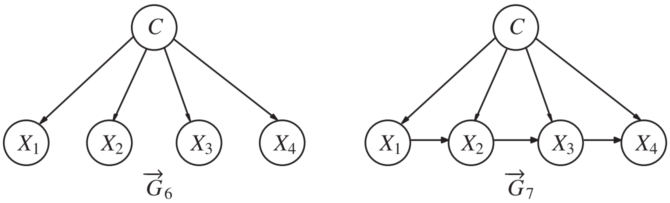

Fig. 5. Naive Bayes ($\vec{G}_6$) and a tree augmented naive Bayes ($\vec{G}_7$) classifier structures with 4 predictor variables.

and credit-rating have two levels. The class label variable, buy-computer, has two levels. Given the training data, estimation of prediction error is our goal.

Generalization error (denoted by $\mathrm{Err}_{\mathcal{T}}$) is the prediction error over an independent test sample. Consider binary classification, if we fit N training points using a class of functions $\{f(x, \alpha)\}$ having VC dimension h, then with probability at least $1 - \eta$ over training sets [page 116 in 7]:

$$\mathrm{Err}_{\mathcal{T}} \le \overline{\mathrm{err}} + \frac{\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4 \cdot \overline{\mathrm{err}}}{\varepsilon}}\right),$$

where $\varepsilon = a_1 \frac{h[log(a_2N/h)+1]-log(\eta/4)}{N}$, $0 < a_1 \le 4$, $0 < a_2 \le 2$; $\overline{\mathrm{err}}$ is the training error, α is a parameter vector. This bound hold simultaneously for all members $f(x, \alpha)$, with $a_1 = 4$ and $a_2 = 2$ corresponding to the worst-case scenarios.

Return to Example 5.1, if one predicts the class label of a tuple using a naive Bayes classifier or a tree augmented naive Bayes classifier, the values of h are 14 and 28, respectively (see $\vec{G}_6$ and $\vec{G}_7$ in Fig. 5, they can be viewed as UGs from the viewpoint of representing conditional independence relationships), and meaningful upper bounds for generalization error can be obtained when the sample size N is sufficient large. To offer an upper bound for prediction error, our results can be used to UG or DAG augmented naive Bayes classifiers, that is, the class C is considered as a root vertex parent of every predictor variables. Meanwhile, we note that, for a Bayesian network with at least one V-structure, if we use $\dim(I_{A(G)})$ to estimate generalization error, the upper bound may be rather loose.

6. Discussion

In this paper, we mainly provide three equivalent characterizations of VC dimension of concept classes induced by Markov networks through viewing dimension of a toric ideal as a bridge between VC dimension and Euclidean dimension, and present upper bounds for Euclidean dimension of concept classes induced by Bayesian networks. This concrete connection between algebraic geometry and machine learning allows one to compute VC dimension of the concept class induced by a Markov network in terms of a computer algebra system directly. We illustrate the utility of our results in calculating prediction error in binary classification. For Bayesian networks without V-structures, based on levels of each variable, Yang and Wu [37] and Varando et al. [33] offered explicit formula to compute Euclidean dimension and thus VC dimension of induced concept classes. A natural question arises, is there an explicit formula for VC dimension of concept classes induced by Markov networks?

For a general Bayesian network $(\vec{G}, \mathcal{P})$ (with at least one V-structure), as far as we know, no other result illustrates the relationship among dimension of the model, VC dimension and Euclidean dimension of concept class induced by $(\vec{G}, \mathcal{P})$ except Yang and Wu [38] for the Bayesian networks with less than four variables. As mentioned in Remark 4.1, it seems difficult to tackle the problem given in Yang and Wu [36] using dimension of ideals, that is, whether the VC dimension is equal to the Euclidean dimension remains open.

Acknowledgments

The authors wish to thank the Editor, an Associate Editor, and three anonymous referees for their valuable comments which have helped improve the presentation of this paper a great deal. This research was supported by the National Natural Science Foundation of China (Grants 11301408, 61573266) and the Natural Science Basic Research Plan in Shaanxi Province (Grant 2017JQ1013).

References

[1] Y. Altun, I. Tsochantaridis, T. Hofmann, Hidden markov support vector machines, in: Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 3–10.
[2] S. Ben-David, N. Eiron, H.U. Simon, Limitations of learning via embeddings in euclidean half spaces, J. Mach. Learn. Res. 3 (2002) 441–461.
[3] C. Bielza, P. Larrañaga, Discrete bayesian network classifier: a survey, ACM Comput. Surv. 47 (1) (2014) 43.
[4] M. Boullé, Compression-based averaging of selective naive bayes classifiers, J. Mach. Learn. Res. 8 (2007) 1659–1685.
[5] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, D. Koller, Max-margin classification of data with absent features, J. Mach. Learn. Res. 9 (2008) 1–21.
[6] S. Chen, G.J. Gordon, R.F. Murphy, Graphical models for structured classification, with an application to interpreting images of protein subcellular location patterns, J. Mach. Learn. Res. 9 (2008) 651–682.
[7] V. Cherkassky, F. Mulier, Learning from Data: Concepts, Theory, and Methods, Wiley-IEEE Press, 2007.
[8] D.M. Chickering, Learning equivalence classes of bayesian-network structures, J. Mach. Learn. Res. 2 (2002) 445–498.
[9] D. Cox, J. Little, D. O'Shea, Ideals, Varieties, and Algorithms, Springer, 2007.
[10] W. Decker, G. Greuel, G. Pfister, H.S. G, Singular 4-0-1, a computer algebra system for polynomial computations, available at http://www.singular.uni-kl.de/, 2014.
[11] V. Franc, B. Savchynskyy, Discriminative learning of max-sum classifiers, J. Mach. Learn. Res. 9 (2008) 67–104.
[12] N. Friedman, Inferring cellular networks using probabilistic graphical models, Science 303 (2004) 799–805.
[13] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.
[14] L.D. Garcia, M. Stillman, B. Sturmfels, Algebraic geometry of bayesian networks, J. Symb. Comput. 39 (2005) 331–355.
[15] D. Geiger, D. Heckerman, H.P. King, C. Meek, Stratified exponential families: graphical models and model selection, Ann. Stat. 29 (2001) 505–529.
[16] F. Ghofrani, A. Keshavarz-Haddad, A. Jamshidi, A new probabilistic classifier based on decompsable models with applicaiton to internet traffic, Pattern Recognit. 77 (2018) 1–11.
[17] D. Geiger, C. Meek, B. Sturmfels, On the toric algebra of graphical models, Ann. Stat. 34 (2006) 1463–1492.
[18] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.
[19] T. Hastie, R.T. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
[20] M.J. Kearns, R.E. Schapire, Efficient distribution-free learning of probabilistic concepts, J. Comput. Sys. Sci. 48 (1994) 464–497.
[21] P. Kohli, L. Ladický, P.H.S. Torr, Robust higher order potentials for enforcing label consistency, Int. J. Comput. Vision 82 (2009) 302–324.
[22] S.L. Lauritzen, Graphical Models, Oxford University Press, 1996.

[23] A. Nakamura, M. Schmitt, N. Schmitt, H.U. Simon, Inner product spaces for bayesian networks, J. Mach. Learn. Res. 6 (2005) 1383–1403.

[24] R.E. Neapolitan, Learning Bayesian Networks, Prentice Hall, 2004.

[25] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.

[26] G. Pistone, E. Riccomagno, H.P. Wynn, Algebraic Statistics: Computational Commutative Algebra in Statistics, Chapman and Hall, 2001.

[27] R. Settimi, J.Q. Smith, Geometry, moments and conditional independence trees with hidden variables, Ann. Stat. 28 (2000) 1179–1205.

[28] M. Studený, Probabilistic Conditional Independence Structures, Springer, 2005.

[29] B. Sturmfels, Gröbner Bases and Convex Polytopes, American Mathematical Society, 1996.

[30] V.Y.F. Tan, S. Sanghavi, J.W. Fisher, A.S. Willsky, Learning graphical models for hypothesis testing and classification, IEEE Trans. Signal Process. 58 (2010) 5481–5495.

[31] B. Taskar, C. Guestrin, D. Koller, Max-margin markov networks, in sebastian thrun, in: L.K. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, 2004, pp. 25–32.

[32] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theor. Prob. Appl. 16 (1971) 264–280.

[33] G. Varando, C. Bielza, P. Larrañaga, Dicision boundary for discrete bayesian network classifiers, J. Mach. Learn. Res. (2016). In press.

[34] T.T. Wong, A hybrid discretization method for naive bayesian classifiers, Pattern Recognit. 45 (6) (2012) 2321–2325.

[35] T.T. Wong, C.R. Liu, A efficient parameter estimation method for generalized dirichlet priors in naive bayesian classifiers with multinomial models, Pattern Recognit. 60 (2016) 62–71.

[36] Y. Yang, Y. Wu, VC dimension and inner product space induced by bayesian networks, Int. J. Approximate Reasoning 50 (2009) 1036–1045.

[37] Y. Yang, Y. Wu, On the properties of concept classes induced by multivalued bayesian networks, Inf. Sci. 184 (2012) 155–165.

[38] Y. Yang, Y. Wu, VE dimension induced by bayesian networks over the boolean domain, Pattern Anal. Appl. 17 (2014) 799–807.

[39] E. Zheleva, L. Getoor, S. Sarawagi, Higher-order graphical models for classification in social and affiliation networks, in NIPS Workshop on networks Across Disciplines: Theory and Applications, 2010.