بسم الله الرحمن الرحیم

بازشناسی الگو

درس ۷

شبکه‌های بیزی

Bayesian Networks

کاظم فولادی قلعه
دانشکده مهندسی، پردیس فارابی
دانشگاه تهران

http://courses.fouladi.ir/pr

بازشناسی الگو

۱

شبکه‌های بیزی

❖ Bayes Probability Chain Rule

$$p(x_1, x_2, \ldots, x_\ell) = p(x_\ell \mid x_{\ell-1}, \ldots, x_1) \cdot p(x_{\ell-1} \mid x_{\ell-2}, \ldots, x_1) \cdot \ldots \cdot p(x_2 \mid x_1) \cdot p(x_1)$$

- Assume now that the conditional dependence for each x_i is limited to a subset of the features appearing in each of the product terms. That is:

$$p(x_1, x_2, \ldots, x_\ell) = p(x_1) \cdot \prod_{i=2}^{\ell} p(x_i \mid A_i)$$

where

$$A_i \subseteq \{x_{i-1}, x_{i-2}, \ldots, x_1\}$$

- For example, if $\ell = 6$, then we could assume:

$$p(x_6 \mid x_5, \ldots, x_1) = p(x_6 \mid x_5, x_4)$$

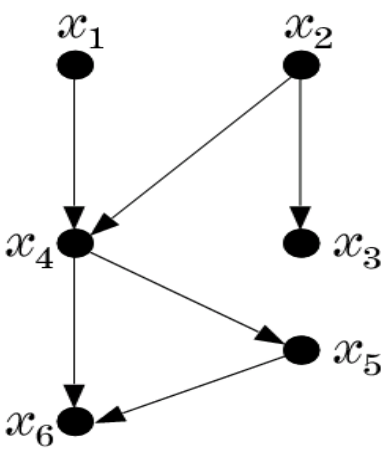
Then:

$$A_6 = \{x_5, x_4\} \subseteq \{x_5, \ldots, x_1\}$$

- The above is a generalization of the Naïve-Bayes. For the Naïve-Bayes the assumption is:

$$A_i = \varnothing, \text{ for } i = 1, 2, \ldots, \ell$$

- A graphical way to portray conditional dependencies is given below



- According to this figure we have that:
  - x_6 is conditionally dependent on x_4, x_5.
  - x_5 on x_4
  - x_4 on x_1, x_2
  - x_3 on x_2
  - x_1, x_2 are conditionally independent on other variables.

- For this case:

$$p(x_1, x_2, \ldots, x_6) = p(x_6 \mid x_5, x_4) \cdot p(x_5 \mid x_4) \cdot p(x_3 \mid x_2) \cdot p(x_2) \cdot p(x_1)$$

شبکه‌های بیزی

Bayesian Networks

شبکه‌ی بیزی

Bayesian Network

یک نمادگذاری ساده و گرافیکی برای بیان استقلال شرطی

(و در نتیجه برای مشخص‌سازی متراکم توزیع‌های توأم کامل)

شبکه‌های بیزی

نحو

Bayesian Networks

شبکه‌ی بیزی
Bayesian Network

یک گراف جهت‌دار بدون دور

گره‌ها *Nodes*	**پیوندها** *Links*
نشان‌دهنده‌ی **متغیرهای تصادفی**	نشان‌دهنده‌ی **رابطه‌ی تأثیر مستقیم**

(link ≈ "directly influences")

توزیع شرطی
Conditional Distribution

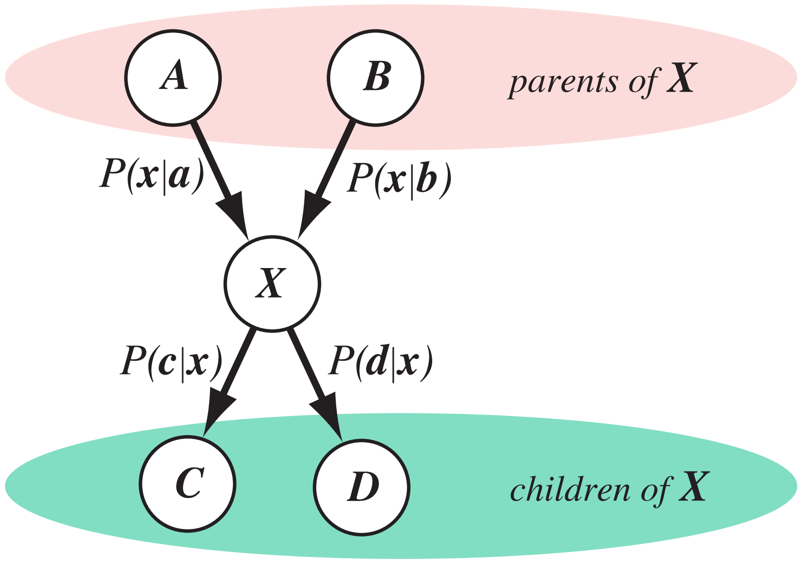
برای هر گره، توزیع شرطی آن گره به‌شرط والدهای آن را داریم:

$$\mathbf{P}(X_i | Parents(X_i))$$

در قالب جدول احتمال شرطی (CPT)

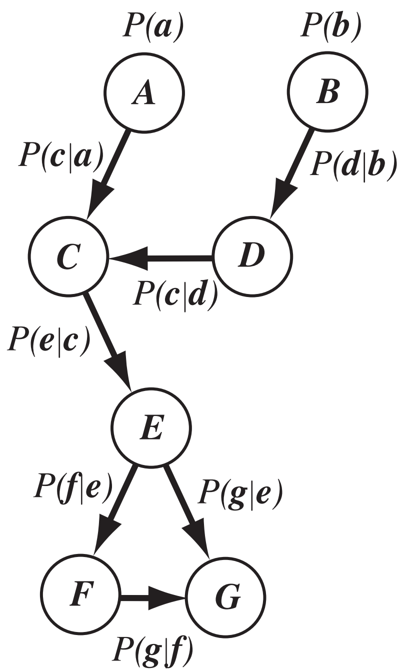
شبکه‌های بیزی

مثال



شبکه‌های بیزی

مثال

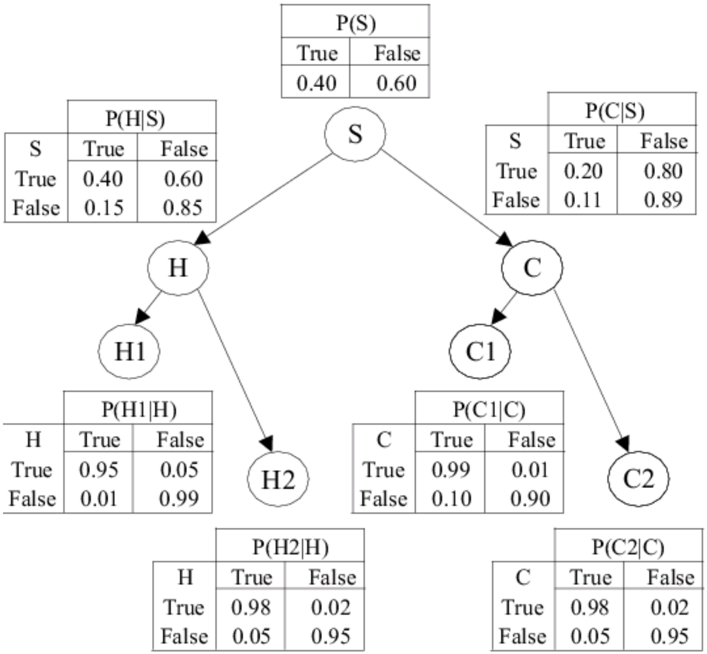


❖ Bayesian Networks

- **Definition:** A Bayesian Network is a directed acyclic graph (DAG) where the nodes correspond to random variables. Each node is associated with a set of conditional probabilities (densities), $p(x_i|A_i)$, where x_i is the variable associated with the node and A_i is the set of its parents in the graph.

- A Bayesian Network is specified by:
  - The marginal probabilities of its root nodes.
  - The conditional probabilities of the non-root nodes, given their parents, for ALL possible combinations.

- The figure below is an example of a Bayesian Network corresponding to a paradigm from the medical applications field.



P(S)	
True	False
0.40	0.60

S	P(H|S) True	P(H|S) False
True	0.40	0.60
False	0.15	0.85

S	P(C|S) True	P(C|S) False
True	0.20	0.80
False	0.11	0.89

H	P(H1|H) True	P(H1|H) False
True	0.95	0.05
False	0.01	0.99

C	P(C1|C) True	P(C1|C) False
True	0.99	0.01
False	0.10	0.90

H	P(H2|H) True	P(H2|H) False
True	0.98	0.02
False	0.05	0.95

C	P(C2|C) True	P(C2|C) False
True	0.98	0.02
False	0.05	0.95

- This Bayesian network models conditional dependencies for an example concerning smokers (S), tendencies to develop cancer (C) and heart disease (H), together with variables corresponding to heart (H_1, H_2) and cancer (C_1, C_2) medical tests.

- Once a DAG has been constructed, the joint probability can be obtained by multiplying the marginal (root nodes) and the conditional (non-root nodes) probabilities.

- Training: Once a topology is given, probabilities are estimated via the training data set. There are also methods that learn the topology.

- Probability Inference: This is the most common task that Bayesian networks help us to solve efficiently. Given the values of some of the variables in the graph, known as evidence, the goal is to compute the conditional probabilities for some of the other variables, given the evidence.

❖ Example: Consider the Bayesian network of the figure:

$P(x1)=0.60$	$P(y1\|x1)=0.40$	$P(z1\|y1)=0.25$	$P(w1\|z1)=0.45$
	$P(y1\|x0)=0.30$	$P(z1\|y0)=0.60$	$P(w1\|z0)=0.30$

$P(x0)=0.40$	$P(y0\|x1)=0.60$	$P(z0\|y1)=0.75$	$P(w0\|z1)=0.55$
	$P(y0\|x0)=0.70$	$P(z0\|y0)=0.40$	$P(w0\|z0)=0.70$
	$P(y1)=0.36$	$P(z1)=0.47$	$P(w1)=0.37$
	$P(y0)=0.64$	$P(z0)=0.53$	$P(w0)=0.63$

a) If x is measured to be $x = 1$ (x_1), compute $P(w = 0|x = 1)$ [$P(w_0|x_1)$].

b) If w is measured to be $w = 1$ (w_1), compute $P(x = 0|w = 1)$ [$P(x_0|w_1)$].

- For a), a set of calculations are required that propagate from node x to node w. It turns out that $P(w_0|x_1) = 0.63$.

- For b), the propagation is reversed in direction. It turns out that $P(x_0|w_1) = 0.4$.

- In general, the required inference information is computed via a combined process of "message passing" among the nodes of the DAG.

❖ Complexity:

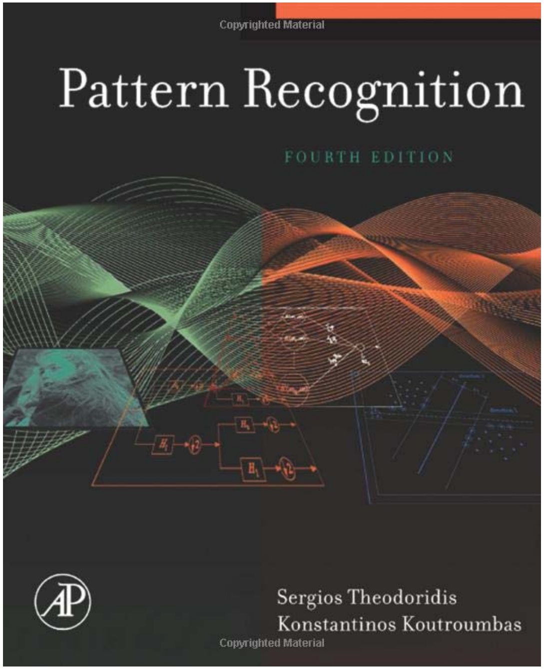
- For singly connected graphs, message passing algorithms amount to a complexity linear in the number of nodes.

شبکه‌های بیزی

۲

منابع

منبع اصلی



S. Theodoridis, K. Koutroumbas,
Pattern Recognition,
Fourth Edition, Academic Press, 2009.

Chapter 2

CHAPTER

Classifiers Based on Bayes Decision Theory

2

2.1 INTRODUCTION

This is the first chapter, out of three, dealing with the design of the classifier in a pattern recognition system. The approach to be followed builds upon probabilistic arguments stemming from the statistical nature of the generated features. As has already been pointed out in the introductory chapter, this is due to the statistical variation of the patterns as well as to the noise in the measuring sensors. Adopting this reasoning as our kickoff point, we will design classifiers that classify an unknown pattern in the most probable of the classes. Thus, our task now becomes that of defining what "most probable" means.

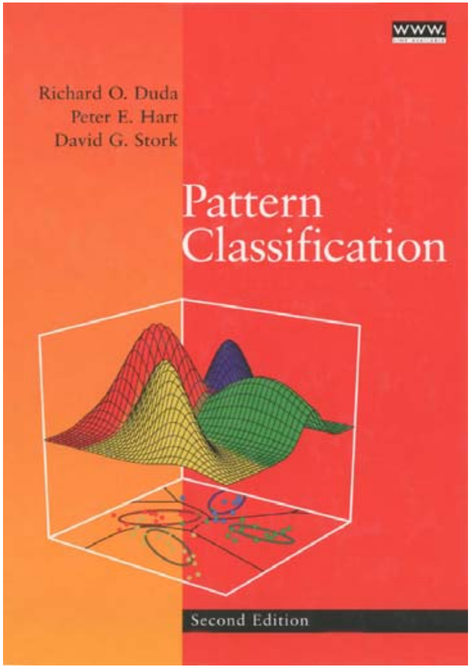
Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector $\boldsymbol{x}$, we form the M conditional probabilities $P(\omega_i|\boldsymbol{x}), i = 1, 2, \ldots, M$. Sometimes, these are also referred to as *a posteriori probabilities*. In words, each of them represents the probability that the unknown pattern belongs to the respective class ω_i, given that the corresponding feature vector takes the value $\boldsymbol{x}$. Who could then argue that these conditional probabilities are not sensible choices to quantify the term *most probable*? Indeed, the classifiers to be considered in this chapter compute either the maximum of these M values or, equivalently, the maximum of an appropriately defined function of them. The unknown pattern is then assigned to the class corresponding to this maximum.

The first task we are faced with is the computation of the conditional probabilities. The Bayes rule will once more prove its usefulness! A major effort in this chapter will be devoted to techniques for estimating probability density functions (pdf), based on the available experimental evidence, that is, the feature vectors corresponding to the patterns of the training set.

2.2 BAYES DECISION THEORY

We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong. In the sequel, we assume that the *a priori probabilities* 13

منبع اصلی



R.O. Duda, P.E. Hart, and D.G. Stork,
Pattern Classification,
Second Edition, John Wiley & Sons, Inc., 2001.

Chapter 2 (2.11)

C H A P T E R 2

BAYESIAN DECISION THEORY

2.1 INTRODUCTION

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. In this chapter we develop the fundamentals of this theory and we show how it can be viewed as being simply a formalization of common-sense procedures; in subsequent chapters we will consider the problems that arise when the probabilistic structure is not completely known.

While we will give a quite general, abstract development of Bayesian decision theory in Section 2.2, we begin our discussion with a specific example. Let us reconsider the hypothetical problem posed in Chapter 1 of designing a classifier to separate two kinds of fish: sea bass and salmon. Suppose that an observer watching fish arrive along the conveyor belt finds it hard to predict what type will emerge next and that the sequence of types of fish appears to be random. In decision-theoretic terminology we would say that as each fish emerges nature is in one or the other of the two possible states: Either the fish is a sea bass or the fish is a salmon. We let ω denote the *state of nature*, with $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon. Because the state of nature is so unpredictable, we consider ω to be a variable that must be described probabilistically.

STATE OF NATURE

If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon. More generally, we assume that there is some *a priori probability* (or simply *prior*) $P(\omega_1)$ that the next fish is sea bass, and some prior probability $P(\omega_2)$ that it is salmon. If we assume there are no other types of fish relevant here, then $P(\omega_1)$ and $P(\omega_2)$ sum to one. These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears. It might, for instance, depend upon the time of year or the choice of fishing area.

PRIOR

Suppose for a moment that we were forced to make a decision about the type of fish that will appear next without being allowed to see it. For the moment, we shall assume that any incorrect classification entails the same cost or consequence, and

20

Prepared by Kazim Fouladi | Fall 2018 | 2nd Edition