Bayesian Learning

Conditional Probability

Bayesian Reasoning

Naïve Bayes Classifiers

Bayesian Networks

Filtering SPAM

The Basic Problem

Our electronic mailboxes fill up with vast numbers of unwanted junk mail trying to sell us things - SPAM

The Basic Solution

A filter that can determine whether an e-mail is SPAM and delete it before delivery.

Subproblem

How does the filter know which e-mails are SPAM?

Solution to Subproblem

Given two sets of example e-mail messages:

Set A comprising only SPAM e-mails

Set B comprising only non-SPAM e-mails

the filter could *learn* to identify SPAM e-mails

CONDITIONAL PROBABILITY

The conditional probability of event X occurring given that event Y has occurred is:

$$P(X \mid Y) \equiv \frac{P(X \wedge Y)}{P(Y)}$$

- Do not confuse $P(X|Y)$ with $P(Y|X)$.

  The probability that someone is the US President, given that they are American, is about 4×10^{-9}.

  The probability that someone is American, given that they are the US President, is 1.0.

Bayes Theorem

There is a simple relationship between $P(X|Y)$ and $P(Y|X)$.

This is Bayes Theorem.

The relationship is also usefully expressed as:

$$P(X \mid Y) \equiv \frac{P(X \wedge Y)}{P(Y)} = \frac{P(X \wedge Y) \times P(X)}{P(Y) \times P(X)} = \frac{P(X \wedge Y)}{P(X)} \times \frac{P(X)}{P(Y)}$$

$$= P(Y \mid X) \times \frac{P(X)}{P(Y)}$$

$$P(X \wedge Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Independence

X and Y are said to be independent if

$$P(X \wedge Y) = P(X) \times P(Y)$$

This means:

- There is no statistical relationship between X and Y.
- Knowing the value of one gives you no help in predicting the value of the other.

Note that if X and Y are independent:

$$P(X \mid Y) = P(X)$$

Conditional Independence

Often we are interested in whether there is a statistical relationship between two variables in the special case that a third variable has a particular value.

X is said to be *conditionally independent* of Y given Z if

$$\forall x_i, y_j, z_k (P(X = x_i \mid Y = y_j \wedge Z = z_k) = P(X = x_i \mid Z = z_k))$$

where x_i, y_j and z_k are possible values of X, Y and Z.

This is usually written as

$$P(X \mid Y \wedge Z) = P(X \mid Z)$$

BAYESIAN REASONING

Bayes theorem is of immense practical value in drawing conclusions from evidence.

An Example

Suppose a person is sneezing and you have 3 hypotheses:

A cold Hay fever Healthy

Which is more likely?

Suppose you also know three unconditional probabilities:

Probability that, at a given time, any member of the population has a cold, $P(\text{cold})$.

Probability that, at a given time, any member of the population has hay fever, $P(\text{hayfever})$.

$P(\text{cold})$ and $P(\text{hayfever})$ are called *prior probabilities*.

Probability that any member of the population sneezes $P(\text{sneeze})$

And two conditional probabilities:

Probability that someone who has a cold will sneeze, $P(\text{sneeze}|\text{cold})$.

Probability that someone who has hay fever will sneeze, $P(\text{sneeze}|\text{hayfever})$.

Then the *posterior probability*, $P(\text{cold}|\text{sneeze})$, is:

$$P(\text{cold} \mid \text{sneeze}) = P(\text{sneeze} \mid \text{cold}) \times \frac{P(\text{cold})}{P(\text{sneeze})}$$

And the *posterior probability*, $P(\text{hay fever}|\text{sneeze})$, is:

$$P(\text{hayfever} \mid \text{sneeze}) = P(\text{sneeze} \mid \text{hayfever}) \times \frac{P(\text{hayfever})}{P(\text{sneeze})}$$

What is $P(\text{healthy}|\text{sneeze})$?

Now let's put in some numbers:

$$P(\text{cold}) = 0.02,\ P(\text{hayfever}) = 0.05,\ P(\text{sneeze}) = 0.08.$$

$$P(\text{sneeze}|\text{cold}) = 0.98,\ P(\text{sneeze}|\text{hayfever}) = 0.75.$$

Then:

$$P(\text{hayfever} \mid \text{sneeze}) = 0.75 \times \frac{0.05}{0.08} = 0.469.$$

$$P(\text{cold} \mid \text{sneeze}) = 0.98 \times \frac{0.02}{0.08} = 0.245.$$

$$P(\text{healthy} \mid \text{sneeze}) = ?$$

So we can conclude that hay fever is the more likely explanation when we observe a sneeze.

What if we don't observe a sneeze?

Maximum A Posteriori Hypotheses

More generally, given a set of mutually exclusive hypotheses H and evidence E,

The *maximum a posteriori* (MAP) hypothesis is given by:

$$h_{\text{MAP}} = \arg\max_{h \in H} P(h \mid E) = \arg\max_{h \in H} P(E \mid h) \times \frac{P(h)}{P(E)}$$

$$= \arg\max_{h \in H} P(E \mid h) \times P(h)$$

IF exact conditional probabilities are known, the performance of a complete Bayesian system is optimal: no better classifier could be built using the same evidence.

More Than One Piece of Evidence

Suppose the patient concerned also had a fever.

How could this be brought into the diagnosis?

We would need to know:

$$P(\text{sneeze} \wedge \text{fever} \mid \text{cold})$$

$$P(\text{sneeze} \wedge \text{fever} \mid \text{hayfever})$$

And the most likely hypothesis would be given by:

$$h_{MAP} = \underset{h \in \{cold, hayfever, healthy\}}{\arg\max} P(\text{sneeze} \wedge \text{fever} \mid h) \times P(h)$$

Thus if we have n pieces of evidence

$$h_{MAP} = \underset{h \in H}{\arg\max} P(E_1 \wedge \ldots \wedge E_i \wedge \ldots \wedge E_n \mid h) \times P(h)$$

But of course, patients may exhibit different combinations of symptoms. To deal with these we would also need to know:

$$P(\neg \text{sneeze} \wedge \text{fever} \mid \text{cold})$$

$$P(\neg \text{sneeze} \wedge \text{fever} \mid \text{hayfever})$$

$$P(\text{sneeze} \wedge \neg \text{fever} \mid \text{cold})$$

$$P(\text{sneeze} \wedge \neg \text{fever} \mid \text{hayfever})$$

$$P(\neg \text{sneeze} \wedge \neg \text{fever} \mid \text{cold})$$

$$P(\neg \text{sneeze} \wedge \neg \text{fever} \mid \text{hayfever})$$

Combinatorial Explosion

This illustrates the major practical limitation of Bayesian methods.

Suppose we were to construct a more realistic diagnostic system using these ideas.

Suppose there n possible symptoms to consider

We would need estimates of all the conditional probabilities $P(E_1 \wedge \ldots \wedge E_n | h)$ in order to determine h_{MAP}.

How many of these would there be?

As many as there are possible combinations of evidence.

This number rises exponentially with n.

Thus, even for simple binary symptoms, we would need 2^n conditional probabilities for each possible hypothesis (i.e. diagnosis)

BAYESIAN LEARNING

Estimating Probabilities

Estimating probabilities is essentially a matter of counting the occurrences of particular combinations of values in the training data set.

Thus $P'(c_j)$, an estimate of $P(c_j)$, is given by

$$P'(c_j) = \frac{F(c_j)}{N}$$

where N is the size of the training set and $F(c_j)$ is the number of examples of class c_j.

Similarly $P'(a_i \wedge b_k | c_j)$ is given by

$$P'(a_i \wedge b_k \mid c_j) = \frac{F(a_i \wedge b_k \wedge c_j)}{F(c_j)}$$

where $F(a_i \wedge b_k \wedge c_j)$ is the number of examples that are in class c_j and have value a_i for attribute A and b_k for attribute B.

Dealing with very low counts:

This method works well when the training set contains sufficient examples in the relevant region of example space to enable such estimates to be made.

It can run into problems as counts approach and reach zero.

The standard solution is to use a correction that estimates the value that would have been obtained with m additional examples:

$$\text{mestimate} = \frac{F(..... \wedge c_j) + mp}{F(c_j) + m}$$

For a variable A_i, p is typically $1/k$ where A_i takes k values.

Complete Bayes Classifiers

Suppose we had a training set of examples in the form of feature vectors $\mathbf{A}$ with categorical attributes $A_1 \ldots A_n$ and a classification variable C.

In principle we could construct a classifier using the relationship:

$$c_{MAP} = \arg\max_{c \in C} P(c \mid a_1 \wedge \ldots \wedge a_n)$$

$$= \arg\max_{c \in C} P(a_1 \wedge \ldots \wedge a_n \mid c) \times \frac{P(c)}{P(a_1 \wedge \ldots \wedge a_n)}$$

$$= \arg\max_{c \in C} P(a_1 \wedge \ldots \wedge a_n \mid c) \times P(c)$$

where $a_1 \ldots a_n$ are the values taken by attributes $A_1 \ldots A_n$.

We could readily estimate the values of $P(c)$ by counting the occurrence of each value of c in the training set.

Estimating all the conditional probabilities $P(a_1 \wedge \ldots \wedge a_n | c_j)$ would only be possible if the data set contained a number of examples (10?) of every feature vector in the example space.

So if there were as few as 10 attributes each taking 4 values we would need at least $4^{10} \text{x} 10 \simeq 10^7$ training examples!

Thus this approach is *not feasible* unless the feature space is very small.

The performance of a complete Bayesian system is optimal (no better classifier could be built using the same evidence) only **IF** exact conditional probabilities are known.

Naive Bayes Classifiers

The complete Bayes classifier is impractical because so much data is required to estimate the conditional probabilities.

Can we get round this problem by finding a much more economical way to estimate them.

Assuming Conditional Independence

Suppose we assume that all the attributes are conditionally independent given the classification variable.

Then

$$P(a_1 \wedge \ldots \wedge a_n \mid c_j) = \prod_{i=1}^{n} P(a_i \mid c_j)$$

For our simple diagnostic system, this assumption would be that

$P(\text{sneeze} \wedge \text{fever} \mid \text{cold}) = P(\text{sneeze} \mid \text{cold}) \times P(\text{fever} \mid \text{cold})$

or, equivalently: $P(\text{sneeze} \mid \text{fever} \wedge \text{cold}) = P(\text{sneeze} \mid \text{cold})$

The number of conditional probabilities required is thus drastically reduced.

If there are 10 attributes, each taking 4 values, and 2 classes, only 80 conditional probabilities need be estimated.

This is quite feasible using a reasonably sized training set.

Filtering Spam Using Naïve Bayes

A reasonable simplification is to regard e-mails as text objects.

Thus each e-mail can be viewed as a sequence of words.

Any given e-mail will include many words

Some are much *more* likely to occur in Spam than in other e-mails. e.g.

P("Viagra"|Spam) >> P("Viagra"|Not Spam)

Some are much *less* likely to occur in Spam than in other e-mails. e.g.

P("Timetable"|Spam) << P("Timetable"|Not Spam)

And some are about equally likely to occur in Spam and in other e-mails. e.g.

P("and"|Spam) ≅ P("and"|Not Spam)

Suppose we knew all these conditional probabilities.

Then we could apply Naïve Bayes to estimate the probability that a given e-mail was or was not Spam.

Given an e-mail comprised of words $w_1 \ldots w_n$

$$P(\text{Spam} \mid w_1 \wedge \ldots \wedge w_n) = P(w_1 \wedge \ldots \wedge w_n \mid \text{Spam}) \times \frac{P(\text{Spam})}{P(w_1 \wedge \ldots \wedge w_n)}$$

$$P(\text{NotSpam} \mid w_1 \wedge \ldots \wedge w_n) = P(w_1 \wedge \ldots \wedge w_n \mid \text{NotSpam}) \times \frac{P(\text{NotSpam})}{P(w_1 \wedge \ldots \wedge w_n)}$$

Hence, using Naïve Bayes approximation,

$$\text{MostLikelyClass} = \underset{c \in \{\text{Spam}, \text{NotSpam}\}}{\arg\max} \; P(c) \times \prod_i P(w_i \mid c)$$

where the product is taken over all the words in the candidate e-mail.

Estimating the Probabilities

$$P(\text{Spam}) = \frac{N_{\text{Spam}}}{N_{\text{Spam}} + N_{\text{NotSpam}}}$$

$$P(\text{NotSpam}) = \frac{N_{\text{NotSpam}}}{N_{\text{Spam}} + N_{\text{NotSpam}}}$$

where N_{Spam} is the number of examples of Spam and N_{NotSpam} is the number of examples that are not Spam.

$$P(w_i \mid \text{Spam}) = \frac{n_{i,\text{Spam}} + 1}{n_{\text{Spam}} + \text{NumWords}}$$

where

Numwords is the total number of distinct words found in the entire set of examples.

n_{Spam} is the total number of words found in the set of Spam examples.

$n_{i,\text{Spam}}$ is the number of times the word w_i occurs in the set of Spam examples.

Values for $P(w_i|\text{NotSpam})$ are obtained similarly.

Note that an m-estimate is used for the conditional probabilities because many words will have very low counts.

NUMERIC ATTRIBUTES

The examples we have discussed have all involved only categorical attributes.

What happens if some (or even all) of the attributes are numeric?

Two types of solution:

1. Discretization

Map the numeric attribute to a set of discrete values and treat the result in the same way as a categorical attribute.

e.g. Temperature ^{o}Celsius > {cold, cool, normal, warm hot}

We will discuss this approach in more detail later.

2. Assume a distribution

Assume the numeric attribute has a particular probability distribution. Gaussian (i.e. normal) is the usual assumption.

Parameters of the distribution can be estimated from the training set.

One such distribution is needed for each hypothesis for each attribute.

These distributions can then be used to compute the probabilities of a given value of the attribute occurring for ach hypothesis.

How Well Do Naive Bayes Classifiers Work?

Comparative studies show that Naïve Bayesian classifiers often perform surprisingly well.

Results obtained for many data sets are typically comparable to those produced by decision trees and back-propagation.

Why is this surprising?

Because the conditional independence assumption is very strong.

It is often not a realistic assumption. e.g.

A person's affluence is likely to depend on both amount of education and parental affluence.

But these two factors do not have independent effects since parental affluence is likely to influence the amount of education.

So why do Naïve Bayesian methods work?

Three hypotheses:

1. In most data sets, the conditional independence assumption is realistic.

   But, interactions between the variables are very common in real data sets

2. Even when the conditional independence assumptions are invalid, the answers produced will be right most of the time.

   That is, the naive Bayesian model is a good approximation.

3. Because the naïve Bayes system makes strong assumptions and builds a simple model, it needs less data than other methods.

BAYESIAN BELIEF NETWORKS

What do we do when the naive Bayesian approach provides too crude a model?

The complete Bayesian model, incorporating all conditional probabilities, is not feasible.

A compromise:

Build a model that

- Specifies which conditional independence assumptions are valid.
- Provides sets of conditional probabilities to specify the joint probability distributions wherever dependencies exist.

Bayesian belief networks achieve these objectives by

- Specifying the conditional independence assumptions in a directed acyclic graph.

  Each node denotes a variable.

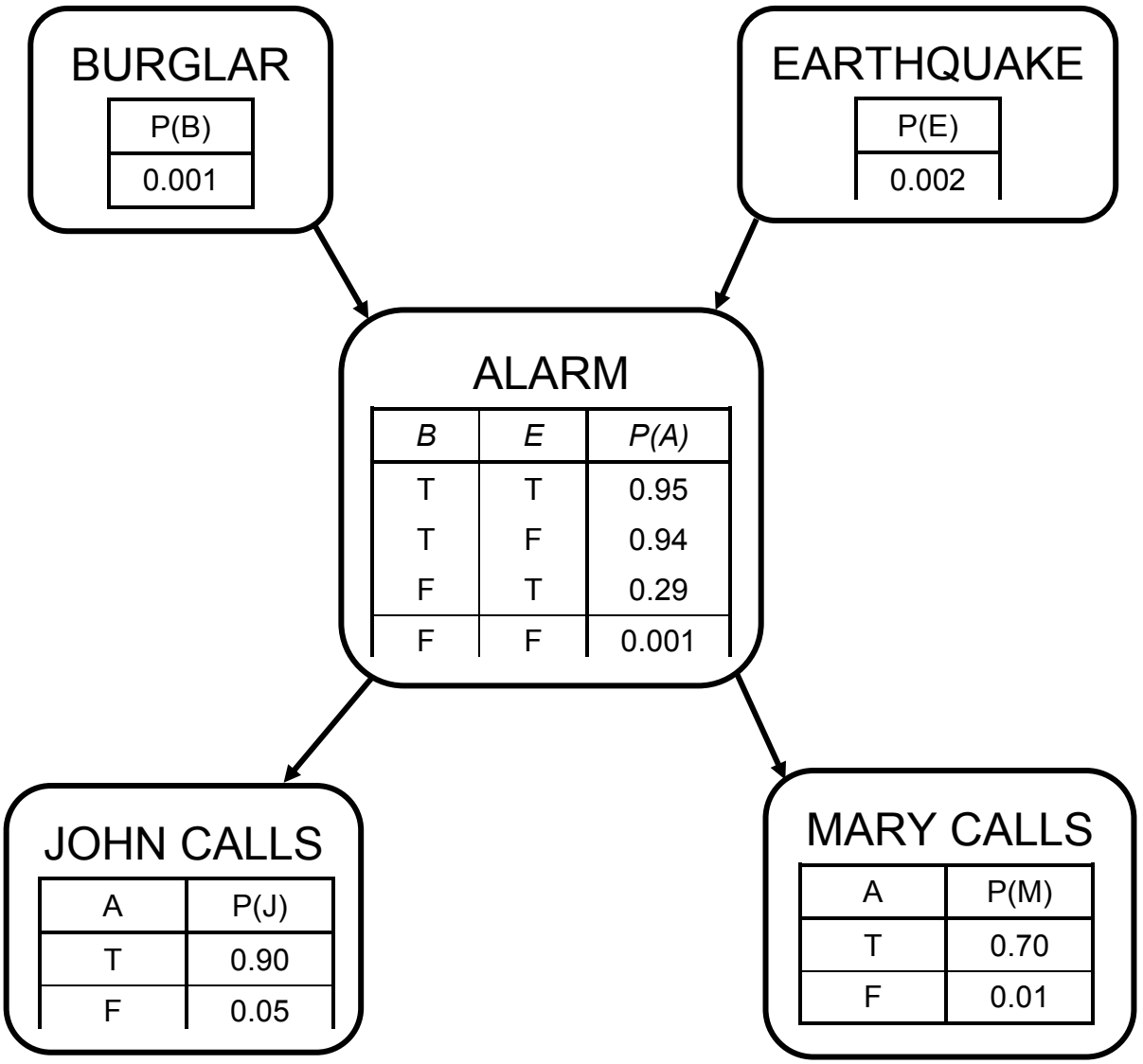
  Each arc indicates a dependency between the nodes at its start and finish.

  Consequently a variable X is conditionally independent of variable Y, given the immediate predecessors of X, if no path exists from X to Y.

- A conditional probability table is provided for each node.

  This specifies the probability distribution of the associated variable given the values of the nodes immediate predecessors.

An Example



There is an alarm which almost always rings if there is a burglary, is sometimes triggered by an earthquake and very occasionally gives a false alarm.

Earthquakes and burglaries are infrequent.

John and Mary are neighbours who will phone the owner if they hear the alarm. They won't always hear it and occasionally imagine they have heard it.

John and Mary's behaviour is conditionally independent given the alarm.

Constructing Belief Networks

Methods for learning Bayesian networks are an active area of current research.

Two situations:

- The structure of the network is known: only the conditional probability tables must be learned.
- The structure of the network is unknown and must be learned.

Network Structure Known

This is obviously the easier learning task.

If data is available on all the variables:

Very straightforward

Simply compute the entries for the conditional probability tables by counting, just as for naive Bayesian classifier.

If data is absent for some variables:

Considerably more difficult.

The space of possible probability assignments must be searched for the one that maximises the likelihood of the available data.

Two established approaches:

Hill climbing along line of steepest ascent.

The EM algorithm.

(See Mitchell pp 188-196 for details).

Network Structure Unknown

A difficult research problem.

UNDERSTANDING OTHER LEARNING PROCEDURES

This lecture has been concerned with the application of Bayesian inference to develop classifier learning procedures.

Bayesian methods are also important because the provide insight into the behaviour of other approaches to learning.

Much of Mitchell's discussion of Bayesian learning is concerned with this second role.

Exercise

The Prisoner Paradox

Three prisoners in solitary confinement, A, B and C, have been sentenced to death on the same day but, because there is a national holiday, the governor decides that one will be granted a pardon. The prisoners are informed of this but told that they will not know which one of them is to be spared until the day scheduled for the executions.

Prisoner A says to the jailer “I already know that at least one the other two prisoners will be executed, so if you tell me the name of one who will be executed, you won’t have given me any information about my own execution”.

The jailer accepts this and tells him that C will definitely die.

A then reasons “Before I knew C was to be executed I had a 1 in 3 chance of receiving a pardon. Now I know that either B or myself will be pardoned the odds have improved to 1 in 2.”.

But the jailer points out “You could have reached a similar conclusion if I had said B will die, and I was bound to answer either B or C, so why did you need to ask?”.

What are A’s chances of receiving a pardon and why? Construct an explanation that would convince others that you are right.

You could tackle this by Bayes theorem, by drawing a belief network, or by common sense. Whichever approach you choose should deepen your understanding of the deceptively simple concept of conditional probability.

Suggested Readings:

*Mitchell, T. M. (1997) "*Machine Learning*", McGraw*-Hill. Chapter 6. (Be careful to distinguish procedures of theoretical importance from those that can actually be used).

Tan, Steinbach & Kumar (2006) "Introduction to Data Mining". Section 5.3

Han & Kamber (2006) "Data Mining: Concepts and Techniques". Section 6.4

Michie, D., *Spiegelhalter, D. J. & Taylor, C.C. (1994) "*Machine learning, neural and statistical classification*" Ellis Horwood. (A series* of articles on evaluations of various machine learning procedures).

An article describing the application of Naïve Bayes learning to filter spam from e-mail can be found at:

http://www.paulgraham.com/spam.html

Implementations

Several implementations of the Naïve Bayes procedure are available as part of the WEKA suite of data mining programs.