

A review of some of the theory and concepts of modern probability theory.

A.1 Some Basics

This section provides a brief overview of some of the concepts used.

A.1.2 Probability

There are two main definitions of probability:

1. Objective (aka Frequentist):

The probability of an event is a real property that can be measured by the *relative frequency* of its occurrence:

$$P(x) = \lim_{n \rightarrow \infty} \frac{f_x(n)}{n} \text{ where } f_x(n) \text{ is the frequency of } x \text{ in } n \text{ independent trials.}$$

2. Subjective (aka Bayesian):

The probability of an event is a measure of an observer's *degree of belief* or *uncertainty* that the event will occur rather than having any external significance.

There are other definitions such as that due to Bruno de Finetti but it should be noted that all are governed by the same rules of probability and that in the following discussion no assumptions are made about which definition is most appropriate.

A.1.3 Conditional Probability

The conditional probability of an event A given that B has occurred is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and where $P(A|B)P(B) = P(A \cap B) = P(A)P(B)$ then A is independent B .

In the context of credit-card fraud we have a binary random variable $Z = \{legal, fraud\}$ and in much that follows it is important to note that although $P(legal) + P(fraud) = 1.0$ in general $P(legal|x) + P(fraud|x) \leq 1$ for some variable x and where the events *legal* and *fraud* are independent of x then the sum is 1.0

$P(legal x) + P(fraud x) = \frac{P(legal \cap x)}{P(x)} + \frac{P(fraud \cap x)}{P(x)}$ <p>and given the independence of x</p> $\frac{P(legal)P(x)}{P(x)} + \frac{P(fraud)P(x)}{P(x)} = P(legal) + P(fraud) = 1$

This is a necessary but not a sufficient condition for independence

(see next section)

A.1.4 Total Probability Theorem

If we assume that the event A is drawn from a binary set of mutually-exclusive events like $\{\text{legal}, \text{fraud}\}$ then we know that $P(\text{legal}) + P(\text{fraud}) = P(A) + P(!A) = 1$. However, as we also know that either *legal* or *fraud* must occur we can also say that $P(A|B) + P(!A|B) = 1$

So, as $P(A \cap B) = P(A|B)P(B)$ and equally $P(!A \cap B) = P(!A|B)P(B)$ we can write

$$P(A \cap B) + P(!A \cap B) = [P(A|B) + P(!A|B)]P(B) = P(B)$$

and then using $P(A|B)P(B) = P(B|A)P(A)$ we have

$$P(B|A)P(A) + P(B|!A)P(!A) = P(B)$$

This result is true because we imposed the two conditions on A :

- 1 It is a set of mutually exclusive events
- 2 One event must occur

We can then generalise from the two-valued event to a multi-valued event A so that

$$P(B) = \sum_i P(A_i \cap B) = \sum_i P(B|A_i)P(A_i)$$

A.1.5 Chain rule of conditional probabilities

From the definition of conditional probability

$$P(x_1 \cap x_2) = P(x_1 | x_2)P(x_2)$$

$$P(x_1 \cap x_2 \cap x_3) = P(x_1 | x_2 \cap x_3)P(x_2 | x_3)P(x_3)$$

$$P(x_1 \cap x_2 \cap x_3 \cap x_4) = P(x_1 | x_2 \cap x_3 \cap x_4)P(x_2 | x_3 \cap x_4)P(x_3 | x_4)P(x_4)$$

and hence more generally we can write:

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | x_{i-1}, \dots, x_1)$$

A.1.6 Conditional Independence

An event A is defined to be conditionally independent of event C given B if

$$P(A|B \cap C) = P(A|B)$$

from this definition we can easily see that

$$P(A \cap B | C) = P(A|C)P(B|C)$$

A.1.7 Marginalisation

Marginalisation is a very useful technique for getting rid of irrelevant random variables from a computation. By using the total probability theorem derived in A.1.4

A.2 Bayes Rule

From the definition of conditional probability we have:

$$P(A|B)P(B) = P(A \cap B) = P(B \cap A) = P(B|A)P(A)$$

and so can write:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ which is the basic form of Bayes rule.}$$

Bayes Rule is often interpreted in terms of updating our belief about a hypothesis A in the light of new evidence B . So our *posterior* belief $P(A|B)$ is calculated by multiplying our *prior* belief $P(A)$ by the *likelihood* $P(B|A)$ that B will occur if A is true.

The basic form of Bayes Rule can be cast into a more generally useful form by using the total probability theorem discussed above.

$$P(B) = \sum_j P(B \cap A_j) = \sum_j P(B|A_j)P(A_j)$$

which gives

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_j P(B|A_j)P(A_j)}$$

A.2.1 Naïve Bayesian Classification

The Naïve Bayesian Classifier seeks to assign a vector \mathbf{X} to a class C

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{\sum_j P(\mathbf{X}|C_j)P(C_j)}$$

If we assume that each component of the vector \mathbf{X} is conditionally independent (see A.1.6) then we can write:

$$P(\mathbf{X}|C) = P(X_1 \cap \dots \cap X_n | C) = P(X_1|C) \dots P(X_n|C) = \prod_j P(X_j|C)$$

and hence we have

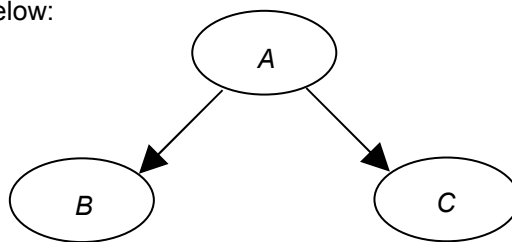
$$P(C|\mathbf{X}) = \frac{P(C) \prod_i P(X_i|C)}{\sum_j P(C_j) \prod_i P(X_i|C_j)}$$

The $P(X_i | C)$ are usually very easy to estimate from the training data. The Naïve Bayesian Classifier has been the subject of a large number of empirical studies in various domains and has been found to be surprisingly effective despite the over simplification inherent in the assumption of conditional independence.

A.3 Bayesian Networks

A Bayesian Network (BN) models the causal relationships of a system or dataset and provides a graphical representation of this causal structure through the use of Directed Acyclic Graphs (DAGs). The DAG representation then provides a framework for inference and prediction.

Consider the BN below:



This represents a causal relationship between 3 random variables A, B and C whereby there is a relationship $A \rightarrow B$ and $A \rightarrow C$ but there is no direct relationship between B and C.

The existence of the causal relationship $A \rightarrow B$ is represented by the arc/edge between the nodes A and B while the strength of the relationship is represented by the conditional probability $P(B | A)$

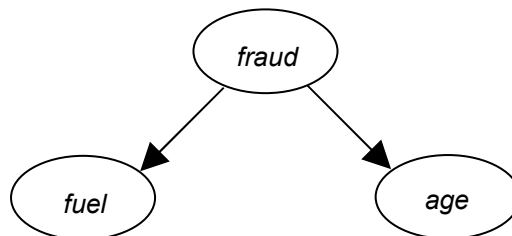
A.3.1 Conditional Independence

Conditional independence is fundamental to the Bayesian Network formalism. It is the central concept from which flows both their causal structure and their ability to reduce the complexity of its representation.

As stated, B and C are conditionally independent (A.1.6)

We can therefore write $P(B | A \cap C) = P(B | A)$ and equally $P(C | A \cap B) = P(C | A)$

A more concrete example of this might be:



Here we have a relationship between *fraud* and the purchase of *fuel* and one between *fraud* and *age*. This BN also captures that fact that we expect *fuel* and *age* to be independent of each other.

Consider the introductory example. Assume we have no prior knowledge about the state of A but we know the state of B. Using Bayes Rule (A.2) we can infer some knowledge about A

and hence increase our knowledge about the state of C. There is then a relationship between the state of B and the state of C. They are not independent.

However, if we now assume prior knowledge about the state of A then whatever we know about B our knowledge of the state of A prevents this from being propagated to C. In this case B and C are independent.

In summary. B and C are conditionally independent given A.

Conditional independence also allows us to prune the number of probability distributions needed to model a process. In general, a system characterised by n random variables will require $2^n - 1$ probability distributions to model it completely. So in the case of the simple 3 variable system above we could need 7. However, by noting the conditional independence of the variables B and C we can reduce this to just 3. This can be seen by considering the joint distribution, which by the chain-rule (A.1.5), can be written:

$$P(A \cap B \cap C) = P(B | A \cap C)P(C | A)P(A)$$

and which can then be simplified using the conditional independence formulae above to just

$$P(A \cap B \cap C) = P(B | A)P(C | A)P(A)$$

From this we can see that the root node's probability $P(A)$ is the *prior* probability that then feeds the conditional probabilities of its child node's B and C.

This leads us to the formal definition.

A.3.2 Definition of a Baysean Network

In general we have by rewriting the chain-rule:

$$P(\mathbf{x}) = P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | x_{i-1}, \dots, x_1) \quad \text{as} \quad P(\mathbf{x}) = \prod_{i=1}^n P(x_i | \text{Parents}(x_i))$$

then by definition we require that:

$P(X_i | \text{Parents}(X_i)) = P(X_i | X_{i-1}, \dots, X_1)$ where X_i is the random variable for which x_i is an instance.

This condition can be satisfied if the following is true:

1. There is an ordering such that $\text{Parents}(X_i) \subseteq \{X_{i-1}, \dots, X_1\}$
2. X_i is conditionally independent of all but its parents

The representation of a BN as a Directed Acyclic Graph ensures these conditions are met.

A.3.3 Types of connection and d-separation

There are three basic types of connection between nodes each representing forms of conditional independence/dependence which motivates the concept of d-separation. Consider in each of the cases below the consequences of introducing evidence about the state of node A, B or C

A.3.3.1 Divergence

This is the example considered in the introduction above.



B and C are conditionally independent given A

B and C are d-separated given A

A.3.3.2 Linear



Knowing the state of A blocks any influence of B on C. Conversely, just knowing B allows us to propagate our belief through A to C.

B and C are d-separated given A

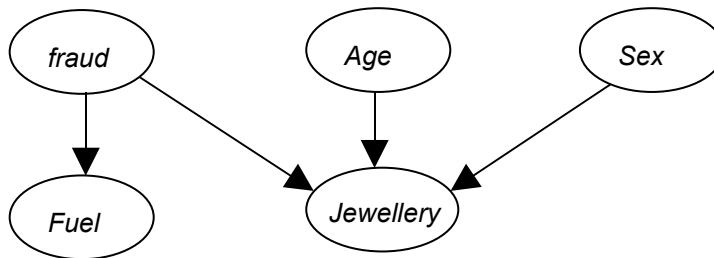
A.3.3.3 Convergent



Knowing the state of A means we can infer knowledge about both B and C. B and C are conditionally dependent on A. If we know nothing about A then B and C are independent.

A.3.4 Inference

To illustrate how inference proceeds in a Bayesian network consider the example below adapted from Heckerman (1999)



If *fraud* is a random variable where $fraud = (true, false)$ and we define $J=Jewellery, F=Fuel, S=Sex, A=Age$

Then using Bayes Rule (A.2) we can write

$$P(fraud | J \cap F \cap S \cap A) = \frac{P(J \cap F \cap S \cap A \cap fraud)}{P(J \cap F \cap S \cap A)}$$

Because the states of *fraud* are mutually exclusive and exhaustive the denominator can be re-written using the total probability theorem (A.1.4) giving:

$$P(fraud | J \cap F \cap S \cap A) = \frac{P(J \cap F \cap S \cap A \cap fraud)}{\sum_j P(J \cap F \cap S \cap A \cap fraud_j)}$$

then applying the chain rule (A.1.5) we have:

$$= \frac{P(J | F \cap S \cap A \cap \text{fraud})P(F | S \cap A \cap \text{fraud})P(S | A \cap \text{fraud})P(A | \text{fraud})P(\text{fraud})}{\sum_j P(J | F \cap S \cap A \cap \text{fraud}_j)P(F | S \cap A \cap \text{fraud}_j)P(S | A \cap \text{fraud}_j)P(A | \text{fraud}_j)P(\text{fraud}_j)}$$

then using the conditional independencies resulting from the causal structure of the BN illustrated above we can simplify this equation to

$$= \frac{P(J | S \cap A \cap \text{fraud})P(F | \text{fraud})P(S)P(A)P(\text{fraud})}{\sum_j P(J | S \cap A \cap \text{fraud}_j)P(F | \text{fraud}_j)P(S)P(A)P(\text{fraud}_j)}$$

we can simplify this equation further as P(S) and P(A) will cancel indicating that the prior probabilities for age and sex make no contribution to the computation of the probability of fraud. So we finally have

$$P(\text{fraud} | J \cap F \cap S \cap A) = \frac{P(J | S \cap A \cap \text{fraud})P(F | \text{fraud})P(\text{fraud})}{\sum_j P(J | S \cap A \cap \text{fraud}_j)P(F | \text{fraud}_j)P(\text{fraud}_j)}$$

All the values in this equation are known from the probability distributions of the nodes in the Bayesian Network and hence the posterior probability for fraud can be calculated.

A.3.5 Network Discovery

The problem of discovering the causal structure of a Bayesian Network is NP-hard (ie. it doesn't get any harder!). Essentially, we need to discover the Directed Acyclic Graph (DAG) that best models the data. However, the number of DAGs for N variables is super-exponential in N. Given the many random variables that characterise a credit-card transaction then the search-space is too big for any automated discovery.

We can reduce the search-space in many ways by introducing constraints:

- Specifying the ordering of the nodes which dictates the hierarchy of dependency in the BN
- Use a priori knowledge about dependencies to build a prototype BN which is then optimised. Dependencies may be suggested by first exploring the data using Association Rule discovery (see A.4)
- Use Monte Carlo methods to explore a bounded set of possibilities

A.4 Association Rules

Association rules have their origins in so called Basket Analysis. By examining the contents of shopping baskets as they pass through a checkout it is possible to determine associations between products that tend to be present in the basket at the same time. An association rule expresses this more formally as:

$$A = a \rightarrow B = b$$

An association rule is expressed in terms of *confidence* and *support* as follows:

$$\text{support}(A \rightarrow B) = \frac{\text{frequency}(A \cup B)}{\text{totalcount}} \text{ and } \text{confidence}(A \rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(B)}$$

An association rule is semantically weak. Nothing can be inferred from the existence of an association rule other than the association exists. This is often all that is required and they are useful for exploratory mining of data. This should be contrasted to a causal model of the data like a Bayesian network.

A.5 Maximum Entropy Methods

A.5.1 Entropy

Entropy can be interpreted in many different ways. Here we view entropy as a measure of the uncertainty (or information) in a system.

Given a discrete system X which consisting of n components x each able to be in one of s micro-states so that $x \in (x_1, \dots, x_s)$ then the number of macro-states of the system X is simply $C = s^n$. C could be regarded as a measure of the ability of the system to store or represent information, so we would expect that by doubling the number of components x to $2n$ that the system would be able to store twice the information. This leads us to define the information as the log of C which is then linear in n , so that:

$$H = \ln(s^n) = n \ln s$$

Now if the probability of each micro-state is equally likely then $p = 1/s$ and we can write:

$$H = -n \ln p$$

More generally the probability of each micro-state is not equal and each state x_i will have a probability p_i . For each micro-state we can write $H_i = -n_i \ln p_i$ where n_i is the expected number of components in state x_i . Then, as for large n we have $n_i \rightarrow np_i$, we have $H_i = -np_i \ln p_i$ and can write

$$H = -n \sum_{i=1}^s p_i \ln p_i$$

This is the well known Shannon entropy measure. There are others measures due to Renyi, Tsallis and others.

A.5.2 Entropy and Random Variables

The entropy for a random variable X can therefore be defined as

$$H(X) = -\sum_i P(x_i) \log P(x_i) \text{ where } X \in \{x_1 \dots x_n\}$$

recalling that the expected value of an arbitrary function of X is $E(f(X)) = \sum_j p(x_j) f(x_j)$

then the entropy can be seen to be

$$H(X) = E(-\sum_j \log P(x_j)) = E(I(X)) \text{ where } I(X) = -\sum_j \log p_j \text{ the Information function.}$$

A.5.2.1 Simple example of entropy

Consider a random variable X that can take just two values with probabilities p and q so that $q = 1 - p$ then the entropy H can be written

$$H = -(p \log p + q \log q) = -(p \log p + (1 - p) \log(1 - p))$$

If we look for the maximum of this equation then we have

$$\frac{dH}{dp} = -(\log p + 1 - \log(1-p) - 1) = \log \frac{1-p}{p} = 0 \text{ and hence } 1-p = p \text{ and } p = 0.5$$

This example illustrates how by maximising the entropy we maximise the uncertainty given no other information. So that here the probability is uniformly distributed among the possible outcomes.

A.5.2.2 Joint Entropy

Joint entropy is simply a generalisation of the definition of entropy and is written

$$H(X \cap Y) = \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j)$$

A.5.2.3 Conditional Entropy

Is the average over Y of the conditional entropy of X given $Y = y_i$ and measures the uncertainty that remains about X once Y is known.

For a particular value of $Y = y_i$ we can write

$$H(X | Y = y_i) = -\sum_j p(x_j | y = y_i) \log p(x_j | y = y_i)$$

as we know Y we can write a weighted average over all Y as

$$H(X | Y) = \sum_i p(y_i) \sum_j p(x_j | y_i) \log p(x_j | y_i) = \sum_i \sum_j p(x_j, y_i) \log p(x_j | y_i)$$

A.5.2.4 Some properties of Entropy

$$H(X) + H(Y) \geq H(X \cap Y)$$

$$H(X \cap Y) = H(X) + H(X | Y)$$

A.5.2.5 Mutual Information

$$I(X \cap Y) = H(X) + H(Y) - H(X \cap Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$$

A.5.2.6 Entropic distance

Given two probability distributions $p(X)$ and $q(X)$ then the entropic distance between them is defined as the difference between their joint-entropy and their mutual-information

$$D(p || q) = H(p \cap q) - I(p \cap q)$$

It satisfies the usual requirements of a metric:

$$D(p || q) \geq 0, D(p || p) = 0, D(p || q) = D(q || p), D(p || z) \leq D(p || q) + D(q || z)$$

A.5.2.7 Kullback-Leibler Distance

Also known as the relative entropy

$$D(p || q) = \sum_j p(x_j) \log \frac{p(x_j)}{q(x_j)}$$

A.5.3 Maximum Entropy Methods

In section A.5.2.1 we saw a very simple example of how maximising the entropy yielded a probability distribution with the greatest uncertainty. Given the available information about the distribution this was the best that could be achieved.

When modelling a random process we seek the probability distribution which fits the training data most uniformly subject to known constraints.

Consider the conditional probability distribution $P(\textit{fraud} | \textit{MCC})$

Initially, given no sample data to constrain the problem, the ME solution would be the uniform distribution so that the probability of fraud would be the same for each Merchant Category Code (MCC) and would simply be $1/N$ where N is the number of MCCs.

To include the sample data we introduce the idea of a feature function:

A.5.3.1 Feature functions

Given a set of random variables $X_1 \dots X_n$ which characterise a system, such as *MCC*, *amount*, *time-of-day*, ... then a feature function is defined as any function that maps a subset of these variables to $[0,1]$. So we could have for example:

$$f_1(\textit{age}, \textit{sex}) = \begin{cases} 1 & \text{if } (\textit{age} < 25) \cap (\textit{sex} = \textit{male}) \\ 0 & \text{otherwise} \end{cases}$$

or identity mappings like $f_2(\textit{age}) = 1$

In general then, a feature is a binary valued function that abstracts aspects of the underlying data and is generally written $f_j : \mathbf{X} \rightarrow [0,1]$ where $\mathbf{X} \subseteq \{X_1 \dots X_n\}$

A.5.3.2 Constraints and Features

The Maximum Entropy Method (MEM) seeks a probability distribution which maximises the entropy subject to constraints. The constraints are formulated by requiring the expected value of the feature function to be the same as that derived from the training data.

So for a two variable feature we would have $E(f(x, y)) = \sum_{ij} p(x_i, y_j) f(x_i, y_j)$

A.5.3.3 Maximum Entropy Principal

We should select a probability distribution from the set of all probability distributions P which satisfy the constraints specified.

$$p^* = \operatorname{argmax} H(p) \text{ where } p \in P$$

This is a constrained optimisation problem which can be addressed using the method of Lagrange multipliers. While not difficult we do not go through this exercise here.

It can be shown that p^* is always well-defined and must be of the form

$$p^*(x) = \theta \prod_{j=1}^k \alpha_j^{f_j(x)}$$

where θ is a scaling constant, k is the number of constraints and each α corresponds to one feature and can be regarded as a weight for that feature.

It should be noted that the literature often presents a different but equivalent form by

substituting $\alpha_j = e^{\lambda_j}$ we can then write $p^*(x) = \theta \exp\left(\sum_{j=1}^k \lambda_j f_j(x)\right)$

A.5.3.4 Parameter Estimation

Estimation of the parameters α or λ is usually done using the Generalise Iterative Scaling (GIS) algorithm or the Improved Iterative Scaling (IIS) algorithm.

It has been suggested recently by Malouf that variable-metric and conjugate gradient descent methods can be more efficient for estimating parameters.

A.6 Kernel Density Estimation

This is a powerful technique when estimating Probability Density Functions (PDF) for sparse data as is the case for *fraud* sample populations. It is very similar to so called Radial Basis Functions.

It is a non-parametric estimator of a PDF as it makes no assumptions about the form.

A kernel function $K(x)$ is any positive-definite function that satisfies $\int_{-\infty}^{+\infty} K(x)dx = 1$

The estimator for a function $f(x)$ would then be $f(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$

where n is the number of data points
 X is a sample data point
 h is the so called bandwidth

There are several common forms of $K(x)$

Name	Properties	K(x)
Normal	nice and smooth, infinitely differentiable expensive to compute especially in a real-time or near real-time environment (min 5 arithmetic ops + an exponential)	$K(x) = \frac{1}{\sigma_f \sqrt{2\pi}} e^{-\frac{(x-\mu_f)^2}{2\sigma_f^2}}$
Triangular	fast to compute spikey, implies slow convergence	
Epanechnikov	inverted parabola, nice and smooth, except at its boundaries very cheap to compute (few arithmetic ops)	$K(x) = \frac{3}{4w} \left(1 - \frac{x^2}{w^2}\right) \quad x \in [-w, w]$ $0 \quad \text{if } x > w$

	very useful because it is zero outside its width (so it has no tails) good convergence properties <ul style="list-style-type: none"> • 	
--	---	--

Software

R-Project	Was GNU S - A language and environment for statistical computing and graphics	http://www.r-project.org/
Goose	C++ library for statistical computation and presentation	http://www.gnu.org/software/goose/goose.html
WEKA	Java Data Mining toolkit	http://www.cs.waikato.ac.nz/ml/weka/

References

- Brause, R (1999) Credit Card Fraud Detection by Adaptive Neural Data Mining.
- Brause, R (1999) Neural Data Mining for Credit Card Fraud Detection.
- Charniak, E (1991). Bayesian Networks without Tears. *AI Magazine*
- Heckerman, D (1999). A tutorial on learning with Bayesian Networks.
- Malouf, R (2002). A comparison of algorithms for maximum entropy parameter estimation
- Pearl, J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc.
- Siverman, B (1986). Density Estimation for Statistics and Data Analysis, *Monographs on Statistics and Applied Probability*, London, Chapman Hall
- Stephenson, T (2000). An Introduction to Bayesian Network Theory and Usage, *IDAP Research Report 00-03*
- Whittaker, J (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons Ltd, Chichester, UK.
- Wolf, D (1994). *Mutual Information as a Bayesian Measure of Independence*