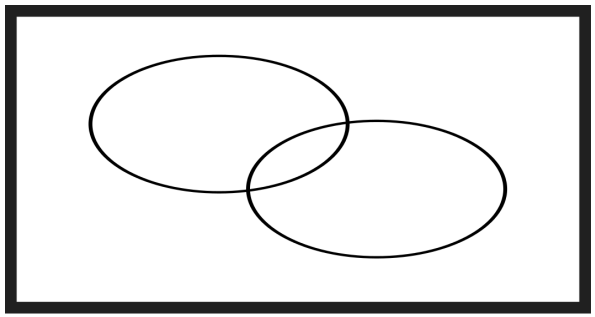
Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Probabilistic Graphical Models: Principles and Applications

Chapter 2: PROBABILITY THEORY

L. Enrique Sucar, INAOE



Outline

1. Introduction
2. Interpretations of Probability
3. Basic Rules
4. Random Variables
   - Two Dimensional Random Variables
5. Information Theory
6. References

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Introduction

- Consider a certain *experiment*, such as throwing a die; this experiment can have different results, we call each result an *outcome*
- The set of all possible outcomes of an experiment is called the sample space, Ω
- An *event* is a set of elements or subset of Ω
- Probability theory has to do with measuring and combining the *degrees of plausibility* of events.

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Interpretations (1)

Classical: probability has to do with equiprobable events; if a certain experiment has N possible outcomes, the probability of each outcome is $1/N$.

Logical: probability is a measure of rational belief; that is, according to the available evidence, a rational person will have a certain belief regarding an event, which will define its probability.

Subjective: probability is a measure of the personal degree of belief in a certain event; this could be measured in terms of a betting factor –the probability of a certain event for an individual is related to how much that person is willing to bet on that event.

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Interpretations (2)

Frequency: probability is a measure of the number of occurrences of an event given a certain experiment, when the number of repetitions of the experiment tends to infinity.

Propensity: probability is a measure of the number of occurrences of an event under repeatable conditions; even if the experiment only occurs once.

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Main approaches

- Objective (classical, frequency, propensity): probabilities exist in the *real* world and can be measured.
- Epistemological (logical, subjective): probabilities have to do with human knowledge, they are measures of belief.
- Both approaches follow the same mathematical axioms defined below; however there are differences in the manner in which probability is applied

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Logical Approach

- Define probabilities in terms of the degree of plausibility of a certain proposition given the available evidence – desiderata:
  - Representation by real numbers.
  - Qualitative correspondence with common sense.
  - Consistency.

  Based on these intuitive principles, we can derive the three axioms of probability:

  1. $P(A)$ is a continuous monotonic function in $[0, 1]$.
  2. $P(A, B \mid C) = P(A \mid C)P(B \mid A, C)$ (product rule).
  3. $P(A \mid B) + P(\neg A \mid B) = 1$ (sum rule).
- Probabilities are always defined with respect to certain context or *background* knowledge: $P(A \mid B)$ (sometimes the background is omitted for simplicity).

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Sum rule

- The probability of the disjunction (logical sum) of two propositions is given by the *sum rule*:
  $P(A + B \mid C) = P(A \mid C) + P(B \mid C) - P(A, B \mid C)$
- If propositions A and B are mutually exclusive given C, we can simplify it to:
  $P(A + B \mid C) = P(A \mid C) + P(B \mid C)$
- Generalized for N mutually exclusive propositions to:
  $P(A_1 + A_2 + \cdots A_N \mid C) = P(A_1 \mid C) + P(A_2 \mid C) + \cdots + P(A_N \mid C)$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Conditional Probabilty

- $P(H \mid B)$ conditioned only on the background B is called *a priori* or prior probability;
- Once we incorporate some additional information D we call it a *posterior* probability, $P(H \mid D, B)$
- The conditional probability can be defined as (for simplicity we omit the background):
  $P(H \mid D) = P(H, D)/P(D)$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Bayes Rule

From the product rule we obtain:

$$P(D, H \mid B) = P(D \mid H, B)P(H \mid B) = P(H \mid D, B)P(D \mid B) \tag{1}$$

From which we obtain:

$$P(H \mid D, B) = \frac{P(H \mid B)P(D \mid H, B)}{P(D \mid B)} \tag{2}$$

This last equation is known as the *Bayes rule*
The term $P(H \mid B)$ is the *prior* and $P(D \mid H, B)$ is the *likelihood*

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Independence

- In some cases the probability of H is not influenced by the knowledge of D, so it is said that H and D are *independent*, therefore $P(H, D \mid B) = P(H \mid B)$
- The product rule can be simplified to: $P(A, B \mid C) = P(A \mid C)P(B \mid C)$

Conditional Independence

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

- If two propositions are independent given only the background information they are *marginally* independent; however if they are independent given some additional evidence, E, then they are *conditionally* independent: $P(A, D \mid B, E) = P(H \mid B, E)$
- Example: A represents the proposition *watering the garden*, D the *weather forecast* and E *raining*

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Chain Rule

- The probability of a conjunction of N propositions, that is $P(A_1, A_2, \ldots, A_N \mid B)$, is usually called the *joint* probability
- If we generalize the product rule to N propositions we obtain what is known as the *chain* rule: $P(A_1, A_2, \ldots, A_N \mid B) = P(A_1 \mid A_2, A_3, \ldots, A_N, B) P(A_2 \mid A_3, A_4, \ldots, A_N, B) \cdots P(A_N \mid B)$
- Conditional independence relations between the propositions can be used to simplify this product

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Total Probability

- Consider a partition, $B = \{B_1, B_2, ...B_n\}$, on the sample space Ω, such that $\Omega = B_1 \cup B_2 \cup ... \cup B_n$ and $B_i \cap B_j = \emptyset$
- A is equal to the union of its intersections with each event B_i, $A = (B_1 \cap A) \cup (B_2 \cap A) \cup ... \cup (B_n \cap A)$
- Then:

$$P(A) = \sum_i P(A \mid B_i) P(B_i) \tag{3}$$

- Given the total probability rule, we obtain Bayes theorem:

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{\sum_i P(A \mid B_i)P(B_i)} \tag{4}$$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Discrete Random Variables

- Consider a finite set of exhaustive and mutually exclusive propositions
- If we assign a numerical value to each proposition x_i, then X is a *discrete random variable*
- The probabilities for all possible values of X, $P(X)$ is the probability distribution of X

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Probability Distributions

- Uniform: $P(x_i) = 1/N$
- Binomial: assume we have an urn with N colored balls, red and black, of which M are red, so the fraction of red balls is $\pi = M/N$. We draw a ball at random, record its color, and return it to the urn, mixing the balls again (so that, in principle, each draw is independent from the previous one). The probability of getting r red balls in n draws is:

$$P(r \mid n, \pi) = \begin{pmatrix} n \\ r \end{pmatrix} \pi^r (1-\pi)^{n-r}, \tag{5}$$

where $\begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{r!(n-r)!}$.

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Mean and Variance

- The expected value or *expectation* of a discrete random variable is the average of the possible values, weighted according to their probabilities:

$$E(X \mid B) = \sum_{i=1}^{N} P(x_i \mid B) x_i \tag{6}$$

- The *variance* is defined as the expected value of the square of the variable minus its expectation:

$$Var(X \mid B) = \sum_{i=1}^{N} P(x_i \mid B)(x_i - E(X))^2 \tag{7}$$

- The square root of the variance is known as the standard deviation

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Continuous random variables

- If we have a continuous variable X, we can divide it into a set of mutually exclusive and exhaustive intervals, such that $P = (a < X \leq b)$ is a proposition, thus the rules derived so far apply to it
- A *continuous random variable* can be defined in terms of a *probability density function*, $f(X \mid B)$, such that:

$$P(a < X \leq b \mid B) = \int_a^b f(X \mid B)dx \tag{8}$$

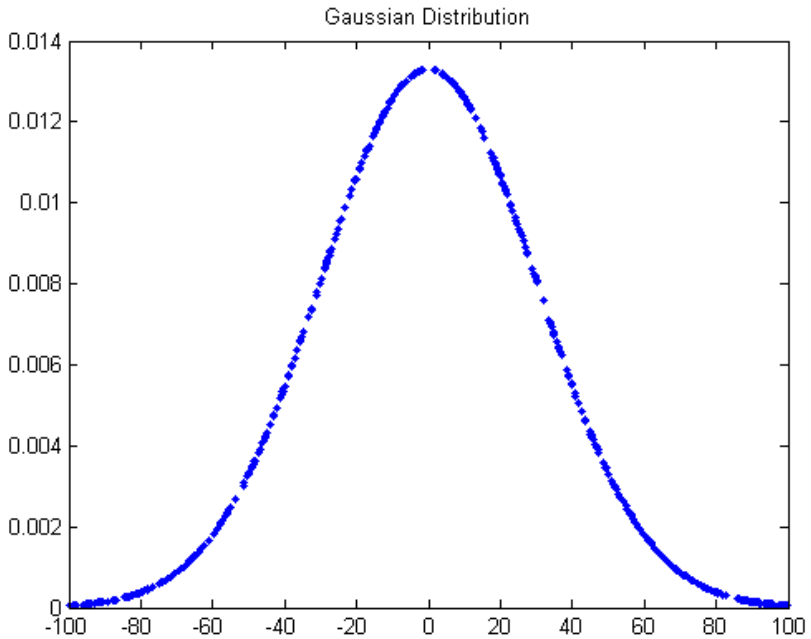
- The probability density function must satisfy $\int_{-\infty}^{\infty} f(X \mid B)dx = 1$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Normal Distribution

A Normal distribution is denoted as $N(\mu, \sigma^2)$, where μ is the *mean* (center) and σ is the *standard deviation* (spread); and it is defined as:

$$f(X \mid B) = \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \tag{9}$$

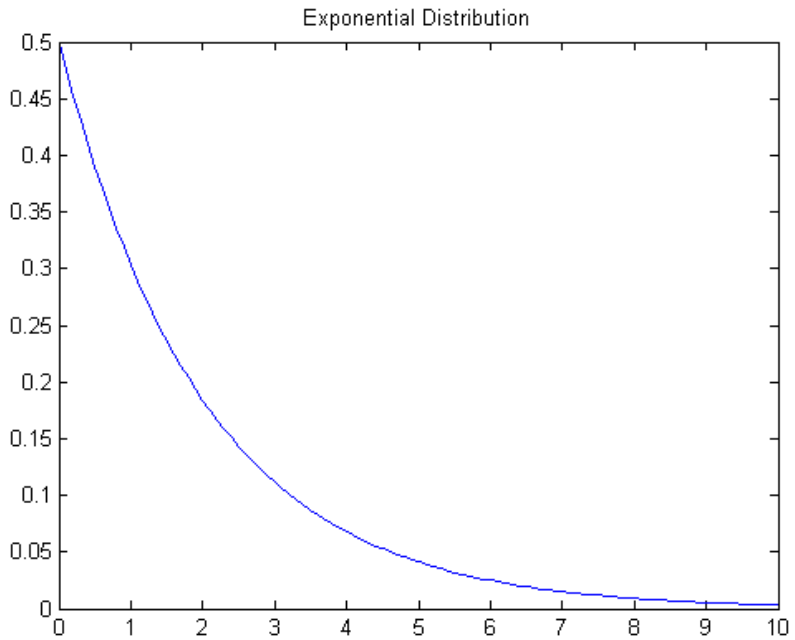


Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Exponential Distribution

The exponential distribution is denoted as $Exp(\beta)$; it has a single parameter $\beta > 0$, and it is defined as:

$$f(X \mid B) = \frac{1}{\beta} e^{-x/\beta}, x > 0 \tag{10}$$



Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Cumulative Distribution

- The cumulative distribution function of a random variable, X, is the probability that $X \leq x$. For a continuous variable, it is defined in terms of the density function as:

$$F(X) = \int_{-\infty}^{x} f(X) \tag{11}$$

- In the case of discrete variables, the cumulative probability, $P(X \leq x)$ is defined as:

$$\mathbf{P}(x) = \sum_{x=-\infty}^{X=x} P(X) \tag{12}$$

Cumulative Distribution Properties

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

- In the interval $[0, 1]$: $0 \leq F(X) \leq 1$
- Non-decreasing: $F(X_1) < F(X_2)$ if $X_1 < X_2$
- Limits: $lim_{x \to -\infty} = 0$ and $lim_{x \to \infty} = 1$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

2D Random Variables

- Given two random variables, X and Y, their joint probability distribution is defined as $P(x, y) = P(X = x \wedge Y = y)$.
- For example, X might represent the number of products completed in one day in product line one, and Y the number of products completed in one day in product line two
- $P(X, Y)$ must follow the axioms of probability, in particular: $0 \leq P(x, y) \leq 1$ and $\sum_x \sum_y P(X, Y) = 1$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

2D Example

- Given two product lines, line one (X) may produce 1, 2 or 3 products per day, and line two (Y), 1 or 2 products, according to the following joint probability distribution:

	X=1	X=2	X=3
Y=1	0.1	0.3	0.3
Y=2	0.2	0.1	0

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Marginal and conditional probabilities

- Given the joint probability distribution, $P(X, Y)$, we can obtain the distribution for each individual random variable:

$$P(x) = \sum_y P(X, Y); P(y) = \sum_x P(X, Y) \tag{13}$$

- From the previous example -
  $P(X = 2) = 0.3 + 0.1 = 0.4$ and
  $P(Y = 1) = 0.1 + 0.3 + 0.3 = 0.7$.
- Conditional probabilities of X given Y and vice-versa:

$$P(x \mid y) = P(x, y)/P(y); P(y \mid x) = P(x, y)/P(x) \tag{14}$$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Independence

- Two random variables, X, Y are independent if their joint probability distribution is equal to the product of their marginal distributions (for all values of X and Y):

$$P(X, Y) = P(X)P(X) \rightarrow \mathit{Independent}(X, Y) \qquad (15)$$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Correlation

- It is a measure of the degree of linear relation between two random variables, X, Y and is defined as:

$$\rho(X, Y) = E\{[X - E(X)][Y - E(Y)]\}/(\sigma_x \sigma_y) \qquad (16)$$

  where $E(X)$ is the expected value of X and σ_x its standard deviation.
- The correlation is in the interval $[-1, 1]$; a positive correlation indicates that as X increases, Y tends to increase; and a negative correlation that as X increases, Y tends to decrease.
- A correlation of zero does not necessarily imply independence

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Information Theory

- Information theory originated in the area of communications, although it is relevant for many different fields
- Assume that we are communicating the occurrence of a certain event. Intuitively we can think that the amount of *information* from communicating an event is inverse to the probability of the event.

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Formalization

- Assume we have a source of information that can send q possible messages, $m_1, m_2, ...m_q$; where each message corresponds to an event with probabilities $P_1, P_2, ...P_q$
- $I(m)$ based on the probability of m - properties:
  - The information ranges from zero to infinity: $I(m) \geq 0$.
  - The information increases as the probability decreases: $I(m_i) > I(m_j)$ if $P(m_i) < P(m_j)$.
  - The information tends to infinity as the probability tends to zero: $I(m) \to \infty$ if $P(m) \to 0$.
  - The information of two messages is equal to the sum of that of the individual messages if these are independent: $I(m_i + m_j) = I(m_i) + I(m_j)$ if m_i independent of m_j.

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Information

- A function that satisfies the previous properties is the logarithm of the inverse of the probability, that is:

$$I(m_k) = log(1/P(m_k)) \tag{17}$$

- It is common to use base two logarithms, so the information is measured in "bits":

$$I(m_k) = log_2(1/P(m_k)) \tag{18}$$

- For example, if we assume that the probability of the message m_r "raining in Puebla" is $P(m_r) = 0.25$, the corresponding information is $I(m_r) = log_2(1/0.25) = 2$

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Entropy

- Given the definition of expected value, the average information of q message or entropy is defined as:

$$H(m) = E(I(m)) = \sum_{i=1}^{i=q} P(m_i) log_2(1/P(m_i)) \qquad (19)$$

- This can be interpreted as that on average H bits of information will be sent

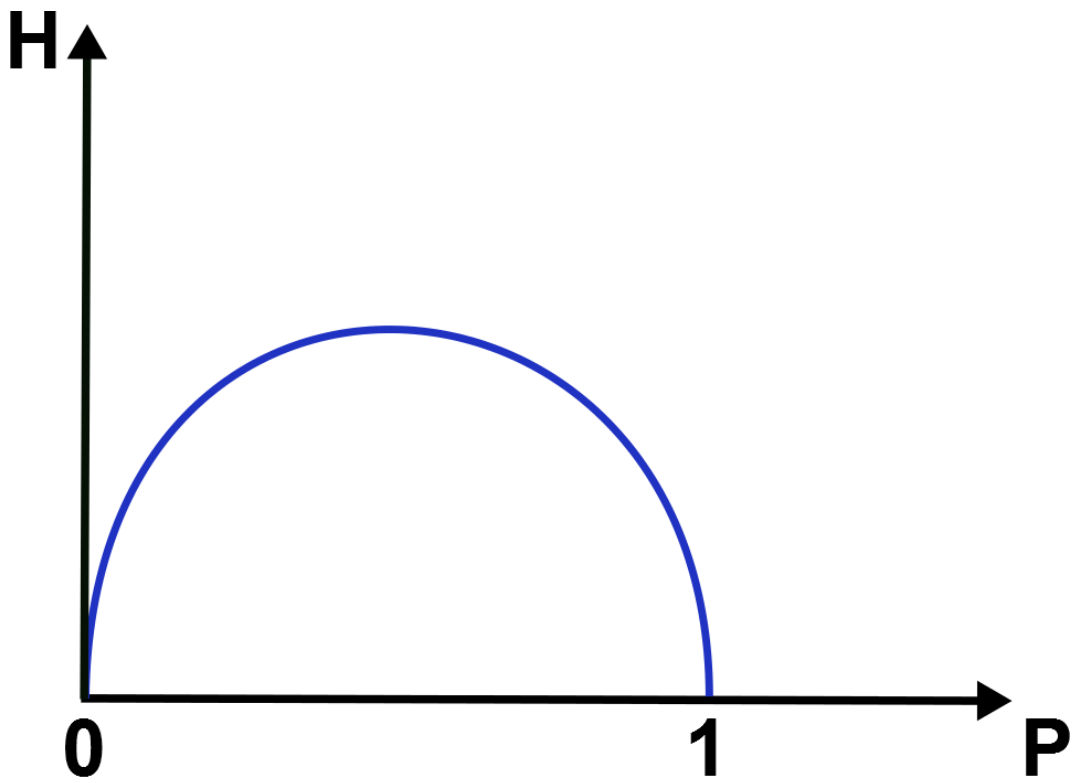
Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Max and Min Entropy

- When will H have its maximum and minimum values?
- Consider a binary source such that there are only two messages, m_1 and m_2; with $P(m_1) = p_1$ and $P(m_2) = p_2$. Given that there are only two possible messages, $p_2 = 1 - p_1$, so H only depends on one parameter, p_1 (or just p)
- For which values of p is H maximum and minimum?

Entropy of a Binary Source

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References



Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Conditional and Cross Enropy

- *Conditional entropy*:

$$H(X \mid y) = \sum_{i=1}^{i=q} P(X_i \mid y) log_2[1/P(X_i \mid y)] \tag{20}$$

- *Cross entropy*:

$$H(X, Y) = \sum_{X} \sum_{Y} P(X, Y) log_2[P(X, Y)/P(X)P(Y)] \tag{21}$$

- The cross entropy provides a measure of the mutual information (dependency) between two random variables

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Additional Reading – Philosophical aspects

Gillies, D.: Philosophical Theories of Probability. Routledge, London (2000)

Sucar, L.E, Gillies, D.F., Gillies, D.A: Objective Probabilities in Expert Systems. Artificial Intelligence **61**, 187–208 (1993)

Introduction
Interpretations of Probability
Basic Rules
Random Variables
Two Dimensional Random Variables
Information Theory
References

Additional Reading – Probability, Statistics, Information theory

Jaynes, E. T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK (2003)

MacKay, D. J.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge, UK (2004)

Wasserman, L.: All of Statistcs: A Concise Course in Statistical Inference. Springer-Verlag, New York (2004)