CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

Administrative

- Any questions

NC STATE UNIVERSITY

Key points from 8/25

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

NC STATE UNIVERSITY

Today

- Introduction to Probability

Probability

- Probability is the mathematical language for quantifying uncertainty or randomness
- Given a random experiment, a probability measure is a population quantity that summarizes the randomness

Statistical Experiment

- All statistical experiments have following common things
  - Has more than one outcome
  - Each possible outcome can be specified in advance
  - Outcome depends on chance
- E.g., coin toss, rolling a die

NC STATE UNIVERSITY

Sample Space and Events

- A sample space Ω, is the set of all possible outcomes of an experiment
- Points ω in Ω are called sample outcomes or realizations
- Events are subsets of Ω

NC STATE UNIVERSITY

Example (1)

- If we toss a coin twice, then
  Ω = {HH, HT, TH, TT}
- Event that the first toss is head
  A = {HH, HT}

Example (2)

- Let ω be the outcome of measurement of some physical quantity, say temperature, then
  $\Omega = R = (-\infty,\infty)$
- Event that the measurement is greater than 10, but less than or equal to 20
  A = (10, 20]

Basic Set Operations

- Given events A, B, …
  - Complement: $A^c = \{\omega \in \Omega;\ \omega \notin A\}$
  - Union: $A \bigcup B = \{\omega \in \Omega;\ \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{both}\}$
  - Intersection: $A \bigcap B = \{\omega \in \Omega;\ \omega \in A \text{ and } \omega \in B\}$
  - Difference: $A - B = \{\omega :\ \omega \in A, \omega \notin B\}$
  - Subset: If every element of A is also contained in B, then A is a subset of B $A \subset B$ or, equivalently, $B \supset A$
  - Disjoint: $A_i \bigcap A_j = \emptyset$
  - A partition of Ω is a sequence of disjoint sets A1, A2, … such that $\bigcup_{i=1}^{\infty} A_i = \Omega$

Summary of Notations

Table 1. Sample space and events.

Ω	sample space
ω	outcome
A	event (subset of Ω)
$\lvert A\rvert$	number of points in A (if A is finite)
A^c	complement of A (not A)
$A \bigcup B$	union (A or B)
$A \bigcap B$ or AB	intersection(A and B)
$A - B$	set difference (points in A that are not in B)
$A \subset B$	set inclusion (A is a subset of or equal to B)
$\emptyset$	null event (always false)
Ω	true event (always true)

Probability

- Probability, P(A), assigns a real number to every event A.
- P is also called a probability distribution or probability measure
- For P to be probability measure, it has to satisfy three axioms

NC STATE UNIVERSITY

Probability

Definition 2.5 *A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A is a* **probability distribution** *or a* **probability measure** *if it satisfies the following three axioms:*

Axiom 1: $\mathbb{P}(A) \geq 0$ *for every* A

Axiom 2: $\mathbb{P}(\Omega) = 1$

Axiom 3: *If* $A_1, A_2, \ldots$ *are disjoint then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

NC STATE UNIVERSITY

Many rules can be derived

- If event A implies occurrence of event B, then P(A) < P(B)

NC STATE UNIVERSITY

Many rules can be derived

- For any two events A and B, the probability that at least one occurs is sum of their probabilities minus their intersection

8/27/15 © Raju Vatsavai CSC-591. 15

NC STATE UNIVERSITY

Group Exercise

- On two tosses of a coin, what is the probability of head on 1st toss or head on 2nd toss (assume all outcomes are equally likely).

8/27/15 © Raju Vatsavai CSC-591. 16

Finite Sample Space

- In general, if Ω is finite, and if each outcome is equally likely, then $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$
- P(A) is also called uniform distribution
- To compute probabilities, we need to count number of points in A
- Methods for counting points are called combinatorial methods

8/27/15 © Raju Vatsavai CSC-591. 17

Counting

- Given n objects, number of ways of ordering these objects is n! = n(n-1)(n-2) … (3)(2)(1)
- For convenience, we denote 0! = 1.
- Let us also define $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Read as “n choose k”, which is the number of distinct ways of choosing k objects from n

8/27/15 © Raju Vatsavai CSC-591. 18

Independent Events

- If we flip a fair coin twice, probability of two heads = ½ x ½ = ¼
- Multiplying because we are assuming two events are independent

Definition *Two events A and B are* **independent** *if*

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$$

and we write $A \amalg B$. A set of events $\{A_i : i \in I\}$ is independent if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

for every finite subset J of I.

Conditional Probability

Definition *If $\mathbb{P}(B) > 0$ then the* **conditional probability** *of A given B is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

- Think P(A|B) as the fraction of times A occurs among those in which B occurs
- Avoid confusion, in general P(A|B) != P(B|A)

Bayes Theorem

Theorem (The Law of Total Probability.) *Let $A_1, \ldots, A_k$ be a partition of Ω. Then, for any event B, $\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$.*

Theorem (Bayes' Theorem.) *Let $A_1, \ldots, A_k$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i. If $\mathbb{P}(B) > 0$ then, for each $i = 1, \ldots, k$,*

$$\boxed{\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.}$$

Example

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

NC STATE UNIVERSITY

Random Variable

- Statistics and DM are concerned with data, so how do we link sample space and events to data?
- A random variable is a numerical outcome of an experiment

Definition *A* **random variable** *is a mapping* $X : \Omega \to \mathbb{R}$ *that assigns a real number* $X(\omega)$ *to each outcome* ω.

8/27/15 © Raju Vatsavai CSC-591. 23

NC STATE UNIVERSITY

Random Variable

- A random variable is a numerical outcome of an experiment
- From previous lecture, numerical data can be either:
  – Discrete or Continuous
- Discrete: you can count (e.g., #of web hits)
  – We can assign probability to every value they can take
- Continuous: real numbers
  – We can assign probability to ranges

8/27/15 © Raju Vatsavai CSC-591. 24

NC STATE UNIVERSITY

Examples

- Simple examples of discrete r.v.
  - The outcome of the flip of the coin
  - The outcome from the roll of a die
- Complex r.v.
  - Number of vehicles on a road network in a given day (discrete but no upper bound)
  - We can use statistical distribution (e.g., Poisson)

NC STATE UNIVERSITY

Examples

- Continuous r.v.; i.e., data can take uncountably infinitely values
  - BMI
  - Satellite imagery
  - Temperature
- We use probability distribution (e.g. Gaussian) to assign probablities

NC STATE UNIVERSITY

Probability Mass Function (PMF)

- A probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.

Definition 3.9 X *is* **discrete** *if it takes countably many values*

$$\{x_1, x_2, \ldots\}.$$

We define the **probability function** *or* **probability mass function** *for* X *by*

$$f_X(x) = \mathbb{P}(X = x).$$

8/27/15 © Raju Vatsavai CSC-591. 27

NC STATE UNIVERSITY

Bernoulli distribution

- Bernoulli distribution is the probability distribution of a random variable which takes value 1 with success probability p and value 0 with failure probability q=1-p.
- Coin toss (H=1, T=0)
- PMF for this distribution can be written as

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

- Can also be written as

$$f(k;p) = p^k(1-p)^{1-k} \quad \text{for } k \in \{0,1\}.$$

8/27/15 © Raju Vatsavai CSC-591. 28

R -- Introduction

- Installation
  - https://www.r-project.org/
- Manuals/FAQs
- Coursera

Acknowledgements

- Vipin Kumar (Minnesota)
- Jiawei Han (UIUC)
- Hanspeter Pfister (Harvard)

NC STATE UNIVERSITY

Acknowledgements

- Vipin Kumar (Minnesota)
- Jiawei Han (UIUC)
- Hanspeter Pfister (Harvard)
- Larry Wasserman (CMU)