ADDIS ABEBA SCIENCE AND TECHNOLOGY UNIVERSITY

COLLEGE ELECTRICAL AND MECHANICAL ENGINEERING

DEPARTEMENT OF SOFTWARE ENGINEERING

Probability and Statistics

NAME ERIMIAS SIRAYE

ID ETS 0253/12

SECTION C

SUBMITED TO TEACHER ASHEBIR FEYISA

SUMMISON DATE 19/8/2021

The Use of Statistics and Probability to Engineering

Specially to Software Engineering

Introduction

Statistics plays an intrinsic role in software engineering and vice versa. Statistics is used in used for data mining speech recognition vision and image analysis and other numerous applications.

A statical background is essential for understanding algorithms and statical properties that form the backbone of software engineering.

Software engineers are often tasked with acquisition/ cleaning, retrieval, mining and reporting of data and all this can be accomplished by the applications of statistics.

Also, when we come to specific engineering part software engineering statistics have much importance. The experimental evaluation of the methods and concepts covered in software engineering has been increasingly valued. This value indicates the constant search for new forms of assessment and validation of the results obtained in Software Engineering research.

Results are validated in studies through evaluations, which in turn become increasingly stringent. As an alternative to aid in the verification of the results, that is, whether they are positive or negative, we suggest the use of statistical methods.

Data Mining

Data Mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules and is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets.

In order to analyze data and learn from the data knowledge of statistics is essential. Using principles of statistics, we can get and generalize the data in to information using the samples.

- Most Data Mining techniques are statistical exploratory data analysis tools.
- Complete understanding of the data and its collection methods are particularly important.
- Database sampling or cluster analysis help in reducing the dimension and size of massive data sets.
- Statistical data visualization tools are used to aid in the analysis of massive data sets.

Process Control

Using statistics in process control also known as statistical process control (SPC) is a method of quality control which employs statistical methods to monitor and control a process. This helps ensure the process operates efficiently, producing more specification-conforming product with less waste (rework or scrap).

SPC can be applied to any process where the "conforming product" (product meeting specifications) output can be measured. Key tools used in SPC include run charts, control charts, a focus on continuous improvement, and the design of experiments. An example of a process where SPC is applied is manufacturing lines.

In the process of software development, the application of SPC involves three main phases of activity:

- Understanding the process and the specification limits.
- Eliminating assignable (special) sources of variation, so that the process is stable.
- Monitoring the ongoing production process, assisted by the use of control charts, to detect significant changes of mean or variation.

Internet Security

Intrusion detection systems DDoS prevention mechanisms and many other areas of security use probability and statistics in detail. In an intrusion detection system, an interesting approach is to group the system calls in pairs, triples etc. and then observe their behavior when the application encounters an attack.

It is also applicable with high-speed networks where specialized network processors capture enormous amounts of traffic. Theoretically such analysis can reveal ongoing DDoS attacks and be used to deploy dynamic filtering.

It also includes:

- Security Incident and Event Management (SIEM) & Security Operations (SOC) - analyzing and correlating logs and alerts
- WAF (web application firewall) - since configuring these systems are difficult, most modern commercial WAFs have incorporated automatic learning mode, based on behavioral analysis
- "risk-based" authentication systems.

Machine learning

Statistics and machine learning are two very closely related fields. In fact, the line between the two can be very fuzzy at times. Nevertheless, there are methods that clearly belong to the field of statistics that are not only useful, but invaluable when working on a machine learning project.

It would be fair to say that statistical methods are required to effectively work through a machine learning predictive modeling project.

Some of the uses of statistics are

- Problem Framing
- Data Understanding
- Data Cleaning
- Data Selection
- Data Preparation

Statistical quality control

The use of statistical methods in the monitoring and maintaining of the quality of products and services. One method, referred to as acceptance sampling, can be used when a decision must be made to accept or reject a group of parts or items based on the quality found in a sample.

A second method, referred to as statistical process control, uses graphical displays known as control charts to determine whether a process should be continued or should be adjusted to achieve the desired quality.

Signal Processing

Statistical signal processing (applying principles of statistics in signal processing) is an approach to signal processing which treats signals as stochastic processes (a collection of random variables), utilizing their statistical properties to perform signal processing tasks. Statistical techniques are widely used in signal processing applications.

For example, one can model the probability distribution of noise incurred when photographing an image, and construct techniques based on this model to reduce the noise in the resulting image. This is typically accomplished using either as a Bayesian or a frequentist model.

Reliability engineering

Reliability engineering is a sub-discipline of systems engineering that emphasizes the ability of equipment to function without failure. Reliability describes the ability of a system or component to function under stated conditions for a specified period of time.

The Reliability function is theoretically defined as the **probability** of success at time t, which is denoted R(t). This probability is estimated from detailed (physics of failure) analysis, previous data sets or through reliability testing and reliability modelling. Availability, testability, maintainability and maintenance are often defined as a part of "reliability engineering" in reliability programs.

Big data

Statistics is fundamental to ensuring meaningful, accurate information is extracted from Big Data. The following issues are crucial and are only exacerbated by Big Data:

- Data quality and missing data o Observational nature of data, so that causal questions such as the comparison of interventions may be subject to confounding.
- Quantification of the uncertainty of predictions, forecasts and models

The scientific discipline of statistics brings sophisticated techniques and models to bear on these issues. Statisticians help translate the scientific question into a statistical question, which includes carefully describing data structure; the underlying system that generated the data (the model); and what we are trying to assess (the parameter or parameters we wish to estimate) or predict.

• Big Data will often not be served well by “off the shelf” methods or black box computational tools that work in low-dimensional and less complicated settings, and therefore require tailored statistical methods.

• Statisticians are skillful at assessing and correcting for bias; measuring uncertainty; designing studies and sampling strategies; assessing the quality of data; enumerating limitations of studies; dealing with issues such as missing data and other sources of non-sampling error; developing models for the analysis of complex data structures; creating methods for causal inference and comparative effectiveness; eliminating redundant and uninformative variables; combining information from multiple sources; and determining effective data visualization techniques.

Software testing

Any large, complex, and expensive process can perform most activities in myriad ways, as can software development, which can significantly improve profitability through the use of statistical science.

Statistics provide a structure for collecting data and turning it into information, which can improve decision-making in uncertain circumstances. The term "statistical test" commonly used in software engineering literature strictly refers to randomly generated test cases. However, the term should be understood as a comprehensive application of statistical science, including operations research methods, to solve problems caused by testing industrial software.

Statistical tests can collect empirical data effectively, eliminating the uncertainty of software-intensive system behavior and supporting economic decisions about further testing, deployment, maintenance, and evolution. The operational usage model is a form of formalism, which can apply many statistical principles to software testing and form the effective evidence base to support decision making.

Role Of Statistics in Neural Network

Traditionally, the term neural network had been used to refer to a network or circuit of biological neurons; the modern term refers to artificial neural networks which are composed of artificial neurons or nodes. Thus, the term has two distinct usages:

1. Biological neural networks are made up of real biological neurons that are connected or functionally related in the peripheral nervous system or the central nervous system. In the field of neuroscience, they are often identified as groups of neurons that perform a specific physiological function in laboratory analysis.

2. Artificial neural networks are made up of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons).

Artificial neural networks may either be used to gain an understanding of biological neural networks, or for solving artificial intelligence problems without necessarily creating a model of a real biological system. The real, biological nervous system is highly complex and includes some features that may seem superfluous based on an understanding of artificial networks.

Role of Statistics in Simulation

Simulation is used in many contexts, such as simulation of technology for performance optimization, safety engineering, testing, training, education, and video games. Training simulators include flight simulators for training aircraft pilots.

Simulation is also used for scientific modeling of natural systems or human systems in order to gain insight into their functioning. Simulation can be used to show the eventual real effects of alternative conditions and courses of action. Simulation is also used when the real system cannot be engaged, because it may not be accessible, or it may be dangerous or unacceptable to engage, or it is being designed but not yet built, or it may simply not exist. There is a major role of statistics to create a discrete –event simulation and queuing system in simulation.

The purpose of creating a discrete-event simulation is to improve understanding of the underlying system being modeled or to use simulation results to help make decisions about the underlying system. Numerical results gathered during simulation can be important tools.

Software Reliability

Is another very important quality factor defined as probability of failure free operation of a computer program in a specified environment for a specified time. For example, a program X can be estimated to have a reliability of 0.96 over 8 elapsed hours.

Software reliability can be measured, directed and estimated using historical and development data. The key to this measurement is the meaning of term failure. Failure is defined as non-conformance to software requirements.

Software safety

Is a software SQA activity that focuses on identification of potential hazards that may affect software negatively and cause and entire system to fail. Modeling and analysis process is conducted as part of software and hazards are identified and categorized by criticality and risk.

Example Let us assume that the following hazards are associated with a computer-based cruise control for an automobile:

- Causes uncontrolled acceleration that can’t be stopped
- Doesn’t respond to depression of brake pedal.

- Doesn’t engage when switch is activated.
- Slowly loses or gains speed

Statistics in Network trafficking model

Advancements in computer technology and storage capabilities have allowed network engineers to collect very large amounts of data obtained from computer networks to address a number of engineering tasks, including network provisioning, providing high quality of-service to users in advanced applications such as Internet telephony and television and online gaming, configuring network protocols, fault diagnosis, traffic forecasting, just to name a few.

Different types of network data can be collected that differ in their granularity, accuracy, volume and delay. We start by providing a brief high-level description of computer network operations. Networks consist of nodes (routers and switches) connected by physical links (optical or copper wires).

Data, in the form of packets, are transmitted over the network from one node (called a source) to another node (called the destination) on predetermined paths, or routes. A stream of packets from a particular source to a particular destination defines a flow. In many applications, flows are examined at a more granular level, such as the protocol level (e.g. http, ftp) .

Data on flow-level traffic can be obtained from NetFlow or similar (tcp dump) technologies that provide very detailed about the flow, including the application, packet and byte volumes, transmission protocol and delays, etc. In principle, such data can be collected for all packets of all flows. However, this is impractical in today’s high-speed networks, which has led to the implementation of sampling strategies for data collection purposes.