Chris Ricchi

12-1-2023

Professor Hoy

**UCI Machine Learning: Breast Cancer Diagnostic Dataset**

**Abstract:** In the world of statistics and computer science, understanding how the topics we cover can be applied in real life is pivotal for a comprehensive understanding of the subject matter. This paper aims to delve into the Breast Cancer Wisconsin Dataset, a dataset from the UCL Machine Learning Repository. This dataset predicts whether a tumor is malignant or benign or benign by analyzing various features. By exploring the intricacies of the dataset in terms of probability distributions, we can showcase the tangible applications of the methods in the context of diagnosing breast cancer. Using statistical formulas in reference to the data can demonstrate something as important as how our approach of understanding and treating diseases like breast cancer can be improved upon. In essence, this exploration not only reinforces the practical relevance of statistics but also highlights their potential to significantly enhance our understanding of treatment approaches in real world problems such as cancer.

**Permutations**

In a genetic study, a team is examining the possible arrangements of a specific set of 6 gene markers within the Breast Cancer Wisconsin Dataset. The order of these gene markers is crucial for understanding their impact on cancer treatment development. How many different ways can the 6 gene markers be arranged in the dataset?

*Solution*

The number of permutations of n distinct items is given by n!

**Therefore, there are 720 different ways the 6 gene markers can be arranged in the Breast Cancer Wisconsin Dataset.**

**Results**

The permutations problem delved into the various ways of arranging a set of 6 gene markets within the dataset. With 720 distinct permutations possible, this result sheds light on the multitude of arrangements that can occur, each representing a single sequence of the gene markers. In the realm of cancer study, the significance of these permutations lie in understanding how the order of these markers may influence the patterns. Permutations provide insight into the importance of the sequence, as they have to take order into consideration. This information proves valuable to researchers studying the specifics of genetic interactions within the specific dataset as it shows the potential variations in gene markers arrangements and their impact on breast cancer.

**Combinations**

In a clinical trial using the Breast Cancer Wisconsin Dataset, a team is selecting a group of 4 patients to study the expression patterns of specific gene markers. The order in which patients are selected is not important; only the composition of the group markers. How many different groups of 4 patients can be selected from a pool of 8 patients in the Breast Cancer Wisconsin Dataset?

*Solution*

The number of combinations of n distinct items taken k at a time is given by

**Therefore, there are 70 different groups of 4 patients that can be selected from the pool of 8 patients in the clinical trial using the Breast Cancer Wisconsin Dataset.**

**Results**

The combinations problem explored the different ways of selecting a group of 4 patients from a pool of 8 in the dataset. The calculated result, 70, signifies the number of unique combinations of patients that can be chosen for a clinical study examining the expression patterns of specific gene markers. Unlike permutations, combinations do not consider the order of selection, demonstrating the importance of the group’s composition as opposed to its sequence. In the context of breast cancer research, this insight is crucial for designing clinical trials that examine diverse patient groups. The number of combinations can provide researchers with a more clear understanding of the possible patient subsets available for study, ensuring unbiased analysis within the dataset.

**Conditional Probability**

In the Breast Cancer Wisconsin Dataset, a diagnostic test for breast cancer is administered. The test results are classified into two categories: positive for patients who have breast cancer, and negative for patients without breast cancer. Additionally, the dataset provides information about the actual diagnosis, where malignant indicates the presence of cancer and benign indicates the absence of cancer. Given that a patient has tested positive in the diagnostic test, what is the conditional probability that the patient’s actual diagnosis is malignant?

*Solution*

equals the number of cases where the actual diagnosis is malignant and the test result is positive.

equals the number of cases where the test result is positive

**= 0.7067**

**Therefore, the conditional probability that a patient’s actual diagnosis is malignant given a positive test result is 0.7067, or 70.67%.**

**Results**

The conditional probability problem investigated the likelihood that a patient’s actual diagnosis is malignant given a positive test result in the dataset. With 212 instances identified as both malignant and positive out of a total 300 positive test results, the calculated conditional probability is approximately 70.67%. This result highlights the considerable probability that a positive test result genuinely indicates the presence of a malignant tumor. In the context of cancer diagnosis, understanding this probability is crucial for healthcare professionals as it shows the test results are very reliable. A high conditional probability signifies a strong association between a positive test result and the actual presence of a malignant tumor.

**Multiplicative Law of Probability**

In a breast cancer screening program, two independent diagnostic tests are administered sequentially to patients. Test A has a sensitivity of 85%, meaning it correctly identifies 85% of patients with breast cancer, and a specificity of 90%, meaning it correctly identifies 90% of patients without breast cancer. Test B is administered only to those who test positive on Test A. Test B has a sensitivity of 80% and a specificity of 95%. What is the probability that a patient tested positive on both Test A and Test B actually has breast cancer?

*Solution*

= **0.233**

**Therefore, the probability that a patient tested positive on both Test A and Test B and actually has breast cancer is 0.233, or 23.3%.**

**Results**

The application of the multiplicative law of probability in the cancer screening scenario yielded a calculated probability of approximately 23.3% that a patient who tested positive on both tests actually has breast cancer. This result takes into account the individual sensitivities of Test A (85%) and Test B (80%), as well as the prevalence of breast cancer in the screened population. The outcomes demonstrate the importance of considering the nature of diagnostic tests and their sensitivities in determining the probability of a true positive result. In the context of screening for cancer, understanding the probabilities aids healthcare professionals in interpreting test outcomes and making informed decisions regarding further potential treatment plans.

**Theorem of Total Probability**

In a breast cancer clinic, diagnostic tests are administered using two different technologies, Machine A and Machine B. Machine A has a sensitivity of 80% and a specificity of 95%, while Machine B has a sensitivity of 90% and a specificity of 85%. The clinic uses Machine A 60% of the time and Machine B 40% of the time. What is the overall probability that a randomly selected patient receives a positive breast cancer diagnosis?

*Solution*

given that:

= **0.84**

**Therefore, the overall probability that a randomly selected patient receives a positive breast cancer diagnosis is 0.84, or 84%.**

**Results**

The application of the Theorem of Total Probability in the breast cancer clinic problem provides an overall probability of 84% that a randomly selected patient receives a positive breast cancer diagnosis. This result takes into account the probabilities of positive diagnosis given the use Machine A (60% of the time) and Machine B (40% of the time), as well as the individual sensitivity of each machine. The weighted sum reflects a fair diagnostic strategy as it incorporates the likelihood of using each machine into the final probability. A probability result such as this is of high importance for healthcare professionals as it can assist them in making decisions about resource allocation. The 84% total probability serves as a valuable piece of information in the clinic’s overall success in detecting breast cancer in specific scenarios.

**Baye’s Theorem**

In a breast cancer screening program, a diagnostic test is administered to detect a specific biomarker associated with breast cancer. The sensitivity of the test is 90%, meaning it correctly identifies 90% of patients with breast cancer, and the specificity is 85%, meaning it correctly identifies 85% of patients without breast cancer. The prevalence of breast cancer in a specific screened population is 15%. If a randomly selected patient receives a positive test result, what is the probability that the patient actually has breast cancer?

*Solution*

whereas

We can use the Law of Total Probability to find by doing the following:

=

= **0.5143**

**Therefore, the probability that a patient actually has breast cancer given a positive test result is approximately 0.5143 or 51.43%.**

**Results**

The application of Baye’s Theorem provided a deep understanding of the probability that a patient actually has breast cancer given a positive test result. With a test sensitivity of 90%, specificity of 85%, and a prevalence of breast cancer in the screened population at 15%, the calculated probability was approximately 51.43%. This result indicates that even with a positive test result, there is not an absolute certainty of that patient having breast cancer. The balance between test accuracy and prevalence plays a crucial role in the way professionals interpret screening results. The calculated probability is very important as it can guide healthcare professionals in making better decisions about follow-up diagnostics and potential treatment interventions. It emphasizes the importance of considering both the accuracy of the test and the prevalence of breast cancer in the given population.

**Chebyshev’s Theorem**

In a study on breast cancer tumor sizes, the mean tumor size is 8 centimeters, and the standard deviation is 2 centimeters. Chebyshev’s theorem states that for any distribution, at least of the data must fall within *k* standard deviations from the mean. What proportion of breast cancer tumors must be within 4 centimeters and 12 centimeters of the mean according to Chebyshev’s theorem?

*Solution*

whereas mean () = 8 centimeters and standard deviation () = 2 centimeters

and

**Therefore, according to Chebyshev’s Theorem, at least 75% of breast cancer tumors must fall within 4 and 12 centimeters of the mean tumor size.**

**Results**

The utilization of Chebyshev’s Theorem in the examination of breast cancer tumor sizes yielded great insight into the dispersion of data around the mean. The theorem predicts that at least 75% of breast cancer tumors must fall within the range of 4 to 12 centimeters from the mean tumor size of 8 centimeters. This wide range accounts for the inherent variability in tumor sizes and emphasizes the importance of Chebyshev’s Theorem in providing a minimum guarantee for the proportion of data within a specified distance from the mean. While more specific results may be obtained from a distribution problem, Chebyshev’s Theorem still offers a fair estimate of the data spread and shows itself as a useful tool for healthcare professionals in understanding the potential range of breast cancer tumor sizes.

**Binomial Distribution**

Consider a medical test designed to detect breast cancer using a specific tool. The test has been proven to correctly identify malignant tumors with a success rate of 90%. However, it also produces false positives, incorrectly indicating the presence of cancer in 5% of cases. A group of 8 patients undergoes this breast cancer diagnostic test. What is the probability that exactly 7 out of the 8 patients will be correctly identified as having malignant tumors?

*Solution:*

whereas

= **0.283**

**Therefore, the probability that exactly 7 out of 8 patients will be correctly identified as having malignant tumors is approximately 0.283, or 28.3%.**

**Results**

The binomial distribution problem presented a scenario where a diagnostic test for breast cancer, with a 90% success rate and a 5% false-positive rate, was administered to a group of 8 patients. The specific question addressed the likelihood of exactly 7 out of 8 patients being correctly identified as having malignant tumors. The calculated probability, approximately 28.3%, illustrates that while the test exhibits a high overall success rate, there remains a significant chance of encountering variability in outcomes, even in a relatively small sample size. This result demonstrates the importance of understanding the distribution outcomes in a diagnostic setting. The binomial distribution provides a valuable tool for assessing the probability of different test result scenarios and informs decision-making processes in healthcare, especially where accurate diagnosis is critical for patient outcomes.

**Geometric Distribution**

Researchers have proposed a new diagnostic tool to identify breast cancer malignancy where they are mainly interested in the number of tests required until the first successful detection of malignancy occurs. The probability of a negative result in a single test using this tool is p = 0.1. What is the probability that a patient will need more than 3 tests before the first successful detection of a malignant tumor using the new tool?

*Solution*

whereas

and

In this case, p = 0.1 (probability of a false-negative result)

and k > 3 (more than 3 tests)

**0.271**

**Therefore, the probability that a patient will need more than 3 tests before the first successful detection of malignancy using the new tool is approximately 0.271, or 27.1%.**

**Results**

In the geometric distribution problem, we explored the probability of a patient requiring more than 3 tests before the first successful detection of a malignant tumor using a new diagnostic tool that has a 10% false-negative rate. The calculated probability, approximately 27.1%, provides insight into the variability that is associated with repeated testing. This result shows that there is a considerable chance of patients undergoing multiple tests before achieving a successful detection, showing the uncertainty in the diagnostic process. The geometric distribution’s application here shows how the number of trials required for success can skew the underlying result of the test. This can be used by healthcare professionals in managing patient expectations and optimizing testing protocols to minimize both the number of tests used and the time involved in accurate diagnosis.

**Hypergeometric Distribution**

A medical research team is testing a revolutionary diagnostic tool designed to identify breast cancer markers. The team has developed a panel of 20 unique markers believed to be associated with malignant tumors. The tool randomly selects 8 markers from the panel for each test. In the Breast Cancer Wisconsin Dataset, 40% of tumors are malignant. The team is interested in understanding the probability that in a single test using this tool, exactly 5 out of the 8 selected markers are associated with malignant tumors.

*Solution*

whereas

= **0.285**

**Therefore, the probability that, in a single test using the new diagnostic tool, exactly 5 out of the 8 randomly selected markers are associated with malignant tumors is approximately 0.285, or 28.5%.**

**Results**

The hypergeometric distribution problem investigated the probability of exactly 5 out of 8 markers selected in a single test using a new diagnostic tool being associated with malignant tumors. The calculated probability, approximately 28.5%, offers insight into the distribution of successful outcomes in the context of real-world breast cancer marker identification. The results demonstrate the significance of understanding the probability of obtaining a specific number of markers, which is crucial for the tool’s overall effectiveness. The hypergeometric distribution provides a nuanced perspective on the randomness of marker selection, showing the complexities of the diagnostic process. These results can guide healthcare professionals in overall optimizing the marker panels and further refining diagnostic strategies for efficiency.

**Poisson Distribution**

In a study focusing on breast cancer occurrences, it has been observed that, on average, 2 new cases of breast cancer are diagnosed in a particular region every day. What is the probability that exactly 4 new cases of breast cancer will be diagnosed in this region on a given day?

*Solution*

whereas

(2 in this case)

(4 in this case)

= **0.0902**

**Therefore, the probability that exactly 4 new cases of breast cancer will be diagnosed in this region on a given day is approximately 0.0902, or 9.02%.**

**Results**The Poisson distribution problem explored the probability of exactly 4 new cases of breast cancer being diagnosed in a specific region on any given day, where the average rate of diagnoses is 2 per day. The calculated probability, approximately 9.02%, provides meaningful insights into the likelihood of observing a specific number of diagnoses in a region with a known average occurrence. This result reflects the randomness and variability of daily diagnoses, with the Poisson distribution serving as a fair model for such scenarios. Understanding the probability of certain events occurring is crucial for healthcare planning, resource allocation, and further research. In a region with a higher known average rate of diagnosis, the Poisson distribution could prove very helpful in determining whether certain spikes of diagnosis results are plausible or not. This could allow researchers to anticipate and respond effectively to the dynamic nature of disease patterns.

**Negative Binomial Distribution**

In a clinical trial for a new breast cancer treatment plan, patients are given a series of doses until they experience a specific adverse reaction. On average, patients in this trial show adverse reactions after an average of 3 doses. The negative binomial distribution is suitable for modeling the number of doses required until the occurrence of the adverse reaction. What is the probability that a patient will experience an adverse reaction after the 5th dose?

*Solution*

whereas

In this example problem,

= **0.197**

**Therefore, the probability that a patient will experience the adverse reaction after the 5th dose in a clinical trial is approximately 0.197, or 19.7%.**

**Results**

The negative binomial distribution problem investigated the probability of a patient experiencing an adverse reaction after the 5th dose in a clinical trial for a new breast cancer treatment. With an average adverse reaction occurring after 3 doses, the calculated probability, approximately 19.7%, sheds light on the unpredictability in the treatment process. This result signifies that, on average, patients may require more doses than anticipated before encountering the adverse reaction. The negative binomial distribution proves valuable for modeling the distribution of the number of doses until something happens, offering healthcare professionals information into the variability of patient reactions. Understanding the results can help manage patient expectations, optimizing treatment protocols and ensuring patient safety in the development of new medical treatment plans.

**Joint Probability Density Function**

In a study on breast cancer, researchers are investigating two continuous variables: tumor size (X) and age at diagnosis (Y). The joint probability density function is given by:

Determine the probability that a randomly selected patient has a tumor size (X) between 5 and 8 centimeters and was diagnosed at an age (Y) between 40 and 50 years old.

*Solution*

Integrate with respect to x:

Integrate with respect to y:

= **0.059**

**Therefore, the probability that a randomly selected patient has a tumor size between 5 and 8 centimeters and was diagnosed at an age between 40 and 50 years is approximately 5.9%.**

**Results**

The exploration of the joint probability density function for breast cancer, considering tumor size (X) and age at diagnosis (Y), revealed very fascinating results. The probability that a randomly selected patient falls within the specified range of having a tumor size between 5 and 8 centimeters and being diagnosed at an age between 40 and 50 years old is approximately 5.9%. This result, found by integration of the joint probability density function, signifies the likelihood of observing patients with specific characteristics in a continuous space. The connection between tumor size and age contributes to our understanding of varied patterns within the breast cancer patient population. Analyses like this are crucial for healthcare professionals to comprehend the joint distribution of multiple variables, as the nature of breast cancer cases is very complex.

**Marginal Probability**

In the joint probability density function for breast cancer tumor size (X) and age at diagnosis (Y):

Determine the marginal probability density functions and for tumor size (X) and age at diagnosis (Y).

*Solution*

and

=

Marginal Probability Density Function for Tumor Size (X):

Marginal Probability Density Function for Age at Diagnosis (Y):

**Results**

The joint probability density function for breast cancer, with tumor size (X) and age at diagnosis (Y), shows the derivation of the marginal probability density function. This highlights the individual distributions of these variables. The marginal probability density function for tumor size takes the form , which shows the probability of observing a particular tumor size independently and shows insight into the distribution of potential tumor size variation in breast cancer patients. Similarly, the marginal probability density function for age at diagnosis takes the form , which shows the probability of a specific age at diagnosis occuring, isolated from the influence of tumor size. These expressions offer a comprehensive understanding of the likelihood of each specific value happening, contributing to the broader understanding of breast cancer characteristics.

**Conditional Probability**

Given the joint probability density function for breast cancer tumor size (X) and age at diagnosis (Y):

Determine the conditional probability density function , representing the probability distribution of tumor size (X) given a specific age at diagnosis (Y).

*Solution*

where is the marginal density function for age at diagnosis.

which simplifies to =

**Therefore, the conditional probability density function is given by the function:**

which represents the probability distribution of tumor size (X) given a specific age (Y).

**Results**

The conditional probability in the breast cancer problems shows the relationship between tumor size (X) and age at diagnosis (Y). The conditional probability density function derived as shows the distribution of tumor size given a specific age at diagnosis. This expression shows that, conditioned on a known age (Y), the likelihood of observing a particular tumor size (X) follows a conditional joint distribution. This offers insight for healthcare professionals as it shows a relation of tumor size variation with age at diagnosis in terms of breast cancer specifically.

**Overall Findings**

Exploring the UCL Machine Learning Repository’s Breast Cancer Wisconsin Dataset has laid the framework for infinite real-world applications. From the foundational basics of probability distributions such as Binomial, Geometric, Hypergeometric, Poisson, and Negative Binomial, to the more intricate functions involving joint and conditional probability density functions, each problem has contributed to a deeper understanding of the dataset. The mock questions for each statistical concept have demonstrated their real-world use, as we have uncovered a predictable tumor malignancy percentage and evaluated the efficiency of certain diagnostic tools. The Breast Cancer Wisconsin Dataset, demonstrating features such as tumor size, texture, and symmetry, provides an array of information that mirrors the strange nature of breast cancer cases. The dataset’s real-world relevance shows the challenges that healthcare professionals face in diagnosing and understanding the disease every day. The Binomial Distribution revealed the probability of the first occurrence of a specified event, helping understand early-stage diagnoses. The Hypergeometric Distribution addressed the likelihood of obtaining a specific set of features in breast cancer cases, emphasizing the challenges associated with predicting specific patient characteristics. Moving to Chebyshev’s Theorem, we assessed the range of tumor sizes, offering a fair estimate of the proportion of tumors within a specified range and recognizing the inherent variability in breast cancer cases. The application of Conditional Probability showed the interdependence of tumor size and age at diagnosis. The Poisson Distribution allows us to model the distribution of events over time, providing a base to learn the occurrence of rare events like breast cancer. This allows researchers to predict the likeliness of rare but critical outcomes such as specific genetic mutations or treatment responses. The Negative Binomial Distribution explored the variability in the number of events until a specified number of failures occurred. This was relevant in the context of breast cancer recurrence, allowing researchers to model the likelihood of a recurrence of breast cancer, allowing better long-term patient management. Each analysis has further contributed to a deeper understanding of the breast cancer dataset, allowing healthcare professionals to make more informed decisions and advancing the pursuit of curing, predicting, and preventing breast cancer overall. This exercise is more than just a mock-up questionnaire of statistics; it creates a bridge between the concepts and their real-world applications to benefit society.

Works Cited

UCL DATASET: <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>