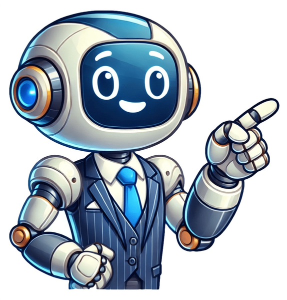

Decision tree testing examples

Statistical tests help researchers evaluate hypotheses or analysis results on their data, determining relationships or differences between variables or groups. However, choosing the right statistical test can be challenging, especially for those without a statistics background. This article provides a decision tree-based guide to navigate this process. Statistical tests are categorized into parametric and non-parametric tests, each with its assumptions and applications. Parametric tests, such as t-tests, ANOVA, and regression analysis, assume data follow a normal distribution. Non-parametric tests, like the Mann-Whitney U test, Kruskal-Wallis test, and Wilcoxon signed-rank test, are used when data do not meet parametric assumptions or have small sample sizes. A decision tree approach is presented to guide choosing the right statistical test based on data characteristics and problem types. Example use cases illustrate how this tool can be applied: - T-test: Comparing average test scores between two high-school classes to evaluate if there's a statistically significant difference. - Paired t-test: Measuring glucose levels before and after treatment in the same group of patients, enabling comparison of means from related groups. - ANOVA: Analyzing exam scores across three classrooms to compare means among independent groups. - Chi-Squared test: Exploring the relationship between gender and product preference to assess if two distributions differ. - Spearman's Rank Correlation: Investigating the relationship between students' performance rankings in math and science. Non-Parametric Tests and Decision Trees: A Guide for Choosing the Right Statistical Tool When dealing with datasets that don't meet the assumptions of a normal distribution, non-parametric tests come into play. These tests are useful for comparing two independent groups of numerical data without requiring a normal distribution. Regression analysis is an example where non-parametric tests can be applied. For instance, predicting sales based on advertising expenditure can benefit from non-parametric methods like the Mann-Whitney U Test or Wilcoxon Signed-Rank Test. The Z-Test is another statistical tool used to compare proportions between two independent samples, determining if the difference in preferences is statistically significant. To choose the right statistical test, it's essential to understand your data and the type of test needed. A decision tree can be a valuable tool in this process. Decision Trees --------------- A decision tree is a supervised machine learning algorithm that works by following a set of if-else conditions to visualize data and classify it according to the conditions. The key components include: * **Root Node**: Splits data into two or more sets based on attribute selection techniques. * **Branch or Sub-Tree**: A part of the entire decision tree that contains a feature attribute selected from Attribute Selection Techniques. * **Splitting**: Divides a node into sub-nodes using if-else conditions. * **Decision Node**: After splitting, it's called the decision node where further splitting occurs. * **Leaf or Terminal Node**: The end of the decision tree where it cannot be split into further sub-nodes. * **Pruning**: Removes a sub-node from the tree. The process begins with selecting the root node feature based on results from Attribute Selection Measure (ASM). The ASM technique, such as Gini index or Information Gain (ID3), is repeated until a leaf node or terminal node cannot be split into sub-nodes. The measure of the degree of probability of a particular variable being wrongly classified when it's randomly chosen is known as the Gini index or impurity. The data is equally distributed based on the Gini index. When using the Gini index as a criterion for selecting features in an algorithm, the feature with the lowest Gini index is selected. This concept is closely related to Information Gain (ID3) Entropy, which helps determine a feature or attribute that provides maximum information about a class. By applying this method, we can reduce the level of entropy from the root node to the leaf node. In a practical example, let's consider predicting loan eligibility based on given data. The problem statement is taken from Analytics Vidhya Hackathon. To solve it, we follow these steps: Step 1: Load the data and finish the cleaning process. This involves either filling null values with some value or dropping all missing values (in this case, we dropped them). Step 2: Take a look at the dataset. We found many categorical values in the dataset, which are not directly supported by decision trees as features. Therefore, feature engineering techniques like label encoding and one-hot label encoding should be used. Step 3: Split the data-set into train and test sets before training a machine learning algorithm. This is crucial for evaluating model performance accurately. Step 4: Build the model and fit the train set. Before visualizing the tree, we calculate the entropy of the total dataset and the gain for every column. In the example provided, the gender column has an average information gain of 0.001, indicating it's not particularly useful in distinguishing between loan statuses. Calculating Entropy: Entropy = E(s) = -p * log2(p) - n * log2(n), where p is the number of positive cases (loan status accepted) and n is the number of negative cases (loan status not accepted). In the example, we have p = 332 and n = 148, giving us an entropy of approximately 0.89. Calculating Gain: Gain(Gender) = E(s) — I(Gender), where I(Gender) represents the information gain in the gender column. The average information gain in this column is very low, indicating it's not a useful feature for distinguishing between loan statuses. Given text: paraphrased text here (No translation or comment) People are continuing the conversation about decision trees on Science Journal on Medium. Published via Towards AI, it's explained how these tree-like structures mimic human decision-making and help solve classification and regression problems in various industries. Decision trees work by splitting data into branches based on conditions or features, with terminal nodes representing outcomes or predictions. There are two main types of decision trees: Classification Trees, which predict discrete outcomes like yes/no or class labels, and Regression Trees, which predict continuous outcomes like numerical values. Decision trees can be applied in different fields to gain actionable insights and unique examples. For example, credit risk analysis uses factors like credit history and income to predict loan default likelihood. This allows banks to segment borrowers into risk categories, giving them tailored interest rates and loan terms. Portfolio managers use decision trees to weigh risk and reward by analyzing company fundamentals and market conditions, predicting stock performance based on indicators like price-to-earnings ratio. Decision trees also help in disease diagnosis, analyzing patient symptoms, test results, and medical history to suggest diagnoses and treatment options. For example, healthcare providers can predict the probability of heart disease based on factors like cholesterol levels and age. Treatment optimization is another application, where decision trees recommend tailored treatment plans by analyzing patient characteristics. This leads to improved survival rates and reduced side effects. Customer segmentation uses decision trees to group customers based on behavior, demographics, and purchase history, identifying high-value segments. For instance, e-commerce platforms segment users by spending patterns and loyalty program engagement, targeting high-value customers with personalized offers and loyalty rewards. Predictive modeling also uses decision trees to predict customer churn and suggest retention strategies. Decision Trees in Real-World Applications Decision trees are widely used in various industries for optimizing processes, predicting outcomes, and making data-driven decisions. Decision trees enable effective irrigation scheduling and fertilizer application, minimizing waste while promoting sustainability. Key tools such as scikit-learn and XGBoost facilitate their implementation. Successful decision tree usage relies on feature selection to enhance interpretability and pruning to prevent overfitting. Ensemble methods like Random Forests and Gradient Boosting can be employed for complex problems. Transparent decision trees provide clear, interpretable decision paths, while they can handle both categorical and numerical data. However, they are prone to memorizing noise in training data and may propagate biases present in the dataset. Data sensitivity is also a concern, as performance can vary with small changes in data. Decision Trees for Decision Making and Problem Solving Decision Trees: A Powerful Tool for Decision Making Online chart creators can help build various graphs and diagrams from scratch, including decision trees. Microsoft products like Excel or PowerPoint can also be used. Additionally, a piece of paper and a pen or writing board can be employed. Decision trees are widely used in business, finance, risk management, project management, healthcare, and other areas to support decision-making. They are also utilized as root cause analysis tools and solutions. Advantages include: - Easy to understand and interpret - Minimal data preparation required - Forces consideration of multiple possible outcomes - Compatible with various decision tools - Helps make informed decisions based on available information - Identifies differences between controlled and uncontrolled events - Estimates likely results of one decision versus another Disadvantages include: - Can become overly complex - Outcomes may be influenced by personal expectations, leading to unrealistic tree structures - Diagrams can narrow focus to critical decisions and objectives

Testing decision tree. Decision tree analysis. Decision tree analysis example problems. Decision testing. Decision tree table examples. Decision tree analysis example.