* Definition of probability / Concerns the study of uncertainty / Fraction of times an event occurs /Degree of belief about an event /Probability arises in two contexts: / In actual repeated experiments / In idealized conceptions of a repeated process
* Random variables and their types / Quantifying uncertainty requires the idea of a random variable / Random Variables can be discrete (Countable or countably infinite) or continuous (takes on all values from neg to pos or all values in possibly disjoint sets, and no single value is positive / all single values are 0) / A random variable is a function that maps outcomes of random experiments to a set of properties we are interested in / A probability distribution is a description of how likely a random variable is to take on each of its possibles states
* Sample Spaces / An experiment is any activity or process whose outcome is subject to uncertainty / The sample space is the set of all possible outcomes of the experiment, denoted by S
* Event Spaces / The event space is the space of potential results of the experiment. Any collection (subset) of outcomes contained in the sample space S / An event is simple if it consists of exactly one outcome, it is compound if it consists of more than one outcome
* **Set Theory** / Union of two events A and B – denoted AUB – A or B, Intersection of two events A and B – denoted A∩B – A and B / Complement of event A – denoted by A’ is the set of all outcomes in sample space that are not contained in A / When A and B have no outcomes in common, they are said to be disjoint or Mutually Exclusive
* **Axioms of Probability**
* **P(A) + P(A’) = 1, P(A)=1-P(A’), P(A) <= 1, P(A U B) = P(A) + P(B) – P(A ∩ B), note that if A and B are disjoint P(A ∩ B) = 0 / Aiom 1: for any event A, P(A) >= 0 / Axiom 2: P(S) = 1 / Axiom 3 If A1, A2, A3 is an infinite collection of disjoint events, then P(A1 U A2 U A3 …) = Sum from 1 to ∞ of P(Ai)**
* **Conditional Probability** / P(A|B) is the probability of A given that B has occurred. This replaces S with B.P(A|B) = P(A ∩ B) / P(B) and P(B|A) = P(A ∩ B) / P(A) – Note that P(A) or P(B) must not be 0.
* **Multiplication Rule** / Since P(A|B) = P(A ∩ B) / P(B), P(A ∩ B) = P(A|B)\*P(B) – problems may specify P(A|B) and P(B). For multiple events: P(A1 ∩ A2 ∩ A3) = P(A3 | A2 ∩ A1) \* P(A1 ∩ A2) = P(A3 | A2 ∩ A1) \* P(A2 |A1) \* P(A1) – A1, then A2, then A3
* **Bayes Theorem / P(A|B) = (P(B|A) \* P(A)) / P(B)**
* **The terms are referred to P(A) the Prior Probability / P(A|B) the Posterior Probability / P(B|A) the likelihood of B given A / P(B) the Evidence / Posterior Probability = (Likelihood \* prior probability) / Evidence**
* Independence / Two events are independent if P(A|B) = P(A) and dependent otherwise. If A and B are independent, then so are the following pairs A’ and B / A and B’ / A’ and B’
* Naïve Bayes Classifiers – Family of classification algorithms, set of features (X1, X2, X3, … Xn) Predicting Class Y, assume that the features are independent, Algorithm steps: Create Frequency Tables / Create Likelihood Tables / Calculate Posterior Probability using Bayes Theorem – From Gemini: Calculate prior probability: Calculate the prior probability for each class label. Find likelihood probability: Find the likelihood probability for each attribute for each class. Use Bayes' formula: Use Bayes' formula to calculate the posterior probability. The formula is: \ (P(class|data) = (P(data|class) \* P(class)) / P(data)\). Determine most probable class: Determine which class has the highest probability, given the input, and assign the input to that class.
* **Text

  Description automatically generated**Decision Trees / Common Algorithms and their Methods: / ID3 (Iterative Dichotomizer 3) – Entropy, Information Gain / C4.5 (Successor of ID3) – Entropy, Information Gain CART (Classification and Regression) – Gini Index / **Entropy** / Entropy is a measure of impurity or randomness / Node is pure when value 0 or 1 A picture containing text

  Description automatically generated where S is the current state, pi is of an event I of state S or percentage of class I in a node of state S / **Information Gain /** How much information an attribute provides for the target variable**** A picture containing text, watch

  Description automatically generated /**Gini Index /** Diagram, schematic

  Description automatically generated pi,k is the ratio of class k instances among the training instances in the ith node / Gini = 0 when the node is “pure” and all training instances it applies to belong to the same class / **Homework** / Calculate Entropy for the target variable, then calculate entropy for the target variable given a value for one non target variable. Then calculate Information Gain for each non target variable. Highest Information gain is selected variable. / Regularlization / The regularization hyperparameters depend on the algorithm used, but generally you can at least restrict the maximum depth of the decision tree • Default max\_depth = None allows for unlimited depth. Reducing max\_depthwill regularize the model and reduce the risk of overfitting
* Random Forests / Random Forest is an ensemble of Decision Trees, generally trained via the bagging method (or sometimes pasting) / Random Forest introduces extra randomness when growing trees; instead of searching for very best feature when splitting a node, it searches for the best feature among a random subset of features / Results in greater tree diversity, which trades a higher bias for a lower variance, generally yielding an overall better model
* Bagging - also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy data set. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.
* **Discrete Distributions / PMF, CDF – a step function / Uniform pi = 1/n Any random variable whose only possible value are 0 and 1 is called a Bernoulli random variable, p and 1-p, 4 Criteria for an experiment to follow the Binomial Distribution: • The experiment consist of a sequence of n smaller experiments called trials, where n is fixed in advance. • Each trial can result in one of the same two possible outcomes, we generically denote by success (S) or failure (F). • Trials are independent, so that the outcome of any particular trial does not influence the outcome on any other trial, • The probability of success P(S) is constant from trial to trial; we denote this as probability p**
* **The probability of getting k success in n trials is P(X = k) = (n choose k) pk (1-p)n-k**
* **Poison Distribution A number of events occurring independently in a fixed interval of time with a known rate / A discrete random variable 𝑋 with states 0, 1, 2, … has probability P(X=k) = (lambdaX x e-lambda)/X! / The rate 𝜆 is the average number of occurrences of the event**
* **Continuous Distributions / Recap: / HW 7: ALL / PDF, CDF / Word problems possible / to find the constant c, set the integral to 1, can pull constants out of integral, for definite integral, evaluate at top and subtract evaluate at bottom, the integral of x to n becomes x to n+1 / n+1. Expected value is U x f(U) for all U, substitute the formula and pull U out as thought it were a constant. Evaluate the definite integral ]. Find P X > z, evaluate the definite integral**

A picture containing application

Description automatically generated

* **Gaussian Naïve Bayes – assumes likelihood of features is normal / Linear Regression – assumes residuals normal / Often an assumption in Statistical Tests ANOVA (Analysis of Variance) – assumes residuals normal / T-tests – assumes populations samples from normal / Featuring scaling – standardization**
* **X is said to have an exponential distribution with parameter lambda (lambda > 0) if the PDF of X is f(x; lambda) = Lambda \* e -lambda x for x >= 0 and 0 otherwise. Mean is 1/Lambda and Sigma Squared = Standard Deviation = 1/ Lambda 2**
* **The exponential distribution is frequently used as a model for the distribution of times between the occurrence of successive events, such as customers arriving at a service facility or calls coming into a switchboard. / The reason for this is that the exponential distribution is closely related to the Poisson process**

**Graphical user interface, text, application, Word

Description automatically generated**

**The standard gamma distribution has beta =1, the exponential distribution results from alpha = 1 and beta = 1/lambda / In Excel there is no inverse for the exponential distribution. But you can use gamma inverse with a parameter of 1, There is no uniform distribution, PDF for uniform is 1/(b-a)**

**Mean: The mean is a measure of central tendency. It represents the average value of a dataset. / Median: Another measure of central tendency. It represents the middle value when data is sorted (or the average of two middle values). Less affected by outliers compared to the mean / Variance: Variance measures the spread or dispersion of data points. A higher variance indicates greater data variability / Standard Deviation: Standard deviation is the square root of the variance, Sx. It provides a standardized measure of data dispersion Formula for variance:**

**A picture containing text, clock, watch

Description automatically generated**

**Text, letter

Description automatically generated**

* **Value of the inverse distribution functions for parameters /** NORM.DIST(X, mu, standard deviation, cumulative) returns probability (‘p’) or the area under the normal curve N(mu, standard deviation) to the left of a value X on the horizontal axis./ NORM.INV(p, mu, standard deviation, cumulative) returns the value X that has probability (‘p’) or the area under the normal curve N(mu, standard deviation) to the left.
* Null Hypothesis/Alternative Hypothesis / Null Hypothesis (H0): Represents the status quo or no effect. / Alternative Hypothesis (Ha): Represents the proposed effect or change. / Type 1 Error (False Positive): Occurs when a null hypothesis that is actually true is rejected. / The probability of Type 1 error is denoted as "alpha" (α) and is typically set as the significance level (e.g., 0.05) in hypothesis testing. / Type 2 Error (False Negative): Occurs when a null hypothesis that is actually false is not rejected. – Beta, important in medical tests. How to decide to accept or reject null hypotheses / If p value is less than level of significance, p value <= to alpha, then reject NULL Hypothesis
* Data Collection: Gather relevant data through experiments, surveys, or observations. / Formulate Hypotheses: Define null and alternative hypotheses based on research questions. / Select Significance Level (α): Determine the acceptable level of Type I error (false positive). / Perform Statistical Test: Choose an appropriate statistical test (e.g., t-test, chisquare, ANOVA) based on data type and research question. Determine P-value: Calculate the p-value, which represents the probability of obtaining results as extreme as those observed, assuming the null hypothesis is true. /  Smaller p-values indicate stronger evidence against the null hypothesis. / Make a Decision: Compare the p-value to the chosen significance level (α).
  + If p-value < α, reject the null hypothesis.
  + If p-value ≥ α, fail to reject the null hypothesis.
* **P-value calculation examples from lectures**
* **T-tests /** If P value is greater than level of significance, p value > alpha, fail to reject NULL Hypothesis / Python TTESTREL does a paired t test and returns the T Statistic and p value. / TTEST\_IND – a two sample t test, independent sample, not before and after. / T test – smaller samples with same mean
* A T-test is a statistical test used to compare the means of two groups. / It helps determine if there are significant differences between the groups. / Types of T-Tests: / Independent samples t-test: Compares means between two different groups.
*  Paired sample t-test: Compares means from the same group at different times.
*  One-sample t-test: Tests the mean of a single group against a known mean.
* • Criteria for Use:
*  Normally distributed data.
*  Scale (interval or ratio) data.
*  Random sampling from the population.
* • Common Applications:
*  Comparing test scores of two different groups of students.
*  Assessing the effect of a treatment in a before-and-after study.
*  Testing hypotheses in experimental research.