What is Data Mining?Many Definitions... Non-trivial extraction of implicit, previously unknown and potentially useful information from data Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns………….Prediction Methods Use some variables to predict unknown or future values of other variablesDescription Method Find human-interpretable patterns that describe the data…….Data Mining Tasks= Classification =  [Predictive].. Clustering= [Descriptive]...Association Rule Discovery = [Descriptive].....Deviation Detection =  [Predictive]..... Regression = [Predictive].....How do we determine the accuracy of our model? Using a test set……Training set is used to build the model… Test set is used to validate it…….Clustering Definition  Given a set of data points, each having a set of attributes, and a similarity measure among them

 Find groups or clusters such that… Data points in one cluster are more similar to one another… Data points in separate clusters are less similar to one another…..Classification: Definition Given a collection of records  Each record (row) contains a set of attributes (fields)...One of the attributes is the class (label) 1. Build a model for the class attribute based on the values of the other attributes 2. Goal: assign a class to previously unseen records as accurately as possible…… Regression  Predict a value of a given continuous valued variable based on the values of other variables  Assuming a linear or nonlinear model of dependency Greatly studied in statistics, neural network fields Examples: Predicting sales amounts of new product based on advertising expenditure Time series prediction of stock market indices…..An attribute: a property or characteristic of an object  Examples: eye color of a person, temperature, etc…….Types of Attributes….Types of values Categorical (qualitative)  Examples: types of protocols for network connection, eye color, gender, marital status, zip codes…. Quantitative Examples: age, income, length/size measurements…Discrete Attribute Has only a finite or countably infinite set of values Often represented as integer variables  Note: binary attributes are a special case of discrete attributes…. Continuous Attribute  Has real numbers as attribute values Practically, real values can only be measured and represented using a finite number of digits Typically represented as floating-point variables….. Types of Attributes Nominal = Just different names to distinguish one from another Examples: ID numbers, eye color, zip codes, sex: {male, female}....Ordinal= Provide enough information to order objects Examples: rankings {good, better, best}, grades, height in {tall,medium, short}....Interval = Differences between values are meaningful Examples: calendar dates, temperatures in Celsius or Fahrenheit…. Ratio=  Differences and ratios are meaningful Examples: age, temperature in Kelvin, length, time, counts……Record=,Data Matrix, Document Data, Transaction Data….Graph= World Wide Web,Molecular Structures….Ordered= Spatial Data, Temporal Data, Sequential Data, Genetic Sequence Data….Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set….Data Preprocessing= Discretization and Binarization, Attribute Transformation…Discretization Example= Transforming a continuous attribute to have discrete values…..Attribute Transformation = Map entire set of values of a given attribute to new set of replacement values Such that each old value can be identified with one of the new values  Examples= Use simple functions: xk, log(x), ex, |x|, Normalization: mean of 0, standard deviation of 1, New values are in range [0,1]: x\_new = (x – min) / (max – min).....Similarity= For many data mining tasks, we would like to know if two objects (records) are similar to each other, and how similar they are.. For example clustering……..Similarity= Numerical measure of how alike two data objects are, Is higher when objects are more alike, Often falls in [0,1]....Dissimilarity= ~ how different…..Example  Eye Color (what kind of attribute is this?)  {blue, brown, green, ...}  To compare eye color for two objects p and q, we can just say for similarity s:  s = 0, if different values (p!=q)  s = 1, if same value (p==q).....Rating (what kind of attribute is this?) {poor=0, fair=1, ok=2, good=3, excellent=4}......Dissimilarity = d = (p – q) / (number\_of\_values – 1).....Euclidean Distance=  If we have continuous attributes, we can calculate the distance between the two objects: dist = k=1n(pk-qk)2  n: number of dimensions (attributes), p, q: two objects, pk , qk: kth attribute value of data object p or q….The frequency of an attribute value is the percentage of time the value occurs in the data set…… The mode of an attribute is the most frequent attribute value….The mean is the most common measure of the location of a set of points= It is sensitive to outliers, The median or a trimmed mean commonly used… mean(x)= x=1mi=1mxi

Range: the difference between the max and min…… The variance is the most common measure of the spread of a set of points  variance(x)= s2x= 1m-1i=1m(xi-x)2

Visualization= Conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported…..Visualization Techniques: Histograms Usually shows distribution of values of a single variable, Divide the values into bins and show a bar plot of the number of objects in each bin,  Height of each bar indicates number of objects, Shape of histogram depends on the number of bins….Discretizing Numeric Attributes= Divide the range of possible values into sub-ranges Called buckets or bins Example: age attribute= child: 0-12, teen: 12-17,  young: 18-35,  middle: 36-59, senior: >= 60…… Equal-width discretization =Divides range into N sub ranges, each one of the same size = bin\_width = (max – min) / N  Example:  If original values are in [0..100], we could create 5 bins:  width = (100 – 0)/5 = 20 bins: [0-20], (20-40], (40-60], (60-80], (80-100] typically, first/last bins are extended (-infinity-20], (20-40], (40-60], (60-80], (80-infinity)..... Equal-frequency discretization..Divides range into N bins, each of holds same number of instances Example: 5, 7, 12, 35, 65, 82, 84, 88, 90, 95 To create 5 bins, we would divide the range of values so that each bin holds 2 of the training examples: 5, 7 / 12, 35 / 65, 82 / 84, 88 / 90, 95……Discretization= Both equal-width and equal-frequency are considered unsupervised methods That is, they do not take into account the class values of the examples…… There are supervised methods for discretization that attempt to take the class values into account  Example: Entropy-based Discretization……A cluster = set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster Centroid: average of all the points in the cluster (continuous) Medoid: the most “representative” point of a cluster(categorical)....Characteristics of Data: Important!  Sparseness Dictates type of similarity: Adds to efficiency…Dimensionality….Noise and Outliers…… K-means Clustering =Partitional clustering approach,  The basic algorithm is very simple… Main idea: Each cluster is associated with a centroid (centerpoint) Each point is assigned to the cluster with the closest centroid Number of clusters, K, must be specified (input parameter)....Partitional Clustering Dividing data objects into non-overlapping subsets (clusters),  such that each data object is in exactly one subset….Hierarchical clustering:  A set of nested clusters organized as a hierarchical tree…Centroid = mean of all points assigned to cluster  Distance = Euclidean distance of data point x to the centroid…. Issues of K-mean How to choose initial centroids One solution: Multiple runs  Helps, but probability is not on your side  Algorithm Variation: Bisecting K-means  Basic K-means algorithm can yield empty clusters  Choose points from other clusters…..K-means Clustering: How many Iterations?  K-means converges for common similarity measures Most of the convergence happens in first few iterations  Often stopping condition is changed to ‘Until relatively few points change clusters’……….Characteristics of Data: Important  Type of proximity or density measure  Attribute type Dictates type of similarity  Type of Data Dictates type of similarity…..A cluster = set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster  Centroid: average of all the points in the cluster (continuous) Medoid: the most “representative” point of a cluster(categorical)…… Classification: Definition Given a collection of records (training set )  Each record contains a set of attributes One of the attributes is the class Find a model for class attribute as a function of the values of other attributes….Classification is a fundamental task in machine learning where the goal is to categorize data points into predefined classes or categories. It is widely used in various applications, such as spam detection, image recognition, and medical diagnosis. Here are some key concepts related to classification: Supervised vs. Unsupervised Classification: Supervised Classification: In supervised classification, the algorithm is trained on a labeled dataset, which means that each data point has a corresponding class label. The goal is to learn a mapping from input features to class labels, enabling the algorithm to make predictions on unseen data….Unsupervised Classification: Unsupervised classification, also known as clustering, involves grouping data points into clusters based on their similarity, without using any predefined class labels. It is exploratory in nature and aims to discover hidden patterns or structures in the data……Labels, Training/Test Sets, etc.:…Labels: Labels are the class or category names assigned to data points. For example, in a spam email classification task, labels could be "spam" or "not spam." Training Set: The training set is a subset of the data used to train the classification model. It consists of input features and their corresponding class labels. Test Set: The test set is a separate subset of the data that the trained model is evaluated on to assess its performance. It contains input features but with the class labels hidden…… Decision Boundary: The decision boundary is the boundary or separation between different classes in the feature space. It represents the regions where the model makes decisions about which class a new data point belongs to. Decision boundaries can be linear or non-linear, depending on the classification algorithm used….Accuracy: Accuracy is a common evaluation metric for classification models. It measures the proportion of correctly classified instances out of all the instances in the test set. Accuracy is calculated as: Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)…..Confusion Matrix: A confusion matrix is a table that is used to evaluate the performance of a classification model. It provides a more detailed view of the model's predictions, breaking them down into different categories:..True Positives (TP): Instances correctly predicted as positive. True Negatives (TN): Instances correctly predicted as negative. False Positives (FP): Instances incorrectly predicted as positive (Type I error). False Negatives (FN): Instances incorrectly predicted as negative (Type II error). The confusion matrix allows you to calculate various performance metrics like precision, recall, and F1-score….k-Nearest Neighbor (k-NN): k-Nearest Neighbor is a simple and intuitive classification algorithm. Nearest Neighbor: A nearest neighbor is a data point in the training set that is most similar to a given test data point. Similarity is typically measured using distance metrics like Euclidean distance. How k-NN Works for Classification: Given a new data point, k-NN finds the k-nearest neighbors to that data point in the training set. It then classifies the new data point by taking a majority vote among the classes of its k-nearest neighbors. For example, if k = 3 and two of the three nearest neighbors belong to class A while one belongs to class B, the new data point will be classified as class A. Steps in Classification (Overall Process/Flow): Data Collection: Collect and preprocess the data. Data Splitting: Split the A screenshot of a computer program

Description automatically generateddata into a training set and a test set. Model Selection: Choose a classification algorithm. Training: Train the model on the training set. Evaluation: Evaluate the model's performance on the test set using metrics like accuracy and the confusion matrix. Fine-Tuning: Adjust hyperparameters or the model itself to improve performance. Deployment: Deploy the trained model for making predictions on new, unseen data. K-means: Parameters: Number of clusters (k): The algorithm requires you to specify the number of clusters in advance. Initial centroids: K-means can be sensitive to the initial placement of centroids, so various initialization methods are used. Assumptions: Assumes that clusters are spherical, equally sized, and have similar densities. Assumes that the features are numerical and continuous. Applicability to Different Data Types: K-means is primarily suitable for numerical data. It can be applied to data with a large number of attributes, but it may not perform well with high-dimensional data (curse of dimensionality). How it Works: K-means partitions data points into k clusters by minimizing the sum of squared distances from data points to the centroids of their respective clusters…. The algorithm iteratively assigns data points to the nearest centroid and updates the centroids until convergence. Advantages: Simple and easy to understand. Efficient and scalable for large datasets. Works well when clusters are approximately spherical and equally sized. Disadvantages: Requires specifying the number of clusters (k) in advance. Sensitive to the initial placement of centroids, leading to different results. Assumes clusters of roughly equal sizes and densities. Characteristics:K-means is an eager learning algorithm, meaning it creates a model during the training phase that can be used to classify new data points….PAM (Partitioning Around Medoids): Parameters: Number of clusters (k): Similar to K-means, PAM requires specifying the number of clusters in advance. Medoids initialization: PAM uses medoids (data points that are most centrally located within a cluster) as representatives. Assumptions: Similar to K-means, PAM assumes that clusters have similar densities but can handle non-spherical clusters better. Applicability to Different Data Types: PAM is also primarily suitable for numerical data. It can handle high-dimensional data better than K-means due to its use of medoids. How it Works: PAM is a variation of K-means that uses medoids instead of centroids. It identifies data points as medoids and assigns other data points to the closest medoid, minimizing a distance metric (usually a dissimilarity measure) such as Manhattan distance. Advantages:Robust to outliers because it uses medoids, which are less affected by extreme values than centroids. Suitable for data with non-spherical clusters. Can handle high-dimensional data better than K-means….Disadvantages: Requires specifying the number of clusters (k) in advance. Can be computationally expensive for large datasets.Characteristics: PAM, like K-means, is an eager learning algorithm that builds a model during training. k-Nearest Neighbors (k-NN): Parameters: Number of neighbors (k): The number of nearest neighbors to consider when making predictions. Assumptions: k-NN assumes that similar data points have similar class labels. It does not make strong assumptions about the data distribution….Applicability to Different Data Types: k-NN can handle both numerical and categorical data. It can work with various data types, making it versatile.How it Works:Given a new data point, k-NN finds the k-nearest neighbors in the training data based on a distance metric (e.g., Euclidean distance). It assigns a class label to the new data point based on majority voting among its k-nearest neighbors.Advantages: Simple and easy to implement. Can work well with both numerical and categorical data.No model training phase; it's a lazy learning algorithm. Disadvantages: Computationally intensive during prediction, especially with large datasets. Sensitive to the choice of k and the distance metric. Can be affected by noisy or irrelevant features.Characteristics: k-NN is a lazy learning algorithm, meaning it doesn't create a model during training. Instead, it stores the training data and uses A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generatedit for classification during prediction. Quizzes: A for loop in python that runs 7 times starts with: for i in range(7):……. Given list alist again shown below, how do you access its first element which has the value 3? 3 11 1 5 13 7 8 9= alist[0]…. what does this print?x = [3, 1, 2] print( x[-1] )= 2…..Given list alist 1 2 3 4 5 and you want to print a subset of this "3 4", which indexing/slicing can you use below?= alist[2:4]….. Given this 2-D numpy array: b = [[1 2 3][4 5 6]]how do you access the last element with value 6?= b[1][2]., b[1,2]….. Given this 2-D numpy array x2 = 4, 2, 3, 13, 2, 1 second row1, 5, 0, 15, 2, 7: what does x2.shape return? = (2,6)…….. The task of classification is to predict the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of given test record(s). (the book also uses the term "unseen test instances") = label….The class label must be this type of variable: categorical…

The \_\_\_\_\_\_ set contains attribute values as well as class labels for each instance= training….What is the accuracy of a classification that results in 144 correctly classified instances using a data set that contains 150 instances?= 96…..k-NN is based on calculating: distances or similarities between records……In k-NN, the "k" stands for the number of: nearest neighbors……n the Voronoi cells and decision boundary k-NN video, the decision boundary is: not a straight line…..

A screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generated

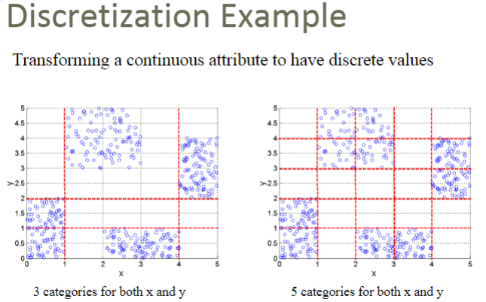
A white background with black text and numbers

Description automatically generatedA diagram of clustering of different types of clustering

Description automatically generatedA diagram of a bird

Description automatically generatedA group of cluster points

Description automatically generatedA chart with text and numbers

Description automatically generatedA screen shot of a graph

Description automatically generatedA screenshot of a computer program

Description automatically generated