Glossary terms from module 3

Terms and definitions from Course 6, Module 3

Affinity: The metric used to calculate the distance between points/clusters

Agglomerative clustering: A clustering methodology that works by first assigning every point to its own cluster, then progressively combining clusters based on intercluster distance

Average: The distance between each cluster's centroid and other clusters' centroids

Centroid: The center of a cluster determined by the mathematical mean of all the points in that cluster

Complete: The maximum pairwise distance between clusters

DBSCAN: A clustering methodology that searches data space for continuous regions of high density; stands for "density-based spatial clustering of applications with noise"

distance_threshold: A hyperparameter in agglomerative clustering models that determines the distance above which clusters will not be merged

eps (Epsilon): In DBSCAN clustering models, a hyperparameter that determines the radius of a search area from any given point

Inertia: The sum of the squared distances between each observation and its nearest centroid

K-means: An unsupervised partitioning algorithm used to organize unlabeled data into groups, or clusters

K-means++: An optimization of K-means that chooses the first centroid uniformly at random from the data points, after which each subsequent centroid is chosen from the remaining points with probability proportional to its squared distance from the point's closest existing centroid

Linkage: The method used to determine which points/clusters to merge

min_samples: In DBSCAN clustering models, a hyperparameter that specifies the number of samples in an ε-neighborhood for a point to be considered a core point (including itself)

n_clusters: In K-means and agglomerative clustering models, a hyperparameter that specifies the number of clusters in the final model

Silhouette analysis: The comparison of different models' silhouette scores

Silhouette score: The mean of the silhouette coefficients of all the observations in a model

Single: The minimum pairwise distance between clusters

Ward: Merges two clusters whose merging will result in the lowest inertia

Terms and their definitions from previous modules

A

Accuracy: Refers to the proportion of data points that were correctly categorized

B

Bayes' Theorem: An equation that can be used to calculate the probability of an outcome or class, given the values of predictor variables

C

Categorical variables: Variables that contain a finite number of groups or categories

Class imbalance: When a dataset has a predictor variable that contains more instances of one outcome than another

Collaborative filtering: A technique used by recommendation systems to make comparisons based on who else liked the content

Content-based filtering: A technique used by recommendation systems tomake comparisons based on attributes of content

Continuous variables: Variables that can take on an infinite and uncountable set of values

Customer churn: The business term that describes how many and at what rate customers stop using a product or service, or stop doing business with a company

D

Decision tree: A flowchart-like structure that uses branching paths to predict the outcomes of events, or the probability of certain outcomes

Discrete features: Features with a countable number of values between any two values

Documentation: An in-depth guide that is written by the developers who created a package that features very specific information on various functions and features

Downsampling: The process of removing some observations from the majority class, making it so they make up a smaller percentage of the dataset than before

F

F1-Score: The harmonic mean of precision and recall

Feature engineering: The process of using practical, statistical, and data science knowledge to select, transform, or extract characteristics, properties, and attributes from raw data

Feature extraction: A type of feature engineering that involves taking multiple features to create a new one that would improve the accuracy of the algorithm

Feature selection: A type of feature engineering that involves selecting the features in the data that contribute the most to predicting the response variable

Feature transformation: A type of feature engineering that involves modifying existing features in a way that improves accuracy when training the model

I

Integrated Development Environment (IDE): A piece of software that has an interface to write, run, and test a piece of code

M

Machine learning: The use and development of algorithms and statistical models to teach computer systems to analyze and discover patterns in data

N

Naive Bayes: A supervised classification technique that is based on Bayes's Theorem with an assumption of independence among predictors

P

Plan stage: The part of the PACE workflow process where a data professional first starts thinking about what the problem actually is and what needs to be done to find a solution

Popularity bias: The phenomenon of more popular items being recommended too frequently

Posterior probability: The probability of an event occurring after taking into consideration new information

Precision: The proportion of positive predictions that were correct to all positive predictions

R

Recall: The proportion of actual positives that were identified correctly to all actual positives

Recommendation systems: Unsupervised learning techniques that use unlabeled data to offer relevant suggestions to users

S

Supervised machine learning: A category of machine learning that uses labeled datasets to train algorithms to classify or predict outcomes

Supervised model: A machine learning model that is used to make predictions about unexperienced events

U

Unsupervised model: A machine learning model that is used to discover the natural structure of the data, finding relationships within unlabeled data

Upsampling: The process of taking observations from the minority class and either adding copies of those observations to the dataset or generating new observations to add to the dataset

Z

“Zero Frequency” problem: Occurs when the dataset has no occurrences of a class label and some value of a predictor variable together