Machine Learning Glossary

Accuracy score is used to evaluate the effectiveness of classification models. It calculates how well a model correctly predicts the input data against the known results. The score is a percentage, e.g., a score of 0.95 means that the model accurately predicted 95% of the results.

Activation function is a function applied to the weighted sum to form the output of a perceptron.

AI hallucination is the phenomenon of an AI model generating content that “feels” like a legitimate output, but is based on unreliable data.

Algorithmic bias refers to situations in which a computer system makes decisions that impact different people (or different groups of people) in different ways.

Algorithms (also known as **models**) are sets of logical, sequentially ordered steps.

Artificial general intelligence (AGI) (also known as **strong AI**) refers to AI that is self-aware and that is generally on par with, or more advanced than, the human brain.

Artificial intelligence (AI) is a field of computer science that specializes in mimicking human intelligence and imparting this knowledge to machines.

Artificial neural networks are algorithms that draw inspiration for their name and structure from neurons in the human brain.

Association algorithms connect pieces of information that are typically either related or used at the same time.

Attention is the process in which a machine learning model develops an internal representation of how each word is related to the next.

Bag-of-words model is an NLP model that converts text into a vector or array of numbers that computers can understand and models can use to train.

Balanced models are machine learning models that include automated sampling for imbalanced datasets.

Bias (ethics) is a situation in which one group, person, or thing is treated differently than other groups, people, or things, respectively.

Bias (mathematics) is a weight given by a perceptron that is not associated with an input value.

Binary classification is classification with only two classes or possible outcomes, e.g., spam/not spam, or approve/deny.

Binary classifier is an algorithm that classifies the input data it receives into two parts by assigning it a numerical value of either 1 or 0.

Boosting is an ensemble learning method that improves upon each iteration by assigning weights to identified incorrect predictions. Example algorithms include AdaBoost, gradient boosting, and XGBoost.

Character tokenization is the most granular form of tokenization where each character in a sentence or word is considered a token.

Chatbots are computer programs that use AI and natural language processing (NLP) to evaluate customer queries and provide personalized feedback in a manner that resembles human interaction.

Classification report shows the precision, recall, f1-score, support, and accuracy of a model's predictions.

Classification occurs when you have a classification problem you want to categorize, i.e., *classify* data into binary or multi-class discrete values.

Cluster centroids is an algorithm for synthetic oversampling.

Clustering is the process of grouping data together based on a particular similarity. Unsupervised learning models are often created by using a clustering algorithm that groups similar objects into clusters.

Confusion matrix is a report showing the total number of true positives, false positives, true negatives, and false negatives for a given set of predictions.

Context vector is a fixed-length vector that represents the source language sentence after the sentence is encoded.

Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space and takes on a value ranging between -1.0 and 1.0.

Data ethics is the application of ethical principles to data and technology.

Data standardization is the process of scaling data, which is a common practice before training a machine learning model.

Decision tree is a supervised learning classification algorithm that builds a tree of decisions based on the provided data to determine predicted classes.

Deep learning is a type of neural network that consists of three or more layers.

Dependent variables are the target variables in supervised learning, i.e., the variable you want to predict because it *depends* on other variables, also known as features.

Dimensionality reduction is the transformation of a large set of features into a smaller set that contains most of the information in the original, large set.

Elbow method is a commonly used heuristic for determining the number of clusters in a dataset.

Embeddings, otherwise known as a vector. These are a list of integers generated, or numerical values of a word or sentences generated from index encoding.

Encoding occurs when you have non-numerical categorical data in your dataset that you want to use to train your model. You must *encode* the data to a numerical value first.

Epoch is a single pass of the entire training dataset through the model.

Ethics refers to a system of principles that determine how people make decisions about what is good to do and what is bad to do.

Euclidean distance measures the distance between two data points in a multi-dimensional space and determines which objects are similar or dissimilar.

Evaluation is the process in which you determine how effective a machine learning model is. There are numerous metrics available to evaluate a model depending on the type of model that needs to be evaluated.

Explained variance is the amount of variability in the data that can be attributed to each individual principal component.

F1-score is a metric that balances both precision and sensitivity.

False negative is a datapoint of the positive class that is incorrectly predicted to be negative.

False positive is a datapoint of the negative class that is incorrectly predicted to be positive.

Features are the attributes that help determine the outcome or result that we want to predict. In statistics, they're known as independent variables. In a DataFrame, features are all column headings except the target column. For example, color, size, and shape could all be features that determine the type of fruit, which in this case is the target variable.

Filtering algorithms either remove or separate information that meets a certain criterion.

Generative AI is a specialized type of AI in which systems use transformers to interpret text-based instructions, then create various types of output.

Generative is a type of transformer that uses a decoder and its knowledge of language patterns and context to predict (i.e., generate) the next word in a sequence, given the previous ones.

Gradio is a Python library that can be used to build a graphical user interface (GUI) for any machine learning model.

GridSearchCV is a tool that will test every combination of user-provided hyperparameter settings and return the set that performs best.

Hallucinations occur where the text of a large language model is nonsensical but may be grammatically correct.

Heuristic is a technique designed for solving a problem quickly and efficiently, especially when classic methods for solving the same problem are slow or inexact.

Hidden layer is any layer in a neural network that lies between the input and output layers.

Hyperparameter tuning is the process of tweaking settings in a model to better fit a specific dataset.

Hyperparameter settings in a machine learning algorithm are set by the programmer.

Imbalanced data occurs where one or more classes are much more or less frequent than the other classes.

Independent variables is the term used for features in statistics. They're the attributes that help determine the outcome or result that we want to predict. For example, color, size, and shape could all be independent variables that determine the type of fruit, which in this case is the dependent variable.

Index encoding are encoded numerical values that represent the tokenization of words and subwords.

Inertia measures how distributed or spread out the data points are within a cluster.

Input layer is the first layer in a neural network where input values are ingested.

Input is the data given to the model.

Keras is an abstraction layer on top of TensorFlow that makes it easier to build models.

Labels are the values we want to predict in a dataset that we use to train a supervised learning model. When we want to predict a continuous value, this might be something like 26.3 for the °C temperature of a room based on certain conditions like the area of a room, outside temperature, and insulation level. When we want to predict a discrete value, the label might be something like 0 or 1, where 0 indicates that a binary classification is false, and 1 indicates that it is true. Labels must be a number, whether you are making a continuous prediction or a discrete one.

Large language model is a type of deep learning algorithm that can recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets

Large language models (LLMs) are an NLP model consisting of a neural network with many parameters.

Linear model is when the data points can either match up fairly closely to a line (regression) or be separated by a line (classification).

Linear regression a supervised learning model that can be used to predict continuous values, such as electricity usage or price.

Linearly separable data is defined by classes that can be separated with a straight line when plotted in two dimensions.

Logistic regression is a supervised learning model used for classification problems. It separates the data in a way that can be visualized by a sigmoid curve.

Loss function provides an indicator of how the model's performance changes over each iteration.

Machine learning optimization is the process of improving the performance of a machine learning model by adjusting its parameters and hyperparameters.

Machine learning is a subset of AI that enables computer algorithms to learn from data and then make decisions or predictions about future data without explicit instructions from programmers.

Mean squared error (MSE) is a metric that indicates how accurately a regression model predicts every data point. Because each error is squared, larger errors will have a disproportionately larger impact, which makes MSE sensitive to outliers. The MSE will always be above 0, but has no upper limit. An MSE of 0 means that a model perfectly predicts every data point, but most of the time the MSE will range well above 0. MSE is a function of the training data, so the MSE will vary widely between projects. Use MSE to compare models trained on the same data, but not to compare models trained on different data.

Model evaluation is when you use various methods to *evaluate* how well a trained model performs. Methods include testing a model's accuracy for classification models and R^2 for regression models.

Model-fit-predict is a three-stage pattern commonly used within supervised learning. Following this pattern, we first consider our data to select a machine learning algorithm (the model stage). Then we

present this data to the algorithm which learns from this data how to form a predictive model (the fit stage). Finally, we can use the resulting predictive model to translate a new set of input data to the correct output (the predict stage).

Model is a machine learning algorithm that may be trained or untrained depending on the stage of its process.

Multi-class or nonbinary classification occurs when you want to classify data into three or more possible categories, such as: fruit/vegetables/dairy/meat/grains.

Narrow AI (also known as artificial narrow intelligence, ANI or weak AI) refers to algorithms that perform specific tasks and make decisions based only on the data it has been trained with or exposed to.

Natural language processing (NLP) combines the rules of human linguistics with machine learning algorithms—specifically deep learning models—not only to translate text into a format that a computer can understand, but to essentially understand the meaning behind the words, including the writer or speaker's intent and sentiment.

Neural networks, also known as artificial neural networks (ANN), are a set of algorithms that are modeled after the human brain, with artificial neutrons serving the same purpose as our biological neurons.

Non-linear model is when the data cannot be modeled to a linear model, as it cannot be separated by a line.

Output layer is the layer that outputs the final predictions from a neural network.

Overfitting occurs when a model finds patterns and relationships in the training data that aren't representative of the dataset as a whole.

Oversampling occurs when more instances of the minority class are created in order to balance a dataset.

Parallelism defines how each vector passes through the encoder at the same time instead of passing through the encoder in sequence.

Parameters are the variables that models update and tune as they learn in order to improve their accuracy. This occurs in LLMs.

Pearson correlation coefficient measures the linear correlation between two variables.

Perceptron is the computational equivalent of a single neuron in the brain.

Precision is a measure of how many of the predicted positive data points were actually positive.

Predictions are the outputs of a trained supervised learning model. For example, once the model is trained on initial data, it can process new data that does not include a target variable and, based on the data's features, it will determine the probability of the result that the target variable should be. In a regression model, this prediction will be a continuous value, whereas in a classification model, this will be a number that represents a class, or category.

Preprocessing is when you make changes to the data prior to using it to train your model. This can include encoding, scaling, splitting the data into training and testing sets, removing missing data, and more.

Principal component analysis (PCA) is a statistical technique that we use to speed up machine learning algorithms when too many features, or dimensions, exist.

Prioritization algorithms help you figure out where to direct your attention. These algorithms sort or rank data based on factors that the algorithm design specifies.

R^2 or R-squared is a metric that indicates how well a regression model accounts for variability of the data. R^2 can fall between -1 and 1, and higher values signify that the model is highly predictive. An R^2 value of 0.85 means that the model accounts for 85% of the variability of the data.

Random forest is an ensemble learning method that combines a specified number of weak learner decision trees to make predictions.

Random sampling involves choosing random data points from the existing data for either over- or undersampling.

RandomizedSearchCV is a tool that will randomly select a certain number of hyperparameter combinations for testing, greatly reducing the time needed for tuning at the expense of skipping combinations.

Regression is a problem that can occur when you want to make predictions about continuous values.

Root mean squared error (RMSE) can be used to compare the predictive capacity of various regression models for the same dataset, just like MSE. It is an aggregate of the magnitude of the errors in prediction for various data points. RMSE, like MSE, has the tendency to amplify the importance of outliers, but has the advantage of using the same units as the underlying training data. That is, if your training data is in meters, then the RMSE will also be meters.

Scale data means to eliminate the measurement units and scale the numeric values to a similar scale. This helps to compare data of differing natures.

Scaling can take place in instances where there is large variation in the numerical values in your dataset (for example, salary values in the thousands versus height measurements, which are much smaller). In these cases, it’s a good idea to scale the data so a model doesn’t overly favor the features with much higher values.

Self-supervised is an approach taken by LLMs where the model starts out as an unsupervised model and evolves into a supervised model through deep learning systems that are able to add any information that's missing.

Sensitivity is a measure of how many positive data points were predicted correctly as positive, also referred to as recall.

Sequence-to-sequence (Seg2Seq) is a transformer model used for translation.

Similarity measures are mathematical methods that calculate distances between vectors. These distances allow us to compare how close vectors are together and thus how "similar" they are from a spatial perspective.

SMOTE, the synthetic minority oversampling technique, is an algorithm for synthetic oversampling.

SMOTEENN is a modified version of SMOTE that includes edited nearest neighbors (ENN).

Specificity is a measure of how many negative data points were predicted correctly as negative.

Standard scaling is a method of centering values around the mean.

Stop words are words that don't add much meaning to the NLP task.

Subword tokenization is a more complex form of tokenization where words are broken down into smaller subwords. Subword tokenization is commonly used in LLMs because they help build the vocabulary of a language model.

Support vector machines (SVM) is a supervised learning model used for classification problems. The model creates a hyperplane split between classes in order to find the optimal boundaries between data points.

Synthetic sampling synthesizes new data points from observations about existing data.

Target variable is another word for the label. It's the outcome or result you want to find after training a model with a dataset's features.

TensorFlow is an open-source platform for machine learning that will allow you to efficiently run code across multiple platforms.

Testing data is a subset of the original dataset used to test how well a trained machine learning model performs.

TinyML is a category of machine learning applications that run on low-power devices, dealing with smaller amounts of data and smaller scale algorithms.

Tokenization is the process of dividing text into smaller units called tokens. It is one of the most crucial preprocessing steps for text-based data used in NLP models.

Training data is a subset of the original dataset used to fit, or train, an untrained model (machine learning algorithm) in order for it to make future predictions.

Transformers are neural networks that learn context and thus meaning by capturing relationships and connections between words that are separated by a significant distance within a sentence or a sequence of text.

True negative is a datapoint of the negative class that is correctly predicted to be negative.

True positive is a datapoint of the positive class that is correctly predicted to be positive.

Underfitting is a common problem in machine learning where a model fails to sufficiently capture the relationship between the input data and the output variable.

Undersampling occurs when instances of the majority class are removed in order to balance a dataset.

Validation can sometimes be the same as model evaluation, although model evaluation often has additional metrics used to determine the efficacy of a model. Validation is the process of determining if a model works as intended to make predictions.

Vectors is an object that has both a magnitude and a direction. It is usually depicted geometrically as a line whose length is the magnitude of the vector, and with an arrow indicating the direction.

Weak learner is a term used for a machine learning algorithm that makes predictions that are only slightly better than random chance. A single weak learner makes inaccurate and imprecise predictions because it is poor at learning adequately due to limited data, like too few features, or data points that can't be classified. Weak learners can be combined within an ensemble learning algorithm in order to improve the accuracy of results, as the predictions that each weak learner makes help determine the final prediction.

Weighted sum is a value formed by adding weighted values together.

Weighting occurs when multiplying input values by weights determined by their predictive importance to the model.

Word tokenization is the simplest form of tokenization where each word is considered a token. Word tokenization is very similar to the bag-of-words model.

X is a variable commonly used in machine learning code to represent the features data.

y is a variable commonly used in machine learning code to represent the target variable, also known as the labels.
