GLOSSARY

BTT Units 1-9

Table of Contents

Unit 1 2

Unit 2 6

Unit 3 10

Unit 4 13

Unit 5 16

Unit 6 19

Unit 7 23

Unit 8 27

Unit 9 30

Unit 1 Glossary

Array

A collection of numbers of a given type, such as float or int, in one or more dimensions.

Axis

A particular dimension (or direction) in an array or a DataFrame.

Broadcasting

A NumPy feature that enables mathematical operations to be applied to arrays of different sizes and dimensions.

Cell

A unit of structure in a Jupyter Notebook that can contain multiple lines of code to be run as a unit.

Classification

One of two classes of methods in supervised learning, where the label is a categorical value. The two types of classification are binary classification and multi-class classification.

Cross-Industry Standard Process for Data Mining (CRISP-DM)

A popular diagram used to represent the process used for building machine learning models.

Data matrix

A structured table consisting of rows and columns.

DataFrame

A data table or spreadsheet with row and column headers, where each column contains data of a particular type but which can be of different types in different columns.

Ethical risk

The likelihood for large-scale and automated decision systems to cause unintended harms.

Example

An instance of data; also known as a data point.

Features

Input variables which are predictive data elements of a machine learning problem. They are the data contained in the columns of a data matrix. One feature value is contained in one column.

Generalization

A model's ability to adapt to new, previously unseen data.

Heuristics

A way of solving problems where the objective is to produce a solution within a reasonable time frame or range of accuracy.

Jupyter Notebook

A web-based interpreter which ties together code, analyses, documentation, and graphics.

Label

In supervised learning, the "answer" or "result" portion of an example. Each example in a labeled data set consists of one or more features and a label. For instance, in a housing data set, the features might include the number of bedrooms, the number of bathrooms, and the age of the house, while the label might be the house's price. In a spam detection data set, the features might include the subject line, the sender, and the email message itself, while the label would probably be either "spam" or "not spam."

Labeled example

An example which contains features and a label.

Labels

What you want to infer about a data point. Your training data has labels so that you can train your function to predict the label of test points.

Library

A set of related software items (e.g., functions, objects) that can be called from within a program but which are defined externally to that program, typically to provide a defined set of operations that are useful to a variety of different applications.

Machine learning (ML)

A broad class of methods and algorithms for building predictive models from data without prescribing the specific form of relationships between inputs and outputs. ML is considered a subfield in the larger field of artificial intelligence but also straddles the world of data science, which is an amalgamation of human insight and automated inference.

Machine learning model

A computer program that has been trained to recognize patterns in data to make predictions on future data.

Notebook

A computational environment that generally combines code, documentation, results, and graphics. Jupyter Notebooks are a widely used platform that supports work with Python as well as several other programming languages.

Package

Within the Python ecosystem, a collection of related software items that are bundled and distributed together to provide specific functionality in a Python program. A package might simply be a library (and sometimes the terms are synonymous), or it might contain additional tools beyond a library that support working with Python.

Recommendation systems

Machine learning systems designed to recommend items to you on various websites and apps.

Regression

One of two classes of methods in supervised learning, where the label is any real valued number.

Supervised learning

A class of machine learning problems in which labeled data are available, enabling an algorithm to learn how to associate data values with data labels so that predictive models for classification or regression on unseen data are possible.

System risk

The likelihood for complex and dynamic systems to have failure points.

Training

Creating or learning the model.

Unlabeled example

An example which contains only features and no label.

Unsupervised learning

A class of machine learning problems in which labeled data are not available, whereby algorithms work to identify various types of patterns in data.

Vectorization

A NumPy feature that performs operations on entire arrays that would normally be performed through the use of loops.

Unit 2 Glossary

Array

A collection of numbers of a given type, such as float or int, in one or more dimensions.

Axis

A particular dimension (or direction) in an array or a DataFrame.

Binary indicators

A data transformation technique for transforming data to binary based on meeting a true/false condition.

Bivariate plots

Plots that show a relationship between two data columns or two dimensions.

Box-and-whisker plot (also known as box plot)

A type of data visualization to characterize the distribution of a set of numerical values, constructed by representing the following statistical quantities of the data set: median (50% level), first quartile (25%), third quartile (75%), minimum, and maximum.

Categorical data type

A data type that is identified based on the label given to it. Two types are ordinal and nominal.

Classification

The process of predicting categorical labels for previously unseen input data based upon prior training of a model (a “classifier”) using labeled data examples.

Classification model

A type of machine learning model for distinguishing among two or more discrete classes. For example, a natural language processing classification model could determine whether an input sentence was in French, Spanish, or Italian.

Column

Each column in a data matrix contains the feature or attribute values. In supervised learning, one column contains an associated label value.

Continuous data type

A data type consisting of standard floats or numbers; the distribution of continuous data types can produce outliers.

DataFrame

A data table or spreadsheet with row and column headers, where each column contains data of a particular type but which can be of different types in different columns.

Exploratory data analysis (EDA)

Looking at your data to answer three key questions: How is the data distributed? Which features are redundant? How do different features correlate with your label?

Feature

An input variable used in making predictions. Features are the relevant characteristics or attributes of the data. For example, the features we might collect to identify fraudulent bank transactions might include dollar amount, type of transaction, country of origin, frequency, etc.

Feature engineering

The process of determining which features might be useful in training a model then preparing and transforming the raw data into features that can be used for the model to train on.

Functional transforms

A data transformation technique that transforms a numeric input X into a new numeric value based on f(X).

Histogram

A type of data visualization to characterize the distribution of a set of numerical values, constructed by putting data into a set of discrete bins and plotting the number of counts in each bin.

Interaction terms

A data transformation technique that takes the multiplication of two numeric types.

Join

In database terminology (and carried over to related areas where one is working with multiple related data tables), the process of connecting different data tables together through a set of shared keys or labels. For example, in a company's employee database, there might be one table that stores employee personal information, one that stores work schedules, and one that stores wage and salary information. They might all share an entry for each employee that encodes a unique employee ID, so one could join these different data tables by matching up employee IDs across tables.

Label

What you want to predict. Your training data has labels so that you can train your model to predict the label of test examples. Each example in a labeled data set consists of one or more features and a label. For instance, in a housing data set, the features might include the number of bedrooms, the number of bathrooms, and the age of the house, while the label might be the house's price. In a spam detection data set, the features might include the subject line, the sender, and the email message itself, while the label would probably be either "spam" or "not spam."

Labeled example

An example containing features and a label.

Matrix

A 2-dimensional array of N vectors of size K.

Nominative data type

A categorical type, often represented as strings. Nominative data types are the typical basis for classification.

One hot encoding

The process of creating binary indicators from categorical value types. Each separate category would have its own binary indicator in one-hot-encoding.

Ordinal data type

Consists of a mix of categorical and numeric. It is usually an integer-based number, but the range is usually so small that we can treat them as discrete categories. The absolute ordering matters in terms of being able to compare different values, but the relative difference doesn't.

Row

In a data matrix, a row corresponds to an "example," also called a data point. In supervised learning, the row consists of features and one label.

Sampling

The process of extracting subsets of examples from some available universe of data.

Scatter plot

A type of data visualization for a pair of associated data sequences, where each pair is represented by a single point in the x-y plane — useful for visualizing the relationship between two sets of data values.

Unit of analysis

A real-life representation of an example.

Vector

A 1-dimensional array of K elements.

Winsorization

The tranforsmation of statistics by setting extreme outliers equal to a specified percentile of the data to reduce the effect of having too many outliers.

Unit 3 Glossary

Bias

Model bias expresses the error that the model makes (how different the prediction is from the training data). High bias means that the model is too simple and failed to capture the relationship between the features and labels; it is a sign that the model is underfitting. This happens, for example, when you make the wrong modeling assumptions, such as training a model on data for which it is not suited.

Decision trees

A popular supervised learning algorithm that relies on recursively splitting the data into partitions. You can keep track of these partitions in a tree structure. During inference, an unlabeled example traverses the tree until it falls into a leaf. Each leaf is associated with one of the data partitions, and you assign the unlabeled example the most common label within that partition (or the average label in the case of regression).

Distance function

A special type of function used to determine nearness in k-nearest neighbors; typically defined between two points.

Entropy

A useful formula that measures dispersion or uncertainty of a discrete random variable; also used in decision trees.

Euclidean distance

The most commonly used distance function; it represents the straightest distance between two points.

Generalization

A model's ability to adapt to new, previously unseen data.

High-dimensional data

A data set with too many features. In such cases, it becomes difficult to train a model that can find the relationship between features and a label.

Hyperparameters

The “knobs” that you tweak during successive runs of training a model; they help guide the learning process. These are parameters in the model that are not learned but set prior to learning. Hyperparameters often trade off complexity vs. simplicity of models.

Information gain

A formula that measures the difference in the average entropy of a variable after segmenting the data into multiple partitions; also used in decision trees.

Instance-based learning

A type of supervised learning model in which training examples are stored in memory. Those examples are utilized on demand to make predictions for a new, previously unseen example.

K-nearest neighbors

A commonly used supervised learning algorithm that makes the assumption that similar points of data share similar labels. The algorithm predicts the label of a test point through a majority vote among its k-nearest neighbors within the training set.

Loss

A measure of how far a model’s predictions are from its label; to phrase it more pessimistically, a measure of how bad the model is. To determine this value, some models use a loss function.

Loss function

A specialized mathematical function that represents how well our models predict the labels; i.e., the “error.”

Model calibration

Setting unique parameters achieved by using the measurements from the model’s predictions. The parameters are used to provide a good description of the system’s behavior.

Neighbor count (k)

The number of nearest neighbors to use in prediction.

Normalization

The methodology used to ensure features are on the same scale.

Overfitting

A model failure mode that occurs when a model is too complex. It learns the training data so closely that it does not generalize well to new data. An overfit model has low training error but poor generalization.

Regression

The process of predicting continuous numerical values of some quantity of interest for previously unseen input data based upon prior training of a model (a "regressor") using labeled data examples. One of three categories for predicted labels, y, used when y is a real value; for example, the price of a house. (The other two categories are binary classification and multiclass classification.)

Regression model

A type of model that outputs continuous (typically, floating point) values. Compare with classification models which output discrete values such as "day lily" or "tiger lily."

Scikit-learn

Software for Python that has a wide range of algorithmic options, covering regression, classification, and unsupervised learning. It also provides rich libraries for data preparation, model selection, and evaluation.

Supervised learning

A class of machine learning problems in which labeled data are available, enabling an algorithm to learn how to associate data values with data labels so that predictive models for classification or regression on unseen data are possible.

Training data

A subset of data used in a supervised learning problem to fit or "train" a predictive model, which can then be used to make predictions about unseen data (e.g., in a testing set).

Underfitting

A model failure mode that occurs when the model is too simple. It is unable to learn important nuances in the training data to properly make predictions. An underfit model has high training error and poor generalization.

Variance

Model variance expresses how consistent the predictions of a model are if it is trained on different sections of the training data set. High variance is a sign that the model is overfitting to the particular data set on which it is trained.

Unit 4 Glossary

Dot product

Also referred to as scalar product, a special operation that takes the element's product of each vector and then sums them together. It is typically represented with what looks like a period or a dot floating between the two vectors.

Gradient descent

A numerical optimization algorithm used to train and optimize a logistic regression model using loss functions. Gradient descent iteratively updates the model parameters until a loss function is minimized.

Hessian matrix

A square matrix of second-order partial derivatives of a scalar-valued function, or scalar field.

Identity matrix

A special matrix whose diagonal elements are one, and all other elements are zero.

Intercept

Sometimes called the "constant." Intercept in a linear regression model is the mean value of the response variable when all of the predictor variables in the model are equal to zero. Intercept in a logistic regression is the "log odds" of the response variable, not the mean.

Inverse Logit

One of three steps in the logistic regression model. The inverse logit step transforms the output of the linear step into a probability prediction $P(y|X)$ between 0-1.

Learning rate

A common logistic regression hyperparameter (also commonly known as the step size), learning rate dictates the speed of gradient descent. The ideal learning rate is one that reaches global minima in a fast and efficient manner.

Linear models

A class of supervised learning models that are represented by an equation: simple to implement, fast to train, and have lower complexity. Logistic regression is an example of a linear model. In a linear model, the form of a model must fit a very specific format.

Linear regression

A supervised machine learning algorithm used for regression problems. Linear regression finds a linear relationship between one or more features and a label (such as a price or an age.) There are two types of linear regression models: Simple linear regression finds the linear relationship between one feature and one label, and multiple linear regression finds the linear relationship between multiple features and one label.

Linear step

One of three steps in the Logistic Regression model. The linear step computes a value $z(X)$ by taking the linear sum of feature values X with model weights W and an intercept term α (also known as bias, β, in some literature).

Log loss

A popular loss function used to measure the performance of a some classification models. This loss function is used for both logistic regression and neural networks.

Logistic regression

A linear classification method that is trained by iteratively tuning a set of weights to minimize the log loss.

Loss function

Specialized mathematical function that represents how well our models predict the labels. A loss function quantifies the amount of error a model makes against the training dataset.

Matrix inverse

A matrix that when multiplied by the original matrix produces the identity matrix.

Matrix multiplication

A binary operation that produces a matrix from two matrices.

Mean squared error

Loss function commonly used to measure the performance of a regression model such as linear regression.

Non-linear models

The opposite of linear models, non-linear models can take many different forms: Non-linear models have more complexity, meaning they can draw more sophisticated curves to fit arbitrary patterns in the data.

NumPy

A Python library that adds support for multi-dimensional arrays and matrices. NumPy also provides an extensive collection of high-level mathematical functions to do element-wise operations on entire arrays.

Optimization algorithm

An algorithm that uses a loss function to evaluate a model's loss and then adjusts the model parameters accordingly to reduce loss. It continues this process until an optimal model is produced. A popular optimization algorithm is gradient descent.

Ordinary least squares (OLS)

A non-iterative method used in linear regression to minimize the sum of the squared errors between the model predictions and the actual values.

Overfitting

A model failure mode that occurs when a model is too complex. It learns the training data so closely that it does not generalize well to new data. An overfit model has low training error but poor generalization.

Regularization

The penalty on a model's complexity; helps prevent overfitting. Different kinds of regularization include L1 regularization and L2 regularization.

Scikit-learn

Software for Python that has a very wide range of algorithmic options, covering regression, classification, and unsupervised learning. It also provides rich libraries for data preparation, model selection, and evaluation.

Zero-one loss

A simple loss function that counts how many mistakes an hypothesis function h makes on the training set. This loss function is not suited for use in the training process to optimize the model but is suited for the evaluation of classification models.

Unit 5 Glossary

Accuracy

A performance metric for classification models that is the number of correct predictions out of the total number of predictions.

Area under the receiver operator curve (AUC)

A commonly used metric for measuring a binary classifier's performance.

Base rate

Pertaining to a model, the percent of cases in your evaluation data where Y equals 1.

Classification

A supervised learning method in which the label is a categorical value. The two types of classification are binary classification and multiclass classification.

Conditional expected value

The likely average future value of Y in cases where X is true.

Empirical risk minimization

Choosing the model that minimizes loss on the training set.

Expected value

The likely average future value of Y.

Expected value estimation

The most likely value of an outcome given known information about an example

Feature selection

The process of empirically testing different combinations of features to choose an appropriate set.

Generalization

A model's ability to adapt to new, previously unseen data.

Heuristic selection

A feature selection method that filters out features using heuristic rules prior to modeling.

Hyperparameters

The “knobs” that you tweak during successive runs of training a model. Hyperparameters often trade off complexity vs. simplicity of models. There are many heuristics to set hyperparameters; a common one is to use grid search.

Implicit feature selection

Reducing feature count as a byproduct of the model training procedure.

K-fold cross-validation

A resampling method that uses different portions of the data to train and validate the model on different partitions of the data.

Model deployment

The process of using a machine learning model in a production environment where it can be used for its intended purpose.

Out-of-sample validation

Computing evaluation metrics on examples that were not part of model training. Out-of-sample validation helps us approximate the expected loss and not rely solely on the training loss.

Precision

Percentage of positive predictions that were actually positive.

Ranking

Sorting examples and choosing top K to fulfill some optimization objective.

Recall

Percentage of actual positives that were correctly classified as positive.

Receiver operator curve (ROC)

A curve that represents the performance of your binary classification model at various classification thresholds.

Regression

A supervised learning method in which the label is any real valued number.

Regularization

The penalty on a model’s complexity; helps prevent overfitting. Different kinds of regularization include L1 regularization and L2 regularization. L1 regularization can be used for feature selection.

Stepwise selection

Feature selection method to iteratively add/reduce features based on empirical model performance.

Supervised learning

A class of machine learning problems in which labeled data are available, enabling an algorithm to learn how to associate data values with data labels so that predictive models for classification or regression on unseen data are possible.

Test set

The subset of the data set that you use as a final test of your model's performance.

Training set

The subset of the data set used to train a machine learning model to make predictions.

Validation set

The subset of the data set that is used to evaluate models' performances when performing model selection.

Unit 6 Glossary

Bagging

A shortening of the phrase “bootstrap aggregating”; an ensemble method that improves the stability and accuracy of models. Bagging powers the random forest algorithm.

Bias

Model bias is a component of the model’s error (how different the prediction is from the training data). A high bias means that the model is too simple and fails to capture the relationship between the features and labels; it is a sign that the model is underfitting. This happens, for example, when you make the wrong modeling assumptions, such as training a model on data for which it is not suited.

Bias-variance tradeoff

Finding the right balance of values between bias and variance.

Boosting

An ensemble modeling technique that combines a set of weak models into a strong model by adding models that fit the residual of prior models.

Bootstrapping

A process that takes multiple or different samples from a data set, computes some quantity or statistic on each sample, then averages them to get a final estimate.

Clustering

An unsupervised learning technique. It is the process of identifying or grouping subsets of data (“clusters”) that are collectively similar to one another, based on some specified criterion for defining similarity.

Decision trees

A popular supervised learning algorithm that relies on recursively splitting the data into partitions. You can keep track of these partitions in a tree structure.

Dimensionality reduction

The process of developing an approximate representation of a dataset that includes fewer features (or dimensions of a dataset), based upon identifying substructure within those data (e.g., correlations) that make such an approximation useful.

Ensemble methods

A class of techniques that train multiple models and aggregate them into a single prediction.

Estimator

A function that estimates a value based on other observations.

Feature extraction

The process of identifying meaningful subsets of data ("features") based on some criterion of interest. Examples might include extracting various facial features from images of people, or identifying interesting astronomical events from large-scale sky surveys.

Gradient boosted decision trees (GBDT)

An ensemble algorithm that is the most popular algorithm that uses boosting. It consists of individual decision trees.

Hierarchical clustering

An algorithm for clustering that involves constructing hierarchical trees relating the input data, such that nearby data points in the branching tree are more similar to each other. The hierarchical tree is built up in an iterative fashion, by accreting (or agglomerating) subtrees to build up larger trees. Thus hierarchical clustering is also known as agglomerative clustering.

Hyperparameters

The "knobs" that you tweak during successive runs of training a model; they help guide the learning process. They are parameters in the model that are not learned but set prior to learning. Hyperparameters often trade off complexity vs. simplicity of models.

K-means clustering

An algorithm for clustering that involves specifying a number of desired clusters (the number k), and which is based on assessing Euclidean distances between data points.

Learning rate

A common GBDT hyperparameter (also typically known as the step size) that dictates the speed of gradient descent. The ideal learning rate is one that reaches global minima in a fast and efficient manner.

Linkage method

A prescription for characterizing the similarity of clusters within a clustering algorithm. There are a handful of widely used linkage methods, such as: “ward linkage,” which minimizes the sum of squares of the differences within all clusters; “maximum or complete linkage”, which minimizes the maximum distance between observations of pairs of clusters; “average linkage,” which minimizes the average of the distances between all observations of pairs of clusters; and “single linkage,” which minimizes the distance between the closest observations of pairs of clusters.

Logistic regression

A linear classification method that is trained by iteratively tuning a set of weights to minimize the log loss.

Overfitting

A model failure mode that occurs when a model is too complex. It learns the training data so closely that it does not generalize well to new data. An overfit model has low training error but poor generalization.

Random forest

An ensemble learning method containing a set of decision trees (typically consisting of dozens to hundreds). The decision trees in a random forest are used for classification and regression problems, along with other tasks that require predictions.

Sampling

The process of extracting subsets of examples from some available universe of data.

Scatter plot

A type of data visualization for a pair of associated data sequences, where each pair is represented by a single point in the x-y plane; useful for visualizing the relationship between two sets of data values.

Similarity measure

A prescription for characterizing the similarity between any two data points in a dataset, for use with clustering. Two data points that are identical should have a maximum similarity. Inversely related to similarity is the notion of distance: The distance between any two identical data points is 0. Euclidean distance is one measure of distance, based on the sum of the squares of the differences between all coordinates. Mathematically, there are many different distances between two data points that can be defined. Clustering algorithms typically define similarity based on an underlying distance metric, but some algorithms are able to use notions of similarity that are not tied to a strictly mathematical measure of distance.

Stacking

An ensemble method that doesn't have a specific supervised learning method attached to it. An implementation of stacking can be done using a combination of any common supervised learning algorithms that you've already learned, such as logistic regression or decision trees.

Supervised learning

A class of machine learning problems in which labeled data are available, enabling an algorithm to learn how to associate data values with data labels so that predictive models for classification or regression on unseen data are possible.

Unsupervised learning

A class of machine learning problems in which labeled data are not available, whereby algorithms work to identify various types of patterns in data.

Variance

Model variance expresses how consistent the predictions of a model are if it is trained on different sections of the training data set. High variance is a sign that the model is overfitting to the particular data set on which it is trained.

Unit 7 Glossary

Activation function

A function which transforms linear inputs into nonlinear forms to help the network learn complex patterns in the data.

Convolutional neural network (CNN)

A deep-learning neural network that makes use of convolutional layers. It learns convolutional filters that process the input image to a particular layer and produce a new image (with many channels). CNNs are particularly useful for processing image data but can also be used for audio signals or videos.

Cosine similarity

A function that computes the similarity between two vectors (or sequences of numbers).

Count vectorizer

A scikit-learn library tool in Python used to convert a collection of text into a vector. It can tokenize a collection of text documents and build a vocabulary of known words.

Data preprocessing

The manipulation or dropping of data before it is used to ensure or enhance performance.

Deep averaging network

A neural network consisting of two components: a word embedding and a traditional neural network (sometimes even a linear classifier, which is a neural network without hidden layers).

Deep neural network

A neural network with many more than two hidden layers, each with many more than two nodes. Traditional networks have a few hidden layers, whereas deep learning models can have hundreds of hidden layers. Empirically deep neural networks perform better on certain machine learning tasks, such as image recognition.

Feedforward neural network

A type of artificial neural network in which nodes' connections do not form a loop. A feedforward neural network typically takes one word or sentence in a text sequence. Still, it loses any context of what was previously input.

Gradient descent

An optimization algorithm that searches for the minimum of a loss function by slightly changing the parameters of a classifier in the direction of the greatest negative gradient of the loss. It is often too hard to calculate the gradient of the loss for all training examples; for this reason, we usually use stochastic gradient descent instead.

Hidden layer

A layer of a neural network located between the input and output layers which applies a linear transformation followed by a nonlinear transformation, called the activation function, to its input values. All neural networks have at least one hidden layer; deep neural networks have many hidden layers.

Hyperparameters

The "knobs" that you tweak during successive runs of training a model; they help guide the learning process. They are parameters in the model that are not learned but set before learning. Hyperparameters often trade off the complexity vs. simplicity of models.

Machine learning (ML)

A broad class of methods and algorithms for building predictive models from data without prescribing the specific form of relationships between inputs and outputs. Machine learning is considered a subfield in the larger field of artificial intelligence but also straddles the world of data science, which is an amalgamation of human insight and automated inference.

N-grams

Combinations of individual word tokens.

Natural language processing (NLP)

A branch of artificial intelligence (AI) that enables machines to understand the human language. NLP interprets raw, arbitrary written text and transforms it into something a computer can understand.

Neural network

A supervised learning algorithm designed to solve complex, real-world problems. It can recognize complex patterns and nonlinear relationships between features and labels. Neural networks are often used in the NLP field.

Output layer

In neural networks, the last layer that outputs a prediction.

Recurrent neural networks (RNN)

A special type of neural network designed so that a given node's output flows back into the same node. In RNN, the information cycles through a loop. When it makes a decision, it considers the current input and what it has learned from the inputs it has previously received.

Scikit-learn pipeline

A scikit-learn utility that orchestrates the flow of data into and out of a machine learning model to automate the machine learning workflow.

Stochastic gradient descent

An approximation of gradient descent. The gradient of the loss function is applied to a subset of all the training examples instead of the whole set, which is much faster to compute.

Stop word

A token that appears very frequently in different examples of text but also adds very little predictive value.

Stop word removal

A data preprocessing step that removes the words that commonly occur. Stop words are usually removed to reduce data size and to speed up computation.

TF-IDF vectorizer

Also known as "term frequency-inverse document frequency"; a process for encoding text that captures the relative importance of a word to a given document.

Text classification

A machine learning technique that assigns a set of predefined categories to open-ended text, categorizing text into organized groups.

Tokenization

The process of parsing text to remove certain words. Tokenization allows you to use textual data for predictive modeling and map every word in training data to a future position.

Vectorization

The simple process of representing a word's binary presence or frequency in a given text example. Vectorization is a common strategy for mapping individual word tokens to a number.

Word embedding

A type of word representation which allows words with similar meanings to have an equal representation. In word embedding, each word can be represented by a k-dimensional vector. Those factors are commonly pre-trained and available as a lookup table.

Unit 8 Glossary

Agile model development

The process of increasing a model's complexity in separate efforts.

Algorithmic accountability

A concept where model owners and developers are accountable for the decisions that their machine learning systems make.

Allocative harm

A discriminatory system that withholds certain opportunities, freedoms, or resources from specific groups.

Bias

Model bias expresses a type of error a model makes (how different the prediction is from the training data). High bias means that the model is too simple and failed to capture the relationship between the features and labels; it is a sign that the model is underfitting. This happens, for example, when you make the wrong modeling assumptions, such as training a model on data for which it is not suited.

Bias-variance tradeoff

The property of a model where the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

Class imbalance

A problem in machine learning where the total number of a class of positive data is far less than the total number of another class of negative data.

Concept drift

Changes in the statistical properties of data over time.

Ethical AI

The subfield of AI that studies model fairness and accountability.

Execution bottleneck

A data preparation or modeling process not terminating in a reasonable amount of time. This common and frustrating issue is often caused by a misalignment among the data, tools, and hardware you use to process the data.

Fairness

The various attempts at correcting algorithmic bias in automated decision processes based on machine learning models.

Feature importance

Techniques that assign a score to input features based on how useful they are at predicting a target variable.

Feature leakage

Data leakage that occurs in information that is used in the model training process that would not be available at prediction time.

Features

Input variables which are predictive data elements of a machine learning problem. They are the data contained in the columns of a data matrix; one feature value is contained in one column.

Learning curve

A plot that shows the relationship between data size and model prediction performance.

Learning curve analysis

A method to empirically measure data size efficiency.

Logistic regression

A linear classification method that is trained by iteratively tuning a set of weights to minimize the log loss.

MapReduce

A programming framework that performs distributed and parallel processing on large data sets in a distributed environment.

Model performance failure

When evaluating a test, a performance failure describes a model that may have poor performance. This can lead to overconfidence in a system and poor generalization performance when the model is applied to live data.

Model variance

Expresses how consistent the predictions of a model are if it is trained on different sections of the training data set. High variance is a sign that the model is overfitting to the particular data set on which it is trained.

Representational harm

A type of harm where a system reinforces negative stereotypes along the lines of identity and protected class.

Reproducibility

The ability to duplicate a model exactly such that given the same raw data as input, both models return the same output.

Societal failures

Failures which happen when a machine learning model produces unintended discrimination or disparate impact and lacks accountability. It is where failure exists between the model's owner and the social context in which the model operates.

Stratified sampling

A sampling method that reduces the sampling error in cases where the population can be partitioned into subgroups.

Unit testing

A software development process in which the smallest testable parts of an application, called units, are individually and independently tested for proper operation.

Upsampling

A strategy of taking 100 percent of the negative classes and sampling the positive class cases with replacement until you get equal sizes for both. Upsampling is a preferred strategy when you have limited data to begin with and can't afford to discard any.

Unit 9 Glossary

Algorithmic accountability

A concept where model owners and developers are accountable for the decisions that their machine learning systems make.

Allocative harm

A discriminatory system that withholds certain opportunities, freedoms, or resources from specific groups.

Ethical AI

The subfield of AI that studies model fairness and accountability

Fairness

The various attempts at correcting algorithmic bias in automated decision processes based on machine learning models.

Model deployment

The process of using a machine learning model in a production environment where it can be used for its intended purpose

Representational harm

A type of harm where a system reinforces negative stereotypes along the lines of identity and protected class.

Societal failures

Failures which happen when a machine learning model produces unintended discrimination or disparate impact and lacks accountability. It is where failure exists between the model's owner and the social context in which the model operates.