Glossary DRAFT

JOEL DAVIS

January 2022

The Very drafty version of a glossary

A/B testing

An A/B test compares (A and B, although there can be more) treatments. Historically this was a current state and some new version. An example may be a website with three offers/items for sale. The operator can remove just one item, and replace it with another, showing half of the website visitors the original (version A) and the other half of visitors the new version (B). After some time, the operator assesses which version performed better on the metric of interest.

accuracy

TBD

action

(reinforcement learning) (SEE ALSO reinforcement learning, agent, environment, state, policy, reward)

In reinforcement learning, software agents take actions within an environment. The output from these actions comes in the form of a reward. The software agent's goal is to learn (use) actions in such a way as to maximize the long-term rewards.

Example: If a software agent were playing a video game with a computer mouse as a controller, the actions available to the agent might be the 2d directional moves on the mouse, the scroll wheel, and the buttons on the mouse

activation function

(neural networks) (SEE ALSO reLU, sigmoid)

In neural networks, a function that takes a weighted sum of inputs from the prior layer and generates (through its function) output forward to the next layer. These are typically (but not necessarily) non-linear functions. Popular activation functions are reLU and sigmoid.

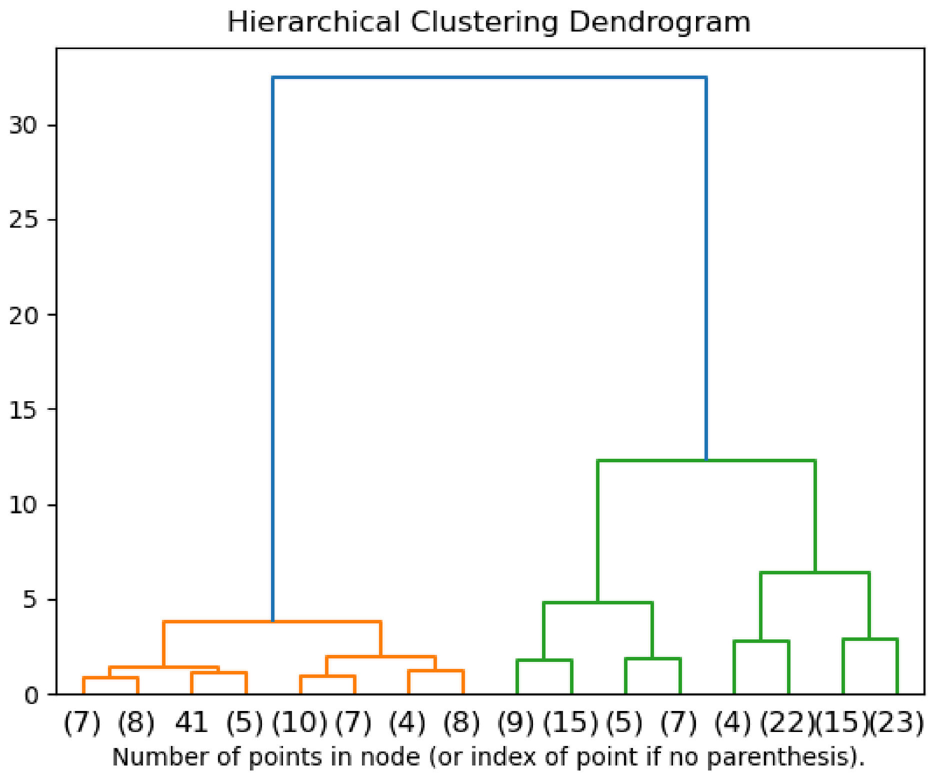
agent

In reinforcement learning, software agent's take actions within an environment. The output from these actions comes in the form of a reward. The software agents goal is to learn (use) actions in such a way as to maximize the long-term rewards. The agent is the entity (software) that uses a policy (rules) to perform some action (does something).

agglomerative clustering

(unsupervised learning, clustering) (SEE ALSO hierarchical clustering)

An Agglomerative clustering algorithm starts from an individual data point (a cluster of 1) and calculates the distance between each cluster. The pair of clusters with the shortest distance to each other are joined (now a cluster of 2), and the process is repeated between all clusters. Visually this can be depicted as a tree diagram. The language of agglomerative clustering leans into this depiction. Each cluster is called a leaf and builds through branches to the root.



An example of agglomerative clustering from the sklearn documentation[1]

anomaly detection

Anomaly detection is the process of finding unexpected observations or events (outliers). These are usually events or observations that deviate from some established norm. An example might be a bank that wishes to understand when a bad actor is using your credit card. They may use anomaly detection to discern whether an incoming transaction is a potential fraud/theft. Perhaps the card is being used almost simultaneously several hundred miles apart.

artificial general intelligence

There are several definitions of artificial general intelligence (AGI), but they share similar themes. AGI is the ability of a system to display the same general intelligence as a human and is not tied to performing a specific task.

Contrast this with “Narrow AI,” which performs specific tasks, like chess or GO, better than any human but can not operate at that level of high performance outside the domain it was trained in (can’t drive a car, for example).

artificial intelligence

A non-human, engineered system that solves or performs tasks. Artificial intelligence is intelligence displayed by engineered systems, as opposed to intelligence exhibited by biological entities (humans, animals, etc.)

Example:

```
- A software system that translates from one language to another
- A software system that drives a vehicle
```

attention

Attention is a technique used in neural networks that allows the model to learn which parts/sequences of the inputs it should pay more attention (focus) to.

B

backpropagation

A common approach/algorithm to perform gradient descent on neural networks.

bag of words

(language)

A way of rendering the value of words within a sequence. The words are stored without reference to the order in the sequence (they are all just put in a bag).

Example: “Cars going fast” and “fast going cars” are represented by the same three words.

baseline

A baseline is a model or the output from a model used to compare the performance of future models. An example might be a simple average of a set of values as a predictor of future values, compared to a linear regression performed on the same values.

batch

(SEE ALSO epoch, hyperparameter)

The examples that are used on one iteration of a model. Batch is often set as a model hyperparameter and defines the number of examples a model will iterate over before updating the model parameters.

batch normalization

Ioffe and Szegedy introduced the idea of batch normalization in 2015[2]. The process normalizes the input or output from a hidden layer, reducing covariate shift and speeding up training on neural networks.

BERT (Bidirectional Encoder Representations from Transformers)

(language) (SEE ALSO Natural Language Processing, attention)

An open-source framework for performing natural language processing developed in 2018[3]. Traditional NLP models read text sequentially in one direction only (left-right, right-left); the significance of this framework is the ability to read in both directions simultaneously (bidirectional). BERT is used as a pre-trained model that can be applied to multiple tasks without training the model from scratch each time.

bias (ethics)

bias (model)

Measures how far off a prediction is from a target/true/correct value

bidirectional

(language) (SEE ALSO BERT) A system that considers text to the left (preceding a target) and to the right (following a target). One such approach is BERT.

binary classification

A model/task that outputs two mutually exclusive results. An example might be a machine that outputs “Anomaly” OR “Steady State” when determining the status of a computer system.

boosting

An approach that iteratively combines a series or set of weak learners (a classifier that is not strongly correlated with a given target) to produce a strong learner.

bounding box

(image recognition)

A bounding box is the coordinates of an area of interest within an image. In an image of faces, for example, bounding boxes may frame the people's faces.

INSERT EXAMPLE OF BOUNDING BOX

bucketing

The process of taking features (usually continuous variables) and grouping them into multiple bins/groups. An example might be age: where 0-5 are grouped together, 6-10, etc.

C

candidate generation

(recommender systems) (SEE ALSO items)

In a recommender system, candidate generation is the first step of the process. The system generates a set of potential items to recommend. The system will generate a subset of all items, using a quick process, and a second step using a more powerful/slower system further reduces those to the final set of items a user sees.

categorical data

SEE ALSO numerical data Features with discrete values.

Example: {Horse, Cow, Pig} are labels for farm animals. They are mutually exclusive (there are no Cow-Horses). Categorical data does not always have to be mutually exclusive (cows may have multiple colors {brown, white, black}

centroid

(clustering) (SEE ALSO K-means)

The center of a cluster.

class

A set of labeled target values.

Example: In a model that predicts if an image is of a cat, the two classes may be {cat, not a cat}

classification

An approach to predict/estimate between different classes.

Example: A model that predicts if a customer will return an item to a store, the classification system may predict one of two classes {will return an item, won't return an item}

class-imbalance (in data)

A problem in a dataset where the classes have very large differences in frequencies.

Example: A system attempting to predict credit card fraud (classes {fraud, not fraud}). The vast majority of transactions are "not fraud," with a very small minority being fraudulent.

clustering

(SEE ALSO unsupervised learning, k-means, hierarchical clustering)

The process of grouping related things together using some measure of similarity or proximity.

collaborative filtering

(recommender systems) (SEE ALSO content-based filtering)

The recommender system makes recommendations to one user based on the interests of other similar users.

Example: If I like the movie "Frozen 2" and "Sing 2" users like me who also liked "Frozen 2" may see a recommendation to watch "Sing 2"

confusion matrix

A matrix/table that summarizes the performance of a classification model. The table illustrates the various types of errors in a classification task (predict true-actually true, predict true-actually false, etc.)

(INSERT EXAMPLE HERE)

content-based filtering

(Recommender systems)

The recommender system uses the features of the items themselves to recommend similar items.

Example: A clothing shop recommends short sleeve shirts in various solid colors to a customer that bought is has in a shopping cart a short sleeve solid blue shirt. The feature set is {short sleeve, solid, blue})

continuous feature

(SEE ALSO discrete feature, numerical features)

A feature or variable with an infinite range of values.

convolution

(image recognition)

A convolution in a machine learning model combines a convolution filter and an input to provide a new value for a pixel.

Example: A convolution filter may be a small matrix that blurs/de-emphasizes sharp contrasts in an image to improve training.

INSERT EXAMPLE OF DIFFERENT CONVOLUTIONS

convolutional filter

(image)

A convolution filter is a matrix used to extract specific features from inputs (images) data.

INSERT BLUR FILTER EXAMPLE HERE

convolutional layer

A later within a neural network that uses a convolutional filter.

convolutional neural network

A neural network in which there is at least one convolutional layer.

cross-validation

An approach used to estimate how well a model may generalize to new data. Cross-validation trains a model on a subset of data that is independent of the training data. In many cases, this training slice/test slice is repeated on various subsets of the data.

INSERT CROSS-VALIDATION IMAGE HERE

D

data augmentation

Synthetically increasing the number of training examples by adjusting or transforming current training samples.

Example: Images of fruit and vegetables can be augmented through rotation, turning an image of an apple upside down. This increases the number of training examples (from 1 to 2) and provides additional information to a learning algorithm (images labeled with class 'apple' may be stem up, stem down)

DataFrame

Language within the popular Python library/package ‘pandas’[4] to describe a table of data.

data parallelism

(SEE ALSO model parallelism)

An approach to run a model on multiple machines concurrently (parallel). The model is passed to every device, and a subset of the input data is sent to each device to run. This allows training on very large data inputs.

data set or dataset

A set of examples put in one place.

decision tree

A model with one or more if-then branching statements to separate the data. In a prediction context, this separation allows for new (unseen) data to be classified.

INSERT IMAGE OF DECISION TREE

deep learning

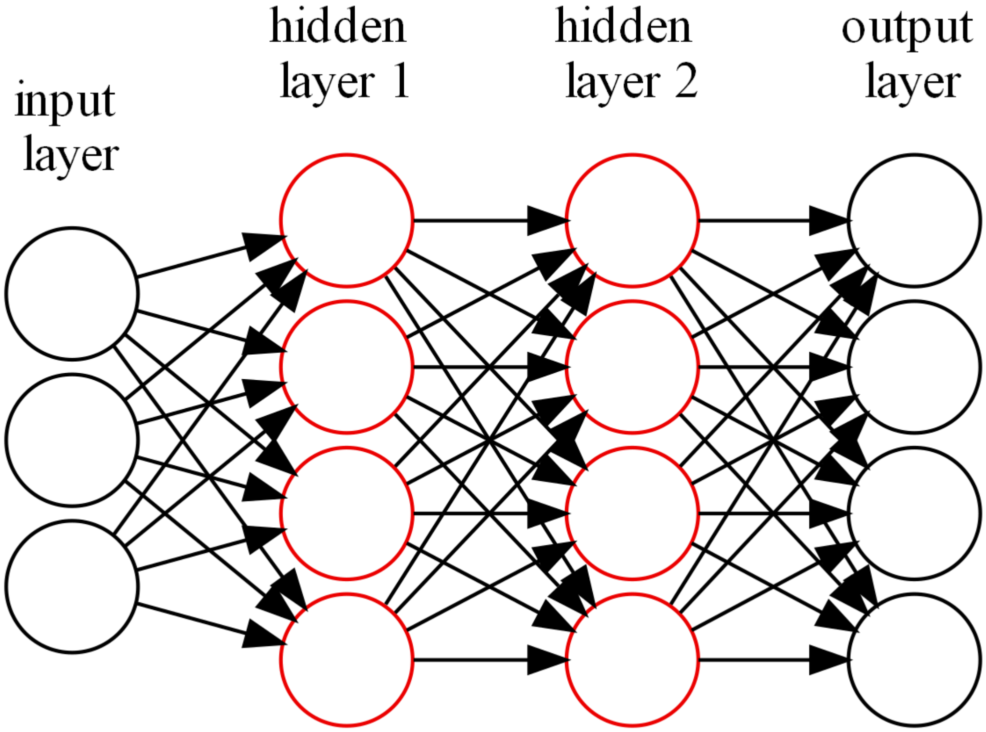
(SEE ALSO deep model)

A neural network learning on a deep model. Deep models have more than one hidden layer.

deep model

(SEE also deep learning)

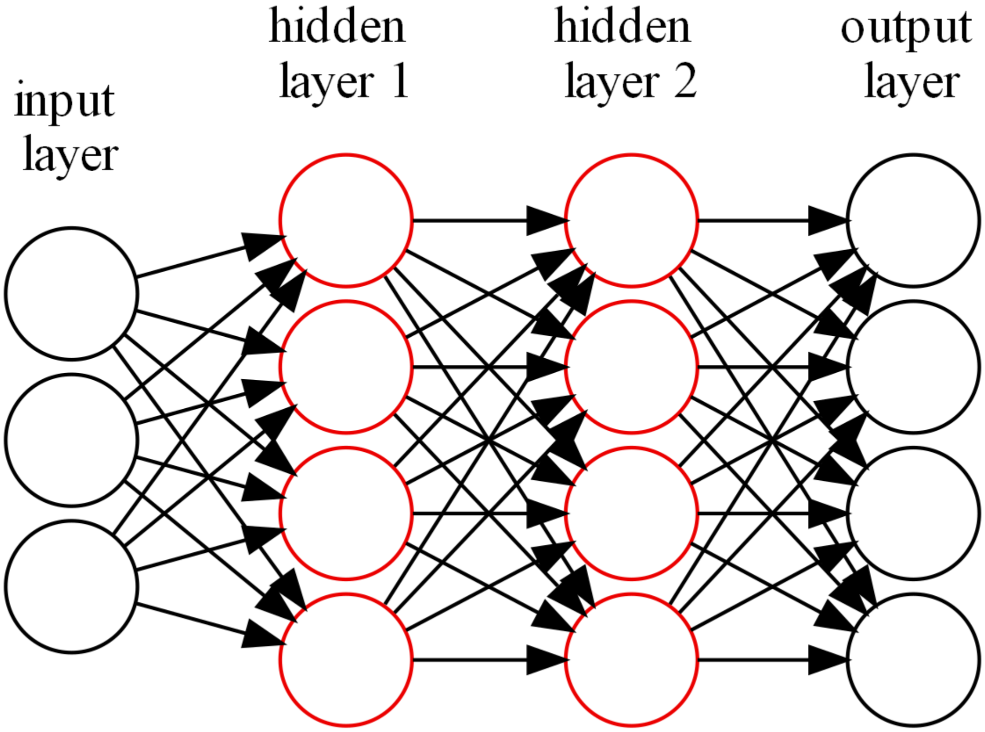
A neural network containing more than one hidden layer.



A simple deep learning network with two hidden layers

dense layer

A fully connected layer in a neural network. This means that each node in the layer is connected to each node in the following layer. This may sometimes be referred to as a "fully connected" layer. If there are n inputs in the layer and y outputs, the connections have n*y weights to estimate.



Each layer is densely connected to (has an arrow to) the following layer

depth

The number of layers in a neural network that learn. Example: A network with 1 input layer, 2 hidden layers, and 1 output layer would have a depth of 3.

dimension reduction

Technique(s) that decrease/reduce the number of features in a model input.

discrete feature

(SEE ALSO continuous feature)

A feature with a finite set of values. Example: cat, dog, mouse.

dropout regularization

A technique to help reduce over-fitting/overspecialization in a type of input. Dropout regularization works by randomly dropping a random number of neurons (ignoring them). This is done to improve a model's generalizability and prevent over-fitting.[5]

E

early stopping

Early stopping is an approach to regularization that ends the training of a model prior to the point at which generalization error increases. Training is usually stopped before training loss has finished decreasing, and as test data (validation data) error starts to increase.

embedding

(SEE ALSO one-hot-encoding)

An embedding is a low-dimensional continuous vector representation of a categorical feature. They are used to translate high-dimensional features into a low-dimensional space.

Example: When encoding written text (a corpus), an embedding is learned during training, and when complete, it roughly describes relationships between words based on the direction of the word vectors.

ensemble

An ensemble is a combination of the predictions of more than one model. The goal of an ensemble is to correct for deficiencies in a single models structure, hyper-parameters, etc., by combining it with another/many models.

environment

In reinforcement learning, software agents take actions within an environment to learn that environment. The output from these actions comes in the form of a reward. The software agents goal is to learn (use) actions in such a way as to maximize the long-term rewards. The space or world that contains the agent is the environment.

Example: Imagine a machine that is playing a video game; the various controllers are the actions the software agent uses to control something in the game; the environment is the game world. The rewards are whatever metrics the world feedbacks (points etc.)

episode

In reinforcement learning, software agents take actions within an environment to learn that environment. The output from these actions comes in the form of a reward. The agent learns by repeating actions with small variations in an attempt to maximize the reward. Each of these repeated attempts is called an episode.

epoch

(SEE ALSO batch)

One full iteration over a dataset, which each example being seen by model 1 time is called an epoch.

example

(SEE ALSO labeled example, unlabeled example)

An example is a single row of a dataset, or a single record. Each example could contain multiple features, in tables represented as columns. If the example comes from a supervised learning data set, it will also contain a label.

Extract-Transform-Load (ETL)

A process for moving or copying data between sources (extract), processing and cleaning the data (transform), and placing it in a target system (load) such as a data warehouse.

F

feature

An input variable that is used in models. If considering a dataFrame or data table, a single row might contain examples; each column of the table may contain features related to that example.

Example: If our table contained information about cars, each row would represent a different car. The columns might be {make, model, year, type}

feature engineering

A process that seeks to determine which features should be included in a model. This process often includes transforming and combining raw input features into new features.

Example: The first three digits of a cell phone number by themselves do not reveal much about an individual, but given that most cell phone users have not changed their number in several years, a model could extract where a person has lived in the past using these three values. This would be an engineered feature.

federated learning

A machine learning model trained using decentralized training sets. In federated learning, a remote system receives a model from a coordinating system; the remote system trains the model locally, then uploads model improvements but not the training data back to the coordinating system. The coordinating system aggregates each remote systems model to create a new global model.

feedback loop

Feedback loops occur when the output from a model, such as a prediction model makes changes or influences data for subsequent runs of that model or new model. A fairly easy-to-understand example would be recommendations on what to watch next on a system such as YouTube. A user’s current viewing behavior leads to recommendations of similar videos, which, when watched, reinforce the genre and lead to more recommendations that are similar.

fully-connected layer

(SEE hidden layer)

G

GAN

(generative adversarial network)

generalization

A model’s ability to make predictions on new, previously unseen data.

generative adversarial network (GAN)

Architecture using two neural networks. The first network generates/creates data (that is the “generative” part of the GAN), and a discriminator validates the data (that is the “adversarial” part of GAN). The method is used to generate synthetic instances of data that closely mimic real data.

generative model

A model that generates new examples of data from a training dataset or determines whether or not a new example comes from a given data set.

gradient descent

An optimization process that iteratively adjusts parameters to find a combination of weights and bias that minimize a model’s loss function.

greedy policy

In reinforcement learning, an agent always chooses the action that has the highest expected return.

ground truth

Refers to the true value of something. The correct output.

H

hidden layer

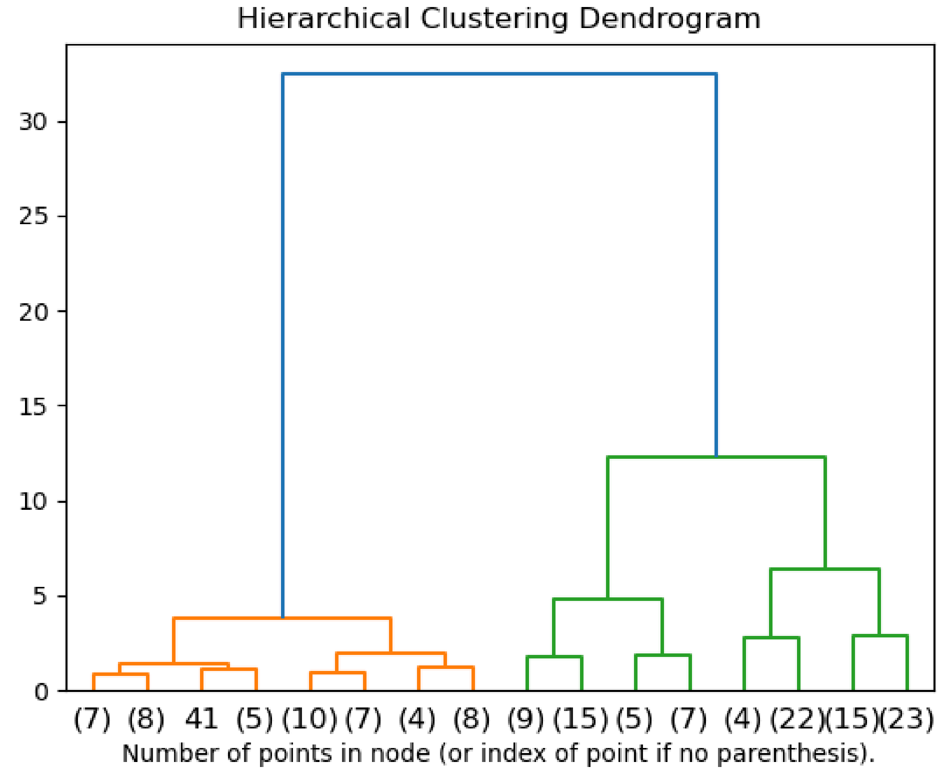
(SEE ALSO ReLU, sigmoid)

A hidden layer in a neural network sits between the input layer and the output layer. Both inputs and outputs can be seen; the layers in the middle of the model are not, hence the term. Hidden layers are used to perform some action on the layer that provides the input, such as applying a non-linear activation function (ReLU, Sigmoid, etc.) on a set of inputs. A deep neural network is called such because it has more than one hidden layer.

hierarchical clustering

(clustering) (SEE ALSO agglomerative clustering)

Hierarchical clustering generates a tree-like structure of clusters. Agglomerative clustering, a type of Hierarchical Clustering algorithm, iteratively joins clusters from the bottom up to a hierarchical tree. Divisive clustering creates a single cluster and iteratively divides from the top down.



An example of agglomerative clustering from the sklearn documentation[1]

holdout data

Examples not used in training a model. Test datasets are an example of held-out data and used to understand how a model will generalize to new data the model has not used in training.

hyperparameter

A hyperparameter is a value used to control the training process of a model. It can be thought of as a dial that can be turned/tweaked during multiple runs of a model.

I

image recognition

(SEE image classification)

A process that classifies parts of an image. A system that classifies certain objects (people, cars, etc.) patterns (shapes, etc.) from image data.

imbalanced dataset

SEE class-imbalanced dataset.

input layer

The first layer, with the input data, of a neural network. Note that the input layer does not have any activation functions and is not counted when considering model depth.

interpretability

This definition is not well accepted. The ability of a human to understand a model's reasoning or output.

items

(recommender systems)

The items in a recommender system are the things that the system recommends.

K

Keras

A python API popular when running on TensorFlow (tf.keras).[6]

k-means

(clustering, unsupervised learning)

A clustering approach that iteratively determines the best k center points (called centroids) and then assigns to a cluster the data point(s) nearest that center. The data points with the same center attachment are considered a single cluster. The algorithm groups examples into k (a value given as a hyperparameter) groups, with each group having about equal variance.

k-median

(clustering)

A clustering algorithm like k-means. The notable difference is that k-means minimizing the sum of the squares of the distance between a potential centroid and its example points, and k-medians minimizes the sum of the distance between a potential centroid and its example points. between a centroid candidate and each of its examples

label

(SEE ALSO supervised learning)

A label is the correct result or answer for a given example in the data. In a labeled data set, supervised learning, each example has a label and one or more features used to estimate that label.

layer

A set of neurons in a neural network that acts as inputs (input layer), hidden layers, or an output layer.

learning rate

A hyperparameter that determines the step size for each iteration of a model as it searches for a minimum of a loss/cost function. If the learning rate is too high, the model may jump the minima, too low, and it may take too long to converge.

linear model

In a linear model: a single weight is assigned to each feature + a bias term to make a prediction.

Linear regression and logistic regression are two popular linear model types. A strength of linear models is interpretability, as an end-user can easily look at the weights a model assigned on the way to a prediction.

linear regression

A type of linear model and a supervised (labeled data) machine learning approach.

logistic regression

A type of linear model. In logistic regression, a sigmoid function converts a linear model into a value between 0 and 1.

loss

A metric that relates to how far a given model predicted outcome is from its label (or truth).

loss curve

A visualization of loss as it relates to iterations on training data.

M

machine learning

TBD

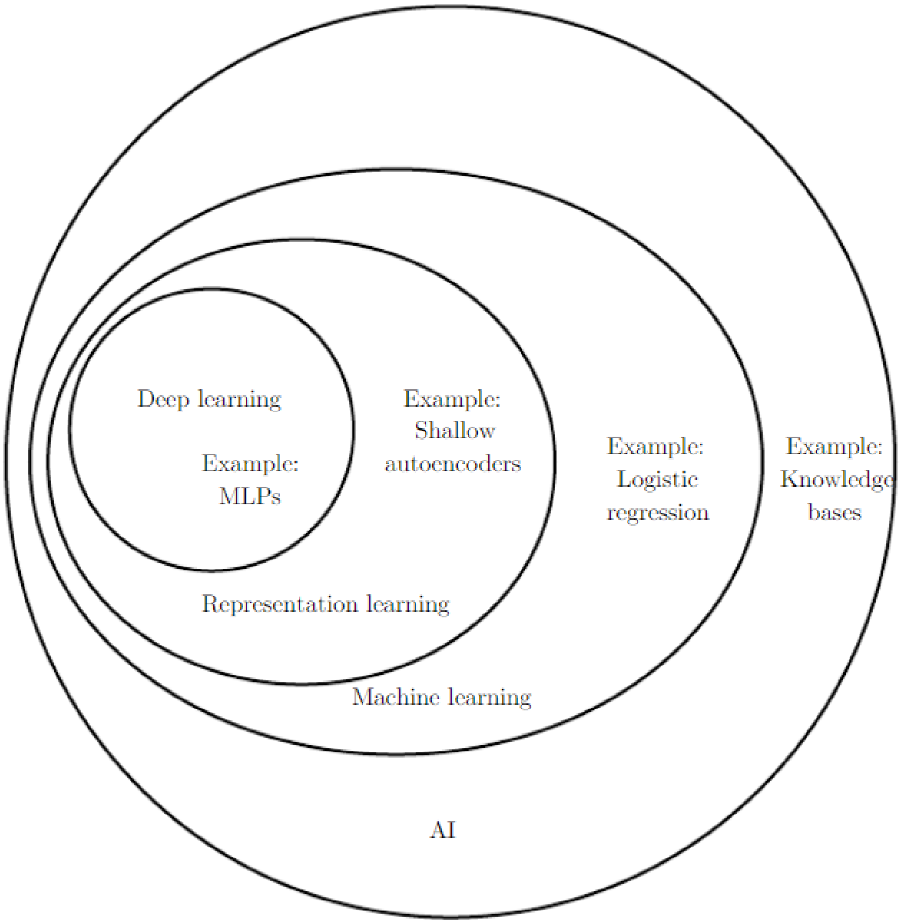


Figure 1.4 from the introduction of Deep Learning[7]

majority class

(SEE ALSO class imbalance, minority class)

The majority class is the more common class in an imbalanced data set. Example: In a dataset containing information on education level, 95% High school graduates, and 5% college graduates, the 95% high school graduates are the majority class.

matplotlib

A library/package within Python[8]. The package is used to create visualizations (charts, graphs, and animations). Many other graphing libraries are built on top of matplotlib.

Mean Absolute Error (MAE)

A common error metric calculated by taking an average of the absolute errors. MAE is defined as:

$$MAE\left(y, \hat{y}\right) = \frac{\sum_{i=0}^{N-1} \left|y_i - \hat{y}_i\right|}{N}$$

Mean Squared Error (MSE)

A common error metric calculated by dividing the squared loss by the number of examples. MSE tells you the average squared difference between the observed and predicted values.

$$MSE\left(y, \hat{y}\right) = \frac{\sum_{i=0}^{N-1} \left(y_i - \hat{y}_i\right)^2}{N}$$

metric

A metric is a value/number that is of interest to the person creating the model, or the system itself. It is not always the value the system attempts to optimize.

minority class

(SEE ALSO class imbalance, majority class)

The minority class is the less common class in an imbalanced date set. Example: In a dataset containing information on education level, 95% High school graduates, and 5% college graduates, the 5% college graduates are the minority class.

MNIST

(image recognition)

A public-domain dataset first introduced by LeCun, Cortes, and Burges[9]. MNIST contains 60,000 images of a handwritten number between 0 and 9. Each image is stored as a 28 x 28 array, where each value is a grayscale value taken from the image.

INSERT MNIST IMAGE HERE

modality

(SEE ALSO multimodal models)

A category of data such as text/corpus, images, video, numerical, etc.

model

There are different definitions of this term, and it is used loosely across this domain. In general, the model can be thought of as what has been learned by a system. In a linear regression, this is the weights + bias etc.

model parallelism

(SEE ALSO data parallelism)

An approach to run a model on multiple machines concurrently (parallel). The model is passed to every device, and a subset of the input data is sent to each device to run. This allows training on very large data inputs.

model training

A system using training data to find the best model (a learned representation of the training data).

multimodal model

A model where the data inputs and/or outputs have more than one modality.

In the example below, an image of a cat is seen as model input, and the caption for the image in the text provides information for a machine learning model.

INSERT MULTIMODAL MODEL EXAMPLE

N

negative class

SEE ALSO positive class

In a binary classification model, one of the two classes is termed negative, the other being positive. Example: In an image recognition problem, if the model is supposed to identify cats, the positive class might be "it is a cat" and the negative class "not a cat."

neural network

A model that, originally inspired by how the brain works, with multiple layers (input, hidden, and output).

neuron

A single node in a neural network. Each neuron takes in multiple inputs from the layer that precedes it and generates a single output. A typical example may be a neuron that takes multiple inputs and transforms them using an activation function (such as reLu, sigmoid).

N-gram

(SEE ALSO bag of words)

An N-gram is a continuous sequence of words or sometimes tokens in a sample of text. N-grams differ from bag-of-words in that N-grams have an order. The "N" in n-grams is an integer starting at 1 and defines how many words are in each token.

```
-uni-gram (N=1) This | is | one | word | per
-bi-gram (N=2) This is | two words
-tri-gram (n=3) this is three | words in each
```

node (neural network)

A neuron in a layer of a neural network.

INSERT IMAGE WITH LABEL FOR NEURON

noise

Noise in a dataset is anything that obfuscates the true model. Noise is often found in "real world" data sets, especially those generated from non-systematic/algorithmic processes (like human activity, etc.)

normalization

A technique to convert values into a standardized range.

See also scaling.

numerical data

(SEE ALSO continuous features)

Features in a data set represented as numbers (integers, floats, etc.) Numerical features are sometimes called continuous features, but not all numerical features are continuous.

NumPy

(SEE ALSO pandas)

A package/library in python[10] that provides array operations.

O

objective

The goal of a given system. Generally a metric the system is attempting to optimize.

objective function

The metric the system is attempting optimize.

one-hot encoding

(SEE ALSO embedding)

Also referred to (outside of this domain) as dummy coding. One hot encoding is a sparse vector in which one element of the vector is set to 1 (on-hot), and all others are 0. One hot encoding is often used to encode strings or factors that have a fixed number of values. The resulting vector can be very large.

outlier detection

(SEE ALSO outliers)

A system, process, or goal that attempts to identify outliers in a set of data.

outliers

Examples outside what a user or system may consider the normal range of examples. For example: in a set of numbers with 2000 examples, with 999 examples ranging between 1 and 4, a value of 400 would be considered an outlier.

output layer

The last layer in a neural network. The output layer contains the model solution/answers.

overfitting, overfit

(SEE ALSO underfitting, generalization)

A model is overfitting when it is mapped so well to the training data that it begins to map noise or random error, as opposed to the true relationships. Overfitting a model results in a model that fails to generalize well to new data.

P

pandas

(see also NumPy)

A package within Python[4]. Pandas DataFrame is a column-oriented structure which looks and feels familiar to most users. Pandas is built on top of the popular NumPy library.

parameter

(SEE ALSO hyperparameters) Parameters are variables the model trains in a system. This is in contrast with hyper-parameters, which are determined outside of a model run. For example: the weights in linear regressions are a parameters.

perceptron

An algorithm used to model binary classifiers. A perceptron is a single-layer neural network. With an input, weights, weighted sum, and an activation function.

INSERT IMAGE OF PERCEPTRON

pipeline

A pipeline generally refers to the processing or structure to complete the processing of a given model. It can, but does not necessarily require, steps to perform data gathering, cleaning, normalization, training, and production of the results.

policy

(reinforcement learning)

An agent's strategy, way of behaving, or action taken to accomplish a task.

positive class

(SEE ALSO negative class)

In a binary classification model one of the two classes is termed negative, the other being positive. Example: In an image recognition problem if the model is supposed to identify cats, the positive class might be "it is a cat" and the negative class "not a cat".

Contrast with negative class.

precision

Used in classification models. Precision illustrates the frequency with which a classification model correctly predicts the positive class. In other words, the model guesses the image contains a cat, and the image contains a cat.

pre-trained model

A model that has already been trained. EXAMPLE: BERT is a pre-trained model, trained originally on a corpus of text from wikipedia and an online bookcorpus. The model can then be used on other language tasks without the need for those latter systems to run the original training data again.

Python

A popular programming language that can be used for machine learning.[11]

R

R

A popular statistical program that can be used for machine learning.[12]

R-Studio

A popular IDE for the programming language R.[13]

random-forest

A random-forest is an ensemble created by using many small decision trees built with a (random) selection of available features and evaluating the average of the many trees (forest)

rater

(SEE ALSO ground truth)

A (usually) human who provides labeled data. Raters may evaluate an image, for example, and determine if it contains certain features. These labels are usually sometimes considered to be the "ground truth," although additional validation steps should be performed before trusting the output.

recommendation system

(recommender system) (SEE ALSO collaborative filtering, content-based filtering)

A system that recommends a subset of items to an individual from a larger available pool of items. An example might be a system that recommends books to read on a website with 50,000 books available.

Recommenders take many forms, but two common approaches are content-based filtering and collaborative filtering.

recurrent neural network

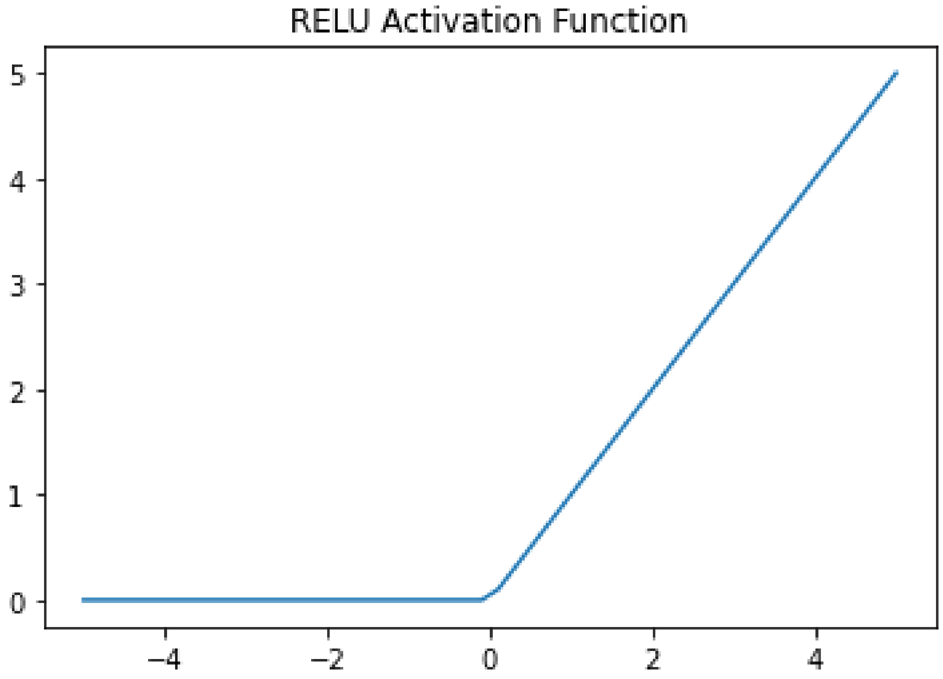
A neural network that is run more than once and outputs from the prior run provide information to the next run of the model. Hidden layers learn from previous model runs.

reinforcement learning (RL)

In reinforcement learning, software agents take actions within an environment. The output from these actions comes in the form of a reward. The software agents goal is to learn (use) actions in such a way as to maximize the long-term rewards. Reinforcement agents learn by repeating (SEE episodes) actions and evaluating which actions lead to the highest reward. There are several recent examples of this in board games (GO etc.). Reinforcement agents can learn extremely complex environments.

ReLU

In neural networks, a function that takes a weighted sum of inputs from the prior layer and generates (through its function) output forward to the next layer. ReLU functions are non-linear



Example of a RELU activation function

re-ranking

In large-scale (many items) recommender systems, items can be re-ranked as a last step prior to an item subset being shown to a user. This is usually done to correct for factors outside the recommender model, such as an item that has been purchased in the past, an item with a low contribution margin to the seller etc.

reward

An agent seeks to find the best policy; the best policy is the one that maximizes some value to the agent. This may be something like the lowest cost, the highest accuracy, etc.

Root Mean Square Error

$$RMSE\left(y,\hat{y}\right)=\sqrt{\frac{\sum_{i=0}^{N-1}\left(y_i-\hat{y}_i\right)^2}{N}}$$

S

scaling

(SEE ALSO normalization)

A technique to convert values into a standard format/range.

scikit-learn

(programming languages)

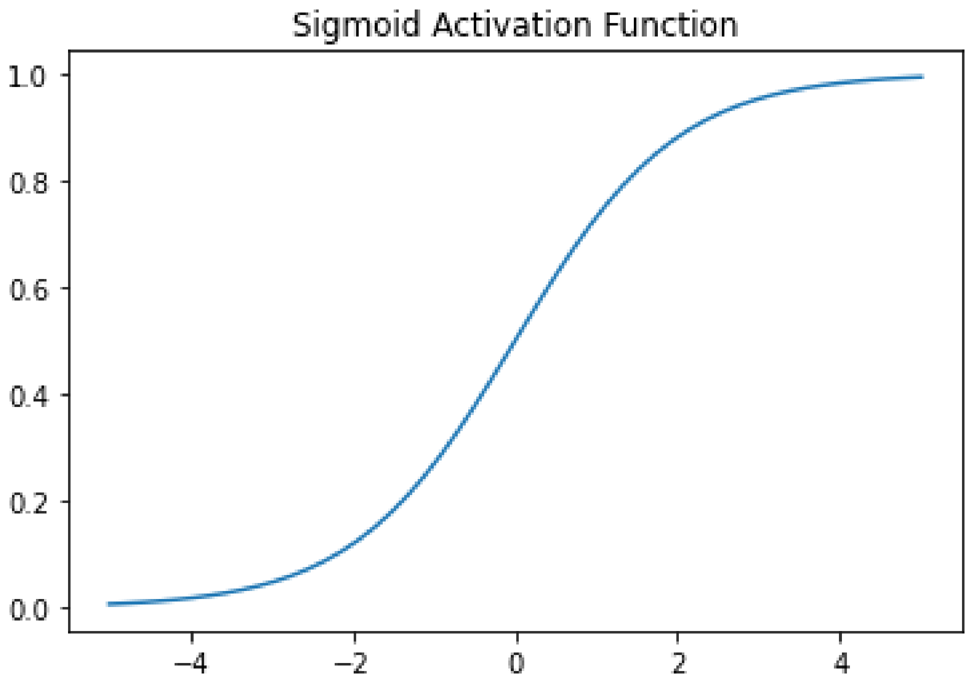
A popular open-source machine learning package within python.[14]

scoring

In a recommender system items are scored. This score becomes a value used to rank/order the items as part of candidate generation.

sigmoid function

In neural networks, a function that takes a weighted sum of inputs from the prior layer, and generates (through its function) output forward to the next layer. Sigmoid functions are non-linear. The sigmoid function maps the inputs to probabilities (values between 0, 1).



Example of a SIGMOID activation function

similarity measure

The metric(s) used in clustering to determine how similar/alike any two given examples in the data are.

sparse feature

A feature vector with mostly 0 or empty values. An example might be a feature engineered from a one-hot encoding algorithm, where 99999 values are 0 and a single value is 1.

state

A parameter in reinforcement learning that describes the current environment.

stride

In a convolution, a stride describes how much each convolution filter moves from one slice to the next across an image.

Example: in a 2d space, a stride of 1,1 moves one position to the right at each step. When a stride reaches the edge (right edge) it resets to the far left, and 1 position down. This description does not account for padding on convolutions.

INSERT GRAPHIC EXAMPLE OF STRIDE

supervised machine learning

(SEE ALSO unsupervised learning)

Supervised learning refers to models trained on labeled data. This is data that has an answer within the data (each example is labeled with an answer). Contrast with unsupervised learning, where the data is not labeled with an answer.

T

target

(SEE label)

TensorFlow

(programming language)

A machine learning platform design by Google, and open sourced[15]

test set

(SEE ALSO training set, validation set)

A subset of data that is held out of the training phase of a machine learning model. The test set is used to validate a model (although occasionally another separate validation is used for this purpose), and to learn how well the model performs on data it has not seen.

token

(language) (SEE ALSO bag of words, N-grams) Typically used to describe a unit within a language model/system that the model trains and predicts on. In a three word sentence, “The window broke” tokens might be “the” “window” “broke”. Tokens can be more complex than this simple example.

Tokens are sometimes used outside of language models, but their use is not as frequent.

training set

(SEE ALSO validation, test set)

A training set is used to train a model on data. this data is held out from the validation and test data sets. Separating data into multiple sets and holding some out for validation and testing helps control for and reduce overfitting.

U

underfitting

(SEE ALSO overfitting)

Overfitting is training a model that captures the noise and randomness in a training set. Underfitting is training a model that does not capture the true complexity in the relationship(s) between features.

unidirectional, unidirectional language models

(language models)

In a language model, an approach that evaluates the text preceding it in a corpus. Bidirectional models evaluate the preceding and following text.

unsupervised machine learning

(example: clustering, kmeans, supervised learning)

Generally speaking unsupervised methods seek to find relationships and patterns in a data set that is unlabeled. One of the most common approaches is clustering.

V

validation set

A subset of the data held out from training, and used to perform validation of a models performance.

W

weight

A value for a feature in a model. In linear regression the weight is the coefficient for each variable.

References

1. Plot Hierarchical Clustering Dendrogram. *scikit-learn*.

2. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]* (2015).

3. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (2019).

4. Pandas - Python Data Analysis Library.

5. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).

6. Keras: The Python deep learning API.

7. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).

8. Matplotlib — Visualization with Python.

9. MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges.

10. NumPy.

11. Python.org. *Python.org*.

12. R: The R Project for Statistical Computing.

13. RStudio Open source & professional software for data science teams.

14. Scikit-learn: Machine learning in Python — scikit-learn 1.1.dev0 documentation.

15. TensorFlow.