

Glossary

Last modified: 2026-10-11

This page explains the terms the course uses, in plain words. Each entry says what the term means, gives a concrete picture of it, shows how it is written in symbols, and links to the related terms and to the section of the notes that defines it. The page lists 43 terms. Homework sections link their terms to the same sources.

The entries are in the order in which the course meets them.

1 supervised learning

Learning a rule from examples that come with the right answer.

In a picture: A spam filter trained on emails that people have already marked spam or not spam.

See also: [unsupervised learning](#), [regression](#), [classification](#)

Defined in: the notes

2 unsupervised learning

Finding structure in data when no answers are given.

In a picture: Grouping genes by expression pattern with no groups known in advance.

See also: [supervised learning](#)

Defined in: the notes

3 regression

A supervised problem where the answer is a number.

In a picture: Predicting tomorrow's bicycle rentals from the weather.

See also: [supervised learning](#), [classification](#)

Defined in: the notes

4 classification

A supervised problem where the answer is one of a fixed set of categories.

In a picture: Deciding whether a tumor is malignant or benign.

See also: [supervised learning](#), [regression](#)

Defined in: the notes

5 inference

Using data to learn how the world works (which inputs matter and by how much) rather than only to predict.

In a picture: Asking how much each extra year of age changes the risk of severe influenza.

See also: [prediction](#)

Defined in: the notes

6 prediction

Using a fitted model to guess the answer for a new input.

In a picture: Flagging, at admission, the patients likely to be readmitted.

See also: [inference](#)

Defined in: the notes

7 feature

One measured input to a model; also called a predictor or covariate.

In a picture: Income, loan amount and years of employment for a loan applicant.

In symbols: x_j

See also: [response](#), [parameter](#)

Defined in: the notes

8 response

The quantity a supervised model is trained to predict; also called the target or label.

In a picture: Whether the loan was repaid.

In symbols: y

See also: [feature](#)

Defined in: the notes

9 parameter

A number the learning algorithm chooses from the data; the model's knobs.

In a picture: The slope and intercept of a fitted line.

In symbols: β_j

See also: [feature](#), [intercept](#), [slope](#)

Defined in: the notes

10 intercept

The model's output when every input is zero.

In a picture: Where a fitted line crosses the vertical axis.

In symbols: β_0

See also: [slope](#), [affine function](#)

Defined in: the notes

11 slope

How much the output changes when one input goes up by one and the others stay fixed.

In a picture: A line that rises two units for each unit across has slope 2.

In symbols: β_j

See also: [intercept](#), [affine function](#)

Defined in: the notes

12 affine function

A constant plus a multiple of each input. It need not pass through the origin; with intercept zero it is a linear function.

In a picture: A line shifted up by its intercept so it misses the origin.

In symbols: $\beta_0 + \vec{\beta} \cdot \vec{x}$

See also: [intercept](#), [slope](#), [linear function](#), [linear model](#), [hyperplane](#)

Defined in: the notes

13 linear function

An affine function whose intercept is zero. Doubling the input doubles the output.

In a picture: A line through the origin.

In symbols: $\vec{\beta} \cdot \vec{x}$

See also: [affine function](#), [intercept](#), [slope](#)

Defined in: the notes

14 linear model

A model whose prediction is an affine function of the features.

In a picture: Predicting house price as a base amount plus a fixed amount per square foot.

See also: [affine function](#), [design matrix](#)

Defined in: the notes

15 design matrix

The matrix whose rows are the observations and whose columns are the features (with a column of ones for the intercept).

In a picture: A table of data with one extra column of ones, treated as a matrix.

In symbols: D

See also: [linear model](#), [intercept trick \(bias trick; 1-padding\)](#)

Defined in: the notes

16 intercept trick (bias trick; 1-padding)

Putting a constant 1 in front of each input vector so the intercept becomes one more coefficient.

In a picture: Turning (x_1, x_2) into $(1, x_1, x_2)$.

See also: [design matrix](#), [intercept](#)

Defined in: the notes

17 loss function

A number saying how bad one prediction is.

In a picture: The squared gap between a prediction and the true answer.

In symbols: $L(h(x), y)$

See also: [empirical risk](#)

Defined in: the notes

18 empirical risk

The average loss over a set of labelled data, such as the training set or a test set.

In a picture: The mean squared error of a fitted line on the points it is scored on.

In symbols: $\hat{R}$

See also: [loss function](#), [training error](#), [test set](#)

Defined in: the notes

19 training error

The empirical risk on the training data.

In a picture: The mean squared error of a fitted line on the points it was fitted to.

See also: [empirical risk](#), [overfitting](#)

Defined in: the notes

20 overfitting

A model that matches the training data so closely that it does worse on new data.

In a picture: A wiggly curve that passes through every training point and misses new ones.

See also: [underfitting](#), [generalization](#)

Defined in: the notes

21 underfitting

A model too simple to capture the pattern even in the training data.

In a picture: A straight line through data that clearly bends.

See also: [overfitting](#)

Defined in: the notes

22 generalization

How well a model does on data it was not trained on.

In a picture: The error on a held-out test set.

See also: [overfitting](#), [validation set](#)

Defined in: the notes

23 validation set

Data held out of training and used to choose among models.

In a picture: Picking the polynomial degree with the lowest error on held-out data.

See also: [test set](#), [cross-validation](#)

Defined in: the notes

24 validation error

The empirical risk on the validation set.

In a picture: The mean squared error of each candidate polynomial on the cars held out to choose among them.

See also: [validation set](#), [empirical risk](#)

Defined in: the notes

25 test set

Data held out until the end and used once, to report the chosen model's error.

In a picture: A final exam that is taken once.

See also: [validation set](#)

Defined in: the notes

26 test error

The empirical risk on the test set.

In a picture: The mean squared error of the chosen model on the cars kept back until the end.

See also: [test set](#), [empirical risk](#)

Defined in: the notes

27 cross-validation

Rotating which part of the data is held out so every observation is used for both training and checking.

In a picture: 5-fold: train on four fifths, check on the fifth, five times.

See also: [validation set](#)

Defined in: the notes

28 gradient

The vector of partial derivatives; it points in the direction of steepest increase.

In a picture: The uphill direction on a hillside.

In symbols: ∇C

See also: [gradient descent](#)

Defined in: the notes

29 gradient descent

Minimizing a function by repeatedly stepping opposite to its gradient.

In a picture: Walking downhill in small steps.

In symbols: $\vec{\beta}_{t+1} = \vec{\beta}_t - h \nabla C(\vec{\beta}_t)$

See also: [gradient](#), [step size \(learning rate\)](#)

Defined in: the notes

30 step size (learning rate)

How far each gradient descent step moves.

In a picture: Too small crawls; too large overshoots the valley.

In symbols: h

See also: [gradient descent](#)

Defined in: the notes

31 odds and log-odds

Odds are $p/(1-p)$; log-odds are their logarithm. A logistic model is linear in the log-odds.

In a picture: Probability 0.8 is odds 4 to 1.

In symbols: $\ln \frac{p}{1-p}$

See also: [sigmoid](#), [decision boundary](#)

Defined in: the notes

32 sigmoid

The S-shaped function that turns any number into a probability between 0 and 1.

In a picture: A score of 0 maps to probability 0.5.

In symbols: $\text{expit}(z)$

See also: [odds and log-odds](#), [decision boundary](#)

Defined in: the notes

33 decision boundary

The set of inputs where a classifier switches from one class to the other.

In a picture: The line in a two-feature plot with spam on one side and email on the other.

See also: [sigmoid](#), [decision threshold](#)

Defined in: the notes

34 cross-entropy loss

The loss for predicted probabilities: large when the model was confident and wrong.

In a picture: Predicting 0.01 for something that happened costs far more than predicting 0.4.

See also: [loss function](#), [sigmoid](#)

Defined in: the notes

35 confusion matrix

A table of how many cases of each true class received each predicted class.

In a picture: Counts of true positives, false positives, false negatives and true negatives.

See also: [precision](#), [recall \(sensitivity\)](#), [prevalence](#)

Defined in: the notes

36 precision

Of the cases the model flagged positive, the fraction that really are positive.

In a picture: Of 140 flagged patients, 90 have the disease.

See also: [recall \(sensitivity\)](#), [confusion matrix](#)

Defined in: the notes

37 recall (sensitivity)

Of the cases that really are positive, the fraction the model flagged.

In a picture: Of 100 patients with the disease, 90 were flagged.

See also: [precision](#), [confusion matrix](#)

Defined in: the notes

38 prevalence

The fraction of cases that are truly positive.

In a picture: 0.1% of network connections are intrusions.

See also: [confusion matrix](#), [precision](#)

Defined in: the notes

39 bias, variance and residual variance

Three sources of prediction error: a model that is systematically off (bias); one that changes a lot from sample to sample (variance); and the residual variance, the noise variance no model on the same inputs can remove.

In a picture: A straight line misses a curve (bias); a wiggly curve changes with each sample (variance).

See also: [overfitting](#), [underfitting](#)

Defined in: the notes

40 ridge regression

Least squares with a penalty on the size of the coefficients.

In a picture: Shrinking coefficients toward zero to steady a model with many features.

In symbols: $\lambda \|\vec{\beta}\|^2$

See also: [loss function](#), [overfitting](#)

Defined in: the notes

41 null space

The set of vectors a matrix sends to zero.

In a picture: For one row (1, -2, 3), the plane of vectors perpendicular to it.

See also: [hyperplane](#)

Defined in: the notes¹

42 hyperplane

A flat surface one dimension smaller than the space it sits in: a line in 2-D, a plane in 3-D.

In a picture: The decision boundary of a linear classifier.

See also: [affine function](#), [decision boundary](#)

Defined in: the notes

43 decision threshold

The predicted probability above which a classifier answers positive; 0.5 unless there is a reason to move it.

In a picture: Raising it from 0.5 to 0.9 flags fewer cases but fewer false alarms.

See also: [decision boundary](#), [precision](#), [recall \(sensitivity\)](#)

Defined in: the notes

¹<https://morrison-lab.github.io/mds/linear-algebra-rank-nullity.html#def-null-space>