# Week 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.