Week 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.