Glossary of machine learning jargon in bioinformatics

Foundational Terms

Machine learning (ML) - In general, a poorly-defined term. Here, it is used to refer to the field of statistics that involves training any of a large family of models in order to best find, classify, or predict patterns in data based on a carefully selected set of assumptions.

Supervised learning - ML approaches that explicitly associate observations' input variables (often called "predictors") with outcome variables such as survival or treatment response.

Unsupervised learning - ML approaches that organize observations - either into groups or along a continuum - based solely on their input features and without access to outcome labels of any kind.

Dimensionality Reduction

A family of analytical approaches in which high-dimensional data are embedded in lower-dimensional space. Commonly used to visualize data with greater than 3 dimensions.

Factor loadings - The correlations between each of a dataset's original variables and each of its principal components (PCs); equivalently, the projection of each original variable onto each PC. Can be useful for interpreting the information represented by each PC.

Nonlinearity - In most cases, a dataset is described as "nonlinear" when its variables are not well-described as a linear function of its other variables. A dataset may also be described as "nonlinear" when an *external* outcome variable related to the dataset is not easily computed as a linear function of its variables. Often, this complexity is a result of interaction terms between variables and/or higher-order polynomial/power-law relationships.

Singular Value Decomposition (SVD) - A matrix factorization method that, when computed, can be used to find the best low-dimensional representation for a dataset under the assumption that the data's variables have linear relationships with one another. Can be calculated rapidly for even very large matrices (i.e., those that represent very large datasets).

Clustering

A type of unsupervised learning in which observations are placed into groups such that similar observations are grouped together and dissimilar observations are not. Often applied to single-cell data to detect distinct cellular subpopulations.

Jaccard Coefficient - A measure of how "connected" two nodes in a graph are to one another. Specifically, the Jaccard Coefficient is the number of neighbors that two nodes share divided by the total number of neighbors of both nodes. For two nodes with sets of neighbors A and B, the Jaccard Coefficient is given by $\frac{|A \cap B|}{|A \cup B|}$.

Minimum-Spanning Tree (MST) - A *spanning tree* is a type of graph in which all nodes are connected to one another without any loops or breaks in continuity. A *minimum spanning tree*, then, is the spanning tree with the smallest total edge length of all possible spanning trees for a given set of nodes.

Self-organizing map (SOM) - A type of artificial neural network capable of dimensionality reduction and clustering of high-dimensional data. Computed using a series of recursive, non-parametric regression computational steps.

Prediction and Correlative Biology

Generalized Linear Models (GLMs) - A class of statistical models in which linear combinations of input variables are transformed via a "link function" that allows them to predict nonlinear response variables. In other words, GLMs are an extension of linear regression such that nonlinear relationships can be predicted. For example, the log-odds (or "logit") link function allows the GLM framework to "extend" linear regression, which can only explicitly predict continuous outcomes, to logistic regression, which can be used to predict the probability of an observation belonging to one of two classes.

Feature selection - A term referring to a general analytical approach in which a subset of a dataset's input variables are identified (manually or automatically) as the most important for predicting or explaining an outcome of interest.

Significance Analysis of Microarrays (SAM) - An analytical approach in which differences in marker expression distributions are detected by permuting sample labels randomly in order to define a null distribution of "expected" differences resulting from chance variation. Originally published in 2001 to analyze microarrays, this approach has since been applied to differential expression analyses on a variety of transcript- and protein-level data.

Miscellaneous

Deep learning - A subfield within ML in which artificial neural networks are used to solve both supervised and unsupervised learning problems. An area of rapid growth in bioinformatics.

Distance metrics - We discuss several distance metrics in the text. For the n-dimensional vectors x and y, these distances are given by...

Manhattan (L1-norm): $\sum_i^n |x_i - y_i|$. Used in Lasso regularization and elastic net.

Euclidean (L2-norm): $\sum_i^n \sqrt{x_i^2 - y_i^2}$. Most commonly-used distrance metric across single-cell data types. Used by many algorithms, including PhenoGraph.

Pearson: $1 - r_{xy}$, where r_{xy} is the Pearson correlation coefficient between x and y

Cosine: $1 - \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}}$. Represents the angle between the vectors x and y, and is invariant to their size. Used by Scaffold.

Mahalanobis: $\sqrt{(x-y)S^{-1}(x-y)}$, where S is the covariance matrix for the distribution that contains x and y. Often thought of as a "multidimensional Z-score."

Hyperparameters (tuning parameters) - Numeric values that influence how a model is computed, but are not estimated directly from input data. Generally speaking, users specify hyperparameter values through trial and error or through an exhaustive "grid search" of many values.