Supporting Information

A User's Guide to Machine Learning for Polymeric Biomaterials

*Travis A. Meyer, Cesar Ramirez, Matthew J. Tamasi, Adam J. Gormley**

Department of Biomedical Engineering, Rutgers, The State University of New Jersey,
Piscataway, NJ 08854, United States

*Correspondence to:

Adam J. Gormley: adam.gormley@rutgers.edu

Vocabulary and Definitions

Activation Function: A non-linear function applied to the dot-product of prior layers at each node in an artificial neural network, which facilitates modeling of non-linear data and facilitates backpropagation process. Several different activation functions exist, with the most common being tanh, ReLu, and softmax.

Artificial Neural Network (ANN): A versatile machine learning algorithm, used for both supervised and unsupervised learning, which are built from a number of nodes grouped into layers. Information from previous layers is fed into subsequent layers, processed, and then passed along.

Bias: Bias refers to error that results from assumptions a model makes when making predictions about data. An example is the use of a linear regression model (with the assumption that data is linear) on a dataset with clear non-linear relationships. A high bias model will be relatively simple and unable to capture complicated patterns within the data. A model with low bias will be more complex and able to capture more complex patterns. Highly biased models tend to be underfit.

Boosted Tree Models: An ensembled model of decision trees in which numerous simple trees are trained in an iterative manner, with the results of one tree being used to inform the training of subsequent trees. Two common forms of boosted tree models are adaptive boosting and gradient boosting.

Bootstrap Aggregation (Bagging): A type of ensembling technique in which multiple models are trained on different subsets of the training data. In each training iteration, a random assortment of observations from the training set are chosen with replacement. Bagging is commonly used as one step in the generation of random forest models.

Class Imbalance: A phenomenon where the dataset used for training has an uneven distribution of labels, such that an observation with a label of a minority class will only be used for training infrequently. For example, if you wanted to perform a classification task identifying benign versus malignant tumors based on certain features, but the training data set only included ten malignant tumor labels out of 1,000 observations.

Classification: A type of machine learning task in which new, unlabeled data needs to be sorted into one of two or more classes. When only two classes exist, the task is known as binary classification. Any model which works with classification tasks is known as a classifier.

Convolution Neural Network (CNN): A type of artificial neural network in which single-layer convolution layers apply a filter to local groups of input features prior to their processing through a traditional ANN. These kinds of neural networks work best when data possesses a local structure, such as within images.

Cross Validation: A validation technique in which the full training data set is split into X-distinct subsets, known as folds. Training is then run X times, with one fold held out for validation and the remaining used for model training. For example, 5-fold cross validation consists of splitting the data into five subsets with 20% of the data in each, followed by five independent training/validation steps. The final performance of the model can thus be assessed by aggregating the accuracy score from each sub-run.

Data Leakage: A phenomenon in which data used for testing a machine learning model has direct connections to data used to train the model. Data leakage leads to validation scores which overestimate the predictive performance of the model on truly new data.

Decision Tree: Supervised machine learning algorithm which utilizes a series of branching if/else decisions to appropriately segment data based on shared features. Decision trees can be used for both classification and regression tasks, and are frequently developed with ensemble methods, such as random forests or boosted trees to improve their performance.

Deep Learning: A term used to refer to the use of artificial neural networks which utilize multiple hidden layers between the input layer and the output layer for processing and analyzing data.

Dimensionality Reduction: Unsupervised machine learning task which works to reduce the number of features/dimensions that are considered while still maintaining relationships between the data points as much as possible. A common example of dimensionality reduction is principal component analysis (PCA).

Ensemble Methods: The process of building multiple different models and using their combined outputs to develop a single, robust model. Ensemble methods can refer to utilizing the same type of machine-learning algorithm (i.e., decision tree, neural network, support vector machine, etc.) but training on different data sets, using the same algorithm on the same data but with different hyperparameters, or using completely different algorithms trained on the same data such that the strengths of each algorithm complement each other.

Epochs: The number of times that a neural network has undergone the training process on the entire training data set. Two epochs refer to two separate passes through the data, four epochs refers to four passes, etc.

Features: The predictive variables or data that are fed into a machine learning algorithm as inputs, which are then used to assign a prediction or determine patterns. Features are analogous to independent variables in the traditional empirical research process.

Gaussian Process Model (GPM): A supervised machine learning algorithm in which the optimal predictive function that is trained is modeled as a distribution of various functions. This process allows for the calculation uncertainty/variance terms within the predicted labels.

Gini-index: A metric for leaf purity for classification tasks which measures how often a randomly chosen element within a node would be labeled incorrectly if the label was chosen randomly based on the distribution of elements within the node. A leaf with a completely homogenous set of labels would thus have a gini-index of 0, since all data would be classified with the same label. Also known as Gini impurity.

Gradient Descent: Optimization algorithm used to find the minimum of a set value within a parameter space. The slope of the function to be minimized is calculated, and the algorithm uses this slope in order to choose which parameters to vary in order to "step" to a lower region of the value. This process is repeated iteratively until a minimum has been reached; there exist no new steps "down" the slope to reach a lower value.

Graph Convolution Neural Network: A form of convolution neural network used when data is connected by specific relationships or interactions, such as with the atoms of a molecule.

Grid Search Method: A process of hyperparameter optimization in which a range of various hyperparameters to be tuned is established and model performance is evaluated with every combination of hyperparameters established.

Hyperparameters: Variables which determine how a model is set up and how the training process proceeds. Hyperparameters are unmodified by the algorithm during the training process, and users can alter the hyperparameters as a means of modulating model performance during a process known as hyperparameter tuning.

Information Gain: Concept in decision trees which helps to drive the optimization of the feature choices at each branching point. Simply stated, information gain refers to the amount of information about labels that is learned when splitting the data based on a certain feature – splits which result in child nodes with a higher label purity give the greatest information gain and are thus favored by decision tree algorithms.

K-Nearest Neighbors (KNN): A supervised machine learning algorithm which assigns labels to new data based on the distance in the feature space from known, labeled data. The number of closest neighbors which contribute to the labeling of new data is controlled by the hyperparameter k. While typically used for classification tasks, KNN models can be used for regression as well.

Kernel Functions: Mathematical functions used to transform data into new feature-spaces, typically through the introduction of new dimensions. Kernel functions are primarily used to convert problems without linear solutions into those which are linear (i.e., a linear decision boundary between two labels in a binary classification problem). This technique is known as the kernel trick, which is where kernel functions get their name.

Labels: The class or value that a machine-learning model seeks to predict. Labels are analogous to dependent variables in the traditional empirical research process.

Lasso Regression: A type of linear regression involving regularization, which acts to reduce or remove the effect of features with lower predictive power in the model.

Latent Space: An abstract, multi-dimensional space which encodes a meaningful representation of external data. Similar data is located close together within the latent space. In general, latent spaces have a lower dimensionality than the feature space of the data to be represented.

Leaf: The final node on a decision tree in which no more branching occurs and a predicted label is established. Leaves are defined by their purity, which refers to the homogeneity of labels from data sorted into that leaf.

Linear Regression: A linear algorithm used for regression tasks in which a line of best-fit is created that models the process underlying the data.

Logistic Regression: A linear algorithm used for binary classification tasks, in which the coefficients of a linear combination of features is optimized to the predict the probability of a certain label occurring.

Loss Function: Mathematical function that is used to calculate the error between predicted and actual labels during the training and validation process. The minimization of the loss function is the primary goal of optimization during training and hyperparameter tuning. Common loss functions include mean squared error for regression tasks and binary cross entropy for classification problems. Loss functions are also known as cost functions.

Machine Learning (ML): A subset of artificial intelligence that focuses on the use of algorithms which enable computer programs to iterate and evolve to accomplish a certain task, without the need for explicit programming.

Random Forest: An ensembled model of decision trees in which multiple individual decision trees, all trained using different subsets of the training data and restricted to the use of different features, are averaged together to produce the final label. Random Forest models help minimize the overfitting common to decision trees.

Recurrent Neural Network (RNN): A type of neural network in which the predicted output from one observation is used as an additional input for the processing of a subsequent observation.

RNNs are used when data is present in a defined sequence, such as with time- or order-dependent data.

Representation Learning: A type of machine learning process in which internal components of the model are built to represent patterns observed within the data. Artificial neural networks are commonly used algorithms that allow for representation learning.

One-Hot Encoding: A technique for converting qualitative classification information into a machine-interpretable output that does not carry inherent sequential or ordinal bias. One-hot encoding produces a vector with a length equal to the number of possible classes for a given variable – the vector is filled with zeros for classes which to not describe the specific observation, while a one is given for correct class labels. For example, one variable could be one of three possible classes (PEG, PVA, PLGA) – rather than assigning 1 to PEG, 2 to PVA, and 3 to PLGA, observations with the PVA class are instead assigned the vector [0,1,0].

Overfitting: Term used to refer to a phenomenon where the models learn to predict both signal and noise within a training dataset, leading to poor generalizable when the model is introduced to unseen data. Overfit models tend to have low error/good validation scores when assessed during training, but high error/low validation scores when unseen data is used.

Parameters: Parameters are values used by the machine learning program in order to perform its task. Examples of parameters include weights, coefficients, etc. These values are pre-set by the algorithm and then iteratively modified throughout the training procedure in order to minimize the loss function. This is in comparison to **Hyperparameters** which are user defined values that are unchanged during training.

Receiver Operating Characteristic Curve (ROC): A graphical plot used to validate the performance of binary classifiers which plots the true positive rate of identification verse the false positive rate of identification.

Regression: A type of machine learning task in which the labels to be predicted are continuous variables which exist on a spectrum, rather than distinct classes. Any model which works with regression tasks is known as a regressor.

Regularization: A mathematical technique used to avoid overfitting which introduces an additional penalty to the loss function that minimizes the introduction of coefficients or parameters with extreme values.

Ridge Regression: A type of linear regression involving regularization, used to minimize the impact of large coefficients during linear regression and useful when features are highly correlated.

Scaling: A data processing technique in which continuous variables are modified so the spanned range is consistent across different kinds of features. A common form of scaling is to modify the data such that the mean of the data is zero and the standard deviation is one, which can be accomplished by subtracting the mean from each observation and dividing by the standard deviation.

Supervised Learning: Type of machine learning in which training data includes labels that the task wishes to predict on new, unlabeled data. Supervised learning can be further broken into classification or regression tasks based on the kind of labels that exist.

Support Vector Machines (SVM): A class of supervised machine learning algorithms which derive an equation to either best split data of different classes (for classification) or a line of best fit (for regression), based on input from data point closest to the boundary (known as support vectors). SVM’s are capable of modeling non-linear problems through the use of kernel functions, which translate the data into a new feature space which can be solved linearly.

Test-Train Split: The process of taking the initial data set and subdividing into groups such that machine learning models are trained on one portion of the data, and then testing to assess model function is done on new, unseen data. This is a critical process to ensure validation processes are judging the model’s ability to generalize to new, unseen data.

Training: The process by which a machine learning model learns to properly accomplish a task. The training process involves generating predictions from or groupings of training data, estimating error using a loss function, and then modification of model parameters in an iterative effort to minimize the loss function.

Underfitting: Term used to refer to a phenomenon in which the patterns or variance within the data are not adequately captured by the model, leading to large errors in model performance and poor validation scores.

Unsupervised Learning: Type of machine learning in which training data is unlabeled and the task wishes to understand patterns or clusters within the data. Examples of unsupervised learning include dimensionality reduction, clustering, and association tasks.

Validation: The process of assessing the performance of the model by comparing outputs to known labels or assessing purity of groupings and calculating the error.

Variance: A measure of how the outputs of a model with vary depending on changes to the training data used to fit the model. Low variance models will tend to perform similarly when trained on different data sets, while high variance models will have different outputs when different training data is used. A high variance model tends to lead to over-fitting.