Supervised Machine Learning Models

datacamp

Category	ALGORITHM	DESCRIPTION & APPLICATION	ADVANTAGES	DISADVANTAGES
Regression Only Models	**Linear Regression**	Linear Regression models a linear relationship between input variables and a continuous numerical output variable. The default loss function is the mean square error (MSE).	1. Fast training because there are few parameters. 2. Interpretable/Explainable results by its output coefficients.	1. Assumes a linear relationship between input and output variables. 2. Sensitive to outliers. 3. Typically generalizes worse than ridge or lasso regression.
Regression Only Models	**Polynomial Regression**	Polynomial Regression models nonlinear relationships between the dependent, and independent variable as the n-th degree polynomial.	1. Provides a good approximation of the relationship between the dependent and independent variables. 2. Capable of fitting a wide range of curvature.	1. Poor interpretability of the coefficients since the underlying variables can be highly correlated. 2. The model fit is nonlinear but the regression function is linear. 3. Prone to overfitting.
Regression Only Models	**Support Vector Regression**	Support Vector Regression (SVR) uses the same principle as SVMs but optimizes the cost function to fit the most straight line (or plane) through the data points. With the kernel trick it can efficiently perform a non-linear regression by implicitly mapping their inputs into high-dimensional feature spaces.	1. Robust against outliers. 2. Effective learning and strong generalization performance. 3. Different Kernel functions can be specified for the decision function.	1. Does not perform well with large datasets. 2. Tends to underfit in cases where the number of variables is much smaller than the number of observations.
Regression Only Models	**Gaussian Process Regression**	Gaussian Process Regression (GPR) uses a Bayesian approach that infers a probability distribution over the possible functions that fit the data. The Gaussian process is a prior that is specified as a multivariate Gaussian distribution.	1. Provides uncertainty measures on the predictions. 2. It is a flexible and usable non-linear model which fits many datasets well. 3. Performs well on small datasets as the GP kernel allows to specify a prior on the function space.	1. Poor choice of kernel can make convergence slow. 2. Specifying specific kernels requires deep mathematical understanding.
Regression Only Models	**Robust Regression**	Robust Regression is an alternative to least squares regression when data is contaminated with outliers. The term “robust” refers to the statistical capability to provide useful information even in the face of outliers.	1. Designed to overcome some limitations of traditional parametric and non-parametric methods. 2. Provides better regression coefficient over classical regression methods when outliers are present.	1. More computationally intensive compared to classical regression methods. 2. It is not a cure-all for all violations, such as imbalanced data, poor quality data. 3. If no outliers are present in the data, it may not provide better results than classical regression methods.
Both Regression and Classification Models	**Decision Trees**	Decision Tree models learn on the data by making decision rules on the variables to separate the classes in a flowchart like a tree data structure. They can be used for both regression and classification.	1. Explainable and interpretable. 2. Can handle missing values.	1. Prone to overfitting. 2. Can be unstable with minor data drift. 3. Sensitive to outliers.
Both Regression and Classification Models	**Random Forest**	Random Forest classification models learn using an ensemble of decision trees. The output of the random forest is based on a majority vote of the different decision trees.	1. Effective learning and better generalization performance. 2. Can handle moderately large datasets. 3. Less prone to overfit than decision trees.	1. Large number of trees can slow down performance. 2. Predictions are sensitive to outliers. 3. Hyperparameter tuning can be complex.
Both Regression and Classification Models	**Gradient Boosting**	An ensemble learning method where weak predictive learners are combined to improve accuracy. Popular techniques include XGBoost, LightGBM and more.	1. Handling of multicollinearity. 2. Handling of non-linear relationships. 3. Effective learning and strong generalization performance. 4. XGBoost is fast and is often used as a benchmark algorithm.	1. Sensitive to outliers and can therefore cause overfitting. 2. High complexity due to hyperparameter tuning. 3. Computationally expensive.
Both Regression and Classification Models	**Ridge Regression**	Ridge Regression penalizes variables with low predictive outcomes by shrinking their coefficients towards zero. It can be used for classification and regression.	1. Less prone to overfitting. 2. Best suited when data suffers from multicollinearity. 3. Explainable & Interpretable.	1. All the predictors are kept in the final model. 2. Doesn't perform feature selection.
Both Regression and Classification Models	**Lasso Regression**	Lasso Regression penalizes features that have low predictive outcomes by shrinking their coefficients to zero. It can be used for classification and regression.	1. Good generalization performance. 2. Good at handling datasets where the number of variables is much larger than the number of observations. 3. No need for feature selection.	1. Poor interpretability/explainability as it can keep a single variable. from a set of highly correlated variables.
Both Regression and Classification Models	**AdaBoost**	Adaptive Boosting uses an ensemble of weak learners that is combined into a weighted sum that represents the final output of the boosted classifier.	1. Explainable & Interpretable. 2. Less need for tweaking parameters. 3. Usually outperforms Random Forest.	1. Less prone to overfitting as the input variables are not jointly optimized. 2. Sensitive to noisy data and outliers.
Classification Only Models	**SVM**	In its simplest form, support vector machine is a linear classifier. But with the kernel trick, it can efficiently perform a non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces. This makes SVM one of the best prediction methods.	1. Effective in cases with a high number of variables. 2. Number of variables can be larger than the number of samples. 3. Different Kernel functions can be specified for the decision function.	1. Sensitive to overfitting, regularization is crucial. 2. Choosing a “good” kernel function can be difficult. 3. Computationally expensive for big data due to high training complexity. 4. Performs poorly if the data is noisy (target classes overlap).
Classification Only Models	**Nearest Neighbors**	Nearest Neighbors predicts the label based on a predefined number of samples closest in distance to the new point.	1. Successful in situations where the decision boundary is irregular. 2. Non-parametric approach as it does not make any assumption on the underlying data.	1. Sensitive to noisy and missing data. 2. Computationally expensive because the entire set of n points for every execution is required.
Classification Only Models	**Logistic Regression (and its extensions)**	The logistic regression models a linear relationship between input variables and the response variable. It models the output as binary values (0 or 1) rather than numeric values.	1. Explainable & Interpretable. 2. Less prone to overfitting using regularization. 3. Applicable for multi-class predictions.	1. Makes a strong assumption about the relationship between input and response variables. 2. Multicollinearity can cause the model to easily overfit without regularization.
Classification Only Models	**Linear Discriminant Analysis**	The linear decision boundary maximizes the separability between the classes by finding a linear combination of features.	1. Explainable & Interpretable. 2. Applicable for multi-class predictions.	1. Multicollinearity can cause the model to overfit. 2. Assuming that all classes share the same covariance matrix. 3. Sensitive to outliers. 4. Doesn't work well with small class sizes.