Machine Learning Cheat Sheet

In this cheat sheet, you'll have a guide around the top machine learning algorithms, their advantages and disadvantages, and use-cases.

Apr 2022 · 8 min read

Richie Cotton
Webinar & podcast host, course and book author, spends all day chit-chatting about data

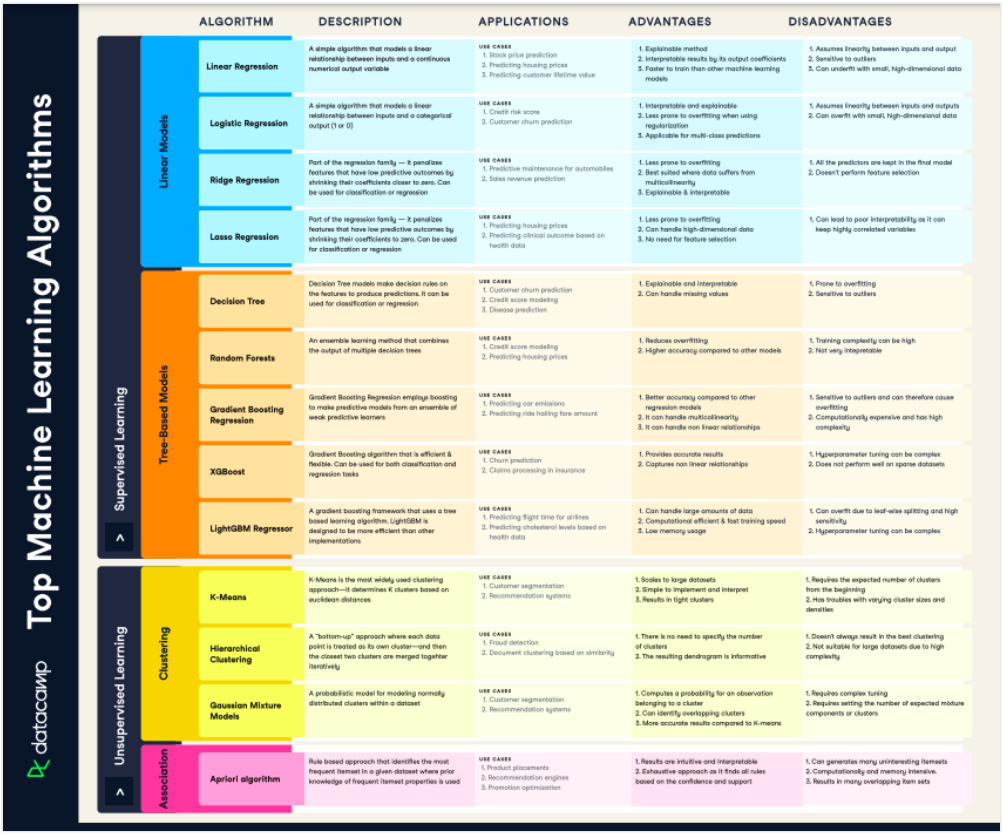
TOPICS

Machine Learning

Data Science

When working with machine learning, it's easy to try them all out without understanding what each model does, and when to use them. In this cheat sheet, you'll find a handy guide describing the most widely used machine learning models, their advantages, disadvantages,

and some key use-cases.



Have this cheat sheet at your fingertips

Download PDF

Supervised Learning

Supervised learning models are models that map inputs to outputs, and attempt to extrapolate patterns learned in past data on unseen data. Supervised learning models can be either regression models, where we try to predict a continuous variable, like stock prices—or classification models, where we try to predict a binary or multi-class variable, like whether a customer will churn or not. In the section below, we'll explain two popular types of supervised learning models: linear models, and tree-based models.

Linear Models

In a nutshell, linear models create a best-fit line to predict unseen data. Linear models imply that outputs are a linear combination of features. In this section, we'll specify commonly used linear models in machine learning, their advantages, and disadvantages.

Algorithm	Description	Applications	Advantages	Disadvantages
Linear Regression	A simple algorithm that models a linear relationship between inputs and a continuous numerical output variable	1. Stock Price Prediction 2. Predicting housing prices 3. Predicting customer lifetime value	1. Explainable method 2. Interpretable results by its output coefficient 3. Faster to train than other machine learning models	1. Assumes linearity between inputs and output 2. Sensitive to outliers 3. Can underfit with small, high-dimensional data
Logistic Regression	A simple algorithm that models a linear relationship between inputs and a categorical output (1 or 0)	1. Predicting credit risk score 2. Customer churn prediction	1. Interpretable and explainable 2. Less prone to overfitting when using regularization 3. Applicable for multi-class predictions	1. Assumes linearity between inputs and outputs 2. Can overfit with small, high-dimensional data

Ridge Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients closer to zero. Can be used for classification or regression	1. Predictive maintenance for automobiles 2. Sales revenue prediction	1. Less prone to overfitting 2. Best suited where data suffer from multicollinearity 3. Explainable & interpretable	1. All the predictors are kept in the final model 2. Doesn't perform feature selection
Lasso Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients to zero. Can be used for classification or regression	1. Predicting housing prices 2. Predicting clinical outcomes based on health data	1. Less prone to overfitting 2. Can handle high-dimensional data 3. No need for feature selection	1. Can lead to poor interpretability as it can keep highly correlated variables

Tree-based models

In a nutshell, tree-based models use a series of "if-then" rules to predict from decision trees. In this section, we'll specify commonly used linear models in machine learning, their advantages, and disadvantages.

Algorithm	Description	Applications	Advantages	Disadvantages

Decision Tree	Decision Tree models make decision rules on the features to produce predictions. It can be used for classification or regression	1. Customer churn prediction 2. Credit score modeling 3. Disease prediction	1. Explainable and interpretable 2. Can handle missing values	1. Prone to overfitting 2. Sensitive to outliers
Random Forests	An ensemble learning method that combines the output of multiple decision trees	1. Credit score modeling 2. Predicting housing prices	1. Reduces overfitting 2. Higher accuracy compared to other models	1. Training complexity ca be high 2. Not very interpretable
Gradient Boosting Regression	Gradient Boosting Regression employs boosting to make predictive models from an ensemble of weak predictive learners	1. Predicting car emissions 2. Predicting ride-hailing fare amount	1. Better accuracy compared to other regression models 2. It can handle multicollinearity It can handle non-linear relationships	1. Sensitive to outliers and co therefore caus overfitting 2. Computationa expensive and has high complexity

XGBoost	Gradient Boosting algorithm that is efficient & flexible. Can be used for both classification and regression tasks	1. Churn prediction 2. Claims processing in insurance	1. Provides accurate results 2. Captures non-linear relationships	1. Hyperparamet tuning can be complex 2. Does not perform well o sparse dataset
LightGBM Regressor	A gradient boosting framework that is designed to be more efficient than other implementations	1. Predicting flight time for airlines 2. Predicting cholesterol levels based on health data	1. Can handle large amounts of data 2. Computational efficient & fast training speed 3. Low memory usage	1. Can overfit du to leaf-wise splitting and high sensitivity 2. Hyperparamet tuning can be complex

Unsupervised Learning

Unsupervised learning is about discovering general patterns in data. The most popular example is clustering or segmenting customers and users. This type of segmentation is generalizable and can be applied broadly, such as to documents, companies, and genes. Unsupervised learning consists of clustering models, that learn how to group similar data points together, or association algorithms, that group different data points based on pre-defined rules.

Clustering models

Algorithm	Description	Applications	Advantages	Disadvantages

K-Means	K-Means is the most widely used clustering approach—it determines K clusters based on euclidean distances	1. Customer segmentation 2. Recommendation systems	1. Scales to large datasets 2. Simple to implement and interpret 3. Results in tight clusters	1. Requires the expected number of clusters from the beginning 2. Has troubles with varying cluster sizes and densities
Hierarchical Clustering	A "bottom-up" approach where each data point is treated as its own cluster—and then the closest two clusters are merged together iteratively	1. Fraud detection 2. Document clustering based on similarity	1. There is no need to specify the numberof clusters 2. The resulting dendrogram is informative	1. Doesn’t always result in the best clustering 2. Not suitable for large datasets due to high complexity

Gaussian Mixture Models	A probabilistic model for modeling normally distributed clusters within a dataset	1. Customer segmentation 2. Recommendation systems	1. Computes a probability for an observation belonging to a cluster 2. Can identify overlapping clusters 3. More accurate results compared to K-means	1. Requires complex tuning 2. Requires setting the number of expected mixture components or clusters

Association

Algorithm	Description	Applications	Advantages	Disadvantages
Apriori Algorithm	Rule based approach that identifies the most frequent itemset in a given dataset where prior knowledge of frequent itemset properties is used	1. Product placements 2. Recommendation engines 3. Promotion optimization	1. Results are intuitive and Interpretable 2. Exhaustive approach as it finds all rules based on the confidence and support	1. Generates many uninteresting itemsets 2. Computationally and memory intensive. Results in many overlapping item sets

TOPICS

Machine Learning Data Science

Related

What is Machine Learning Inference? An Introduction to...

Zoumana Keita

Introduction to Unsupervised Learning

Kurtis Pykes

How the Aviation Industry Leverages Data Science

Adel Nehme

See More →

Grow your data skills with DataCamp for Mobile

Make progress on the go with our mobile courses and daily 5-minute coding challenges.

Download on the App Store GET IT ON Google Play

LEARN

Learn Python

Learn R

Learn SQL

Learn Power BI

Learn Tableau

Assessments

Career Tracks

Skill Tracks

Courses

Data Science Roadmap

DATA COURSES

Python Courses

R Courses

SQL Courses

Power BI Courses

Tableau Courses

Spreadsheet Courses

Data Analysis Courses

Data Visualization Courses

Machine Learning Courses

Data Engineering Courses

WORKSPACE

Get Started

Templates

Integrations

Documentation

CERTIFICATION

Certifications

Data Scientist

Data Analyst

Hire Data Professionals

RESOURCES

Resource Center

Upcoming Events

Blog

Tutorials

Open Source

RDocumentation

Course Editor

Book a Demo with DataCamp for Business

PLANS

Pricing

For Business

For Universities

Discounts, Promos & Sales

DataCamp Donates

SUPPORT

Help Center

Become an Instructor

Become an Affiliate

ABOUT

About Us

Learner Stories

Careers

Press

Leadership

Contact Us