2022 • GIORGIA BERTACCHINI

MLOps

Machine Learning Model Operationalization Management

Part 3

Practical

MLOps • 2022

Supervised Learning

TERMS

In supervised learning, the data analyst works with a collection of **labeled examples** {(x1, y1),(x2, y2), . . . ,(xN, yN)}.
Each element xi among N is called a **feature vector**.
Each such value of a feature vector is called a **feature**.

In supervised learning,

- the problem of predicting a class is called **classification**,
- the problem of predicting a real number is called **regression**.

The value that has to be predicted by a supervised model is called a **target**.

The goal of a supervised learning algorithmis to use a dataset to produce a model that takes a feature vector x as input and outputs information that allows deducing a label forthis feature vector.

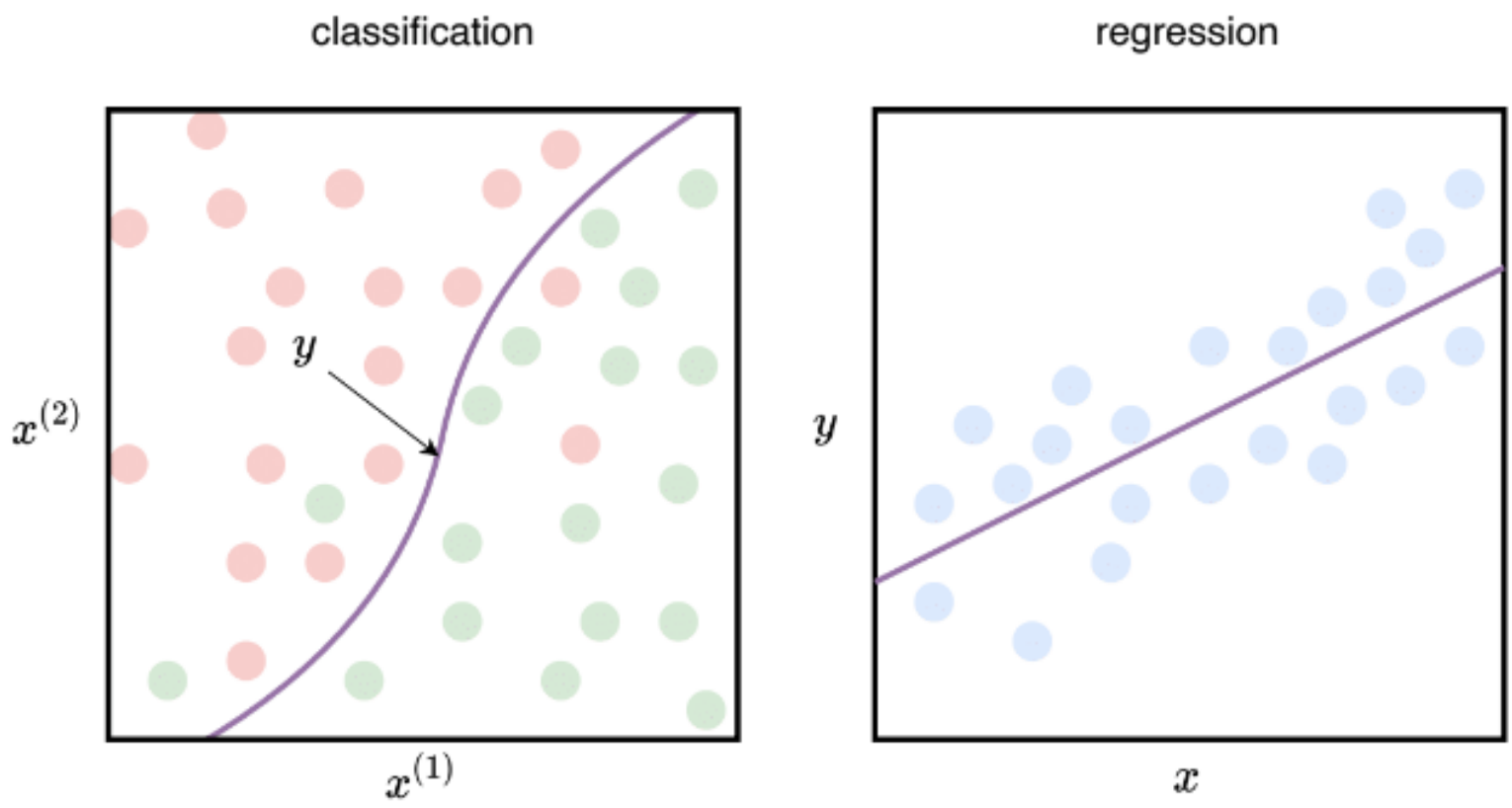
Supervised Learning

DIFFERENCES

In **classification**, the learning algorithm looks for a line (or, more generally, a hypersurface) that separates examples of different classes from one another.

- In a classification problem, a label is a member of a finite set of classes.
- If the size of the set of classes is two, we talk about **binary classification** (also called **binomial**). **Multiclass classification** (also called **multinomial**) is a classification problem with more classes.

In **regression**, on the other hand, the learning algorithm looks to find a line or a hypersurface that closely follows the training examples.



Unsupervised Learning

TERMS

In unsupervised learning, the dataset is a collection of **unlabeled examples** {x1,x2, . . . ,xN}. Again,x is a feature vector, and the goal of an unsupervised learning algorithm is to create a model that takes a feature vector x as input and either transformsit into another vector or into a value that can be used to solve a practical problem.

EXAMPLES

- Clustering
  - the model returns the ID of the cluster for each feature vector inthe dataset. Clustering is useful for finding groups of similar objects in a large collection of objects, such as images or text documents.
- Dimensionality reduction
  - the model's output is a feature vector with fewer dimensions than the input. For example, the scientist has a feature vector that is too complex to visualize (it has more than three dimensions). This new feature vector can be plotted on a graph.
- Outlier detection
  - the output is a real number that indicates how the input feature vectoris different from a "typical" example in the dataset. Outlier detection is useful for solving anetwork intrusion problem...

Data Terminology

TERMS

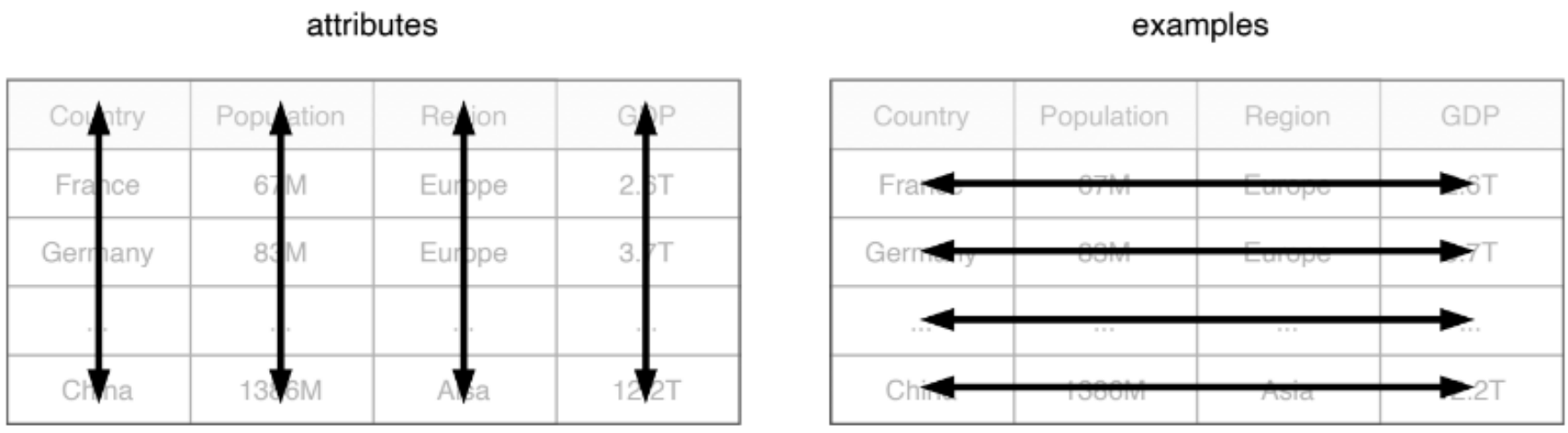
Data can be used **directly or indirectly**. Directly-used data is a basis for forming a dataset of examples. Indirectly-used data is used to enrich those examples.

Raw data is a collection of entities in their natural form; they cannot always be directly employable for machine learning.

To be employable in machine learning, a necessary (but not sufficient) condition for the datais to be tidy. **Tidy data** can be seen as a spreadsheet, in which each row represents one **example**, and columns represent various **attributes** of an example.

In practice, to obtain tidy data from raw data, data analysts often resort to the procedure called **feature engineering,** which is applied to the data with the goal to transform each raw example into a feature vector x.

Data can be tidy, but still not usable by a particular machine learning algorithm. Most machine learning algorithms, in fact, only acceptttraining data in the form of a collection of numerical feature vectors.



Data Terminology

SETS

In practice, data analysts work with three distinct sets of examples:

1. training set
2. validation set
3. test set.

The training set is usually the biggest one; the learning algorithm uses the training set to produce the model. The validation and test sets are roughly the same size, much smaller than the size of the training set.

The learning algorithm is not allowed to use examples from the validation or test sets to train the model. That is why those two sets are also called **holdout sets**.

We need two holdout sets and not one because we use the v**alidation set** to

- choose the learning algorithm, and
- find the best configuration values for that learning algorithm (known as **hyperparameters**).

We use the **test set** to assess the model before delivering it to the client or putting it in production. The test set is used for reporting: once you have your best model, you test its performanceon the test set and report the results.

Data Terminology

PIPELINE

A machine learning pipeline is a sequence of operations on the dataset that goes from its initial state to the model.

PARAMETERS VS. HYPERPARAMETERS

Hyperparameters are inputs of machine learning algorithms or pipelines that influencethe performance of the model. They don’t belong to the training data and cannot be learned from it.

Parameters, on the other hand, are variables that define the model trained by the learning algorithm. Parameters are directly modified by the learning algorithm based on the training data. The goal of learning is to find such values of parameters that make the model optimalin a certain sense.

TRAINING VS. SCORING

When we apply a machine learning algorithm to a dataset in order to obtain a model, we talk about **model training** or simply **training**.

When we apply a trained model to an input example (or, sometimes, a sequence of examples) in order to obtain a prediction (or, predictions) or to somehow transform an input, we talk about **scoring**.

Learning

SHALLOW AND DEEP LEARNING ALGORITHM

A shallow learning algorithm trains a model that makes predictions directly from the input features.
A deep learning algorithm trains a layered model, in which each layer generates outputs by taking the outputs of the preceding layer as inputs.

MACHINE LEARNING PROJECT LIFE CYCLE

A machine learning project life cycle consists of the following stages:

1. goal definition,
2. data collection and preparation,
3. feature engineering,
4. model training,
5. model evaluation,
6. model deployment,
7. model serving,
8. model monitoring,
9. model maintenance.

Data Partitioning

To obtain good partitions of entire dataset into these three disjoint sets:

- Data was randomized before the split
- Split was applied to raw data.

The validation and test data are only used to calculate statistics reflecting the performance of the model.

A small dataset of less than a thousand examples would do best with 90% of the data used for training. In this case, you might decide to not have a distinct validation set, and instead simulate with the cross-validation technique

Three Sets

- The first, the **training set**, is used to train the model. It is the data the machine learning algorithm "sees."
- The second, the **validation set** is not seen by the machine learning algorithm. The data analyst uses it to estimate the performance of different machine learning algorithms (or the same algorithm configured with different values of hyperparameters) or models when applied to new data.
- The remaining **test set**, which is also not seen by the learning algorithm, is used at the end of the project to evaluate and report the performance of the model the best performing on the validation data.

Missing Attributes

Typical approaches of dealing with missing values for an attribute:

- removing the examples with missing attributes from the dataset (if your dataset is big enough to safely sacrifice some data);
- using a learning algorithm that can deal with missing attribute values (such as the decision tree learning algorithm);
- using adata imputation technique.

DATA IMPUTATION TECHNIQUES

Simpler examples:

- To impute the value of a missing numerical attribute, one technique consists in replacing the missing value by the **average value** of this attribute in the rest of the dataset.
- Another technique is to replace the missing value with a value **outside the normal range** of values. For example, if the regular range is [0,1], you can set the missing value to 2 or −1; if the attribute is categorical, such as days of the week, then a missing value can be replaced by the value “Unknown.”
- If the attribute is numerical, another technique is replacing the missing value with a **value in the middle of the range**. Here, the idea is that the value in the middle of the range will not significantly affect the prediction.

DATA AUGMENTATION TECHNIQUES

They are often used to get more labeled examples without additional manual labeling. The techniques usually apply to image data, but could also be applied to text and other types of perceptive data. It consists of applying simple operations, such as crop or flip, to the original images to obtain new images.

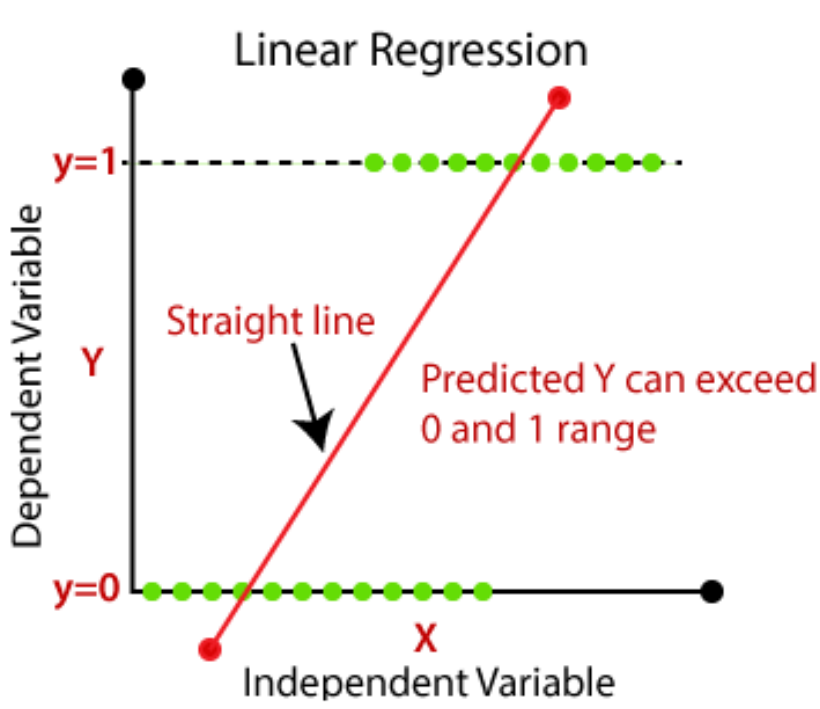
Model Training

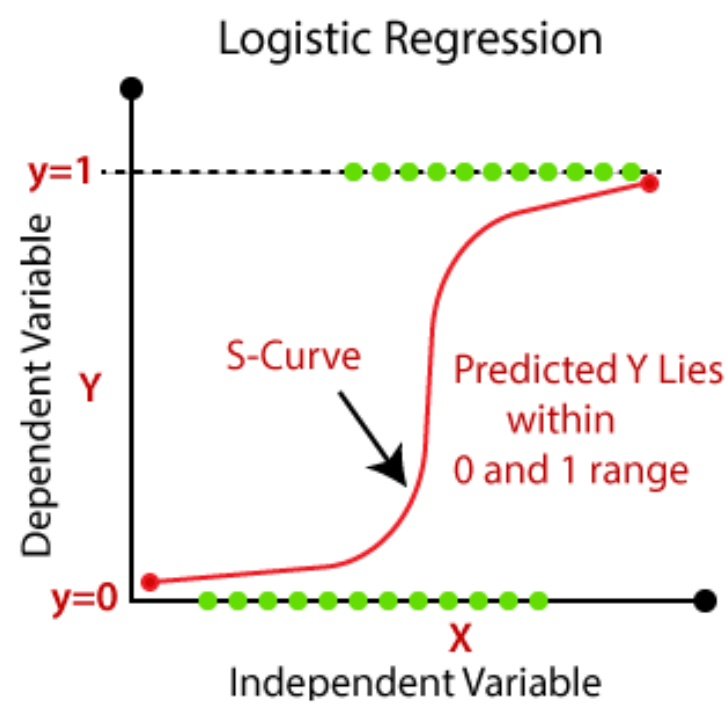
LINEAR REGRESSION

- The outputs produced must be a continuous value, such as price and age.
- Used for solving Regression problems.
- We are finding and using the line of best fit to help us easily predict outputs.
- Least square estimation method is used for the estimation of accuracy.

LOGISTIC REGRESSION

- The outputs produced must be Categorical values such as 0 or 1, Yes or No.
- Used for solving Classification problems.
- We are using the S-curve (Sigmoid) to help us classify predicted outputs.
- Maximum likelihood estimation method is used for the estimation of accuracy.

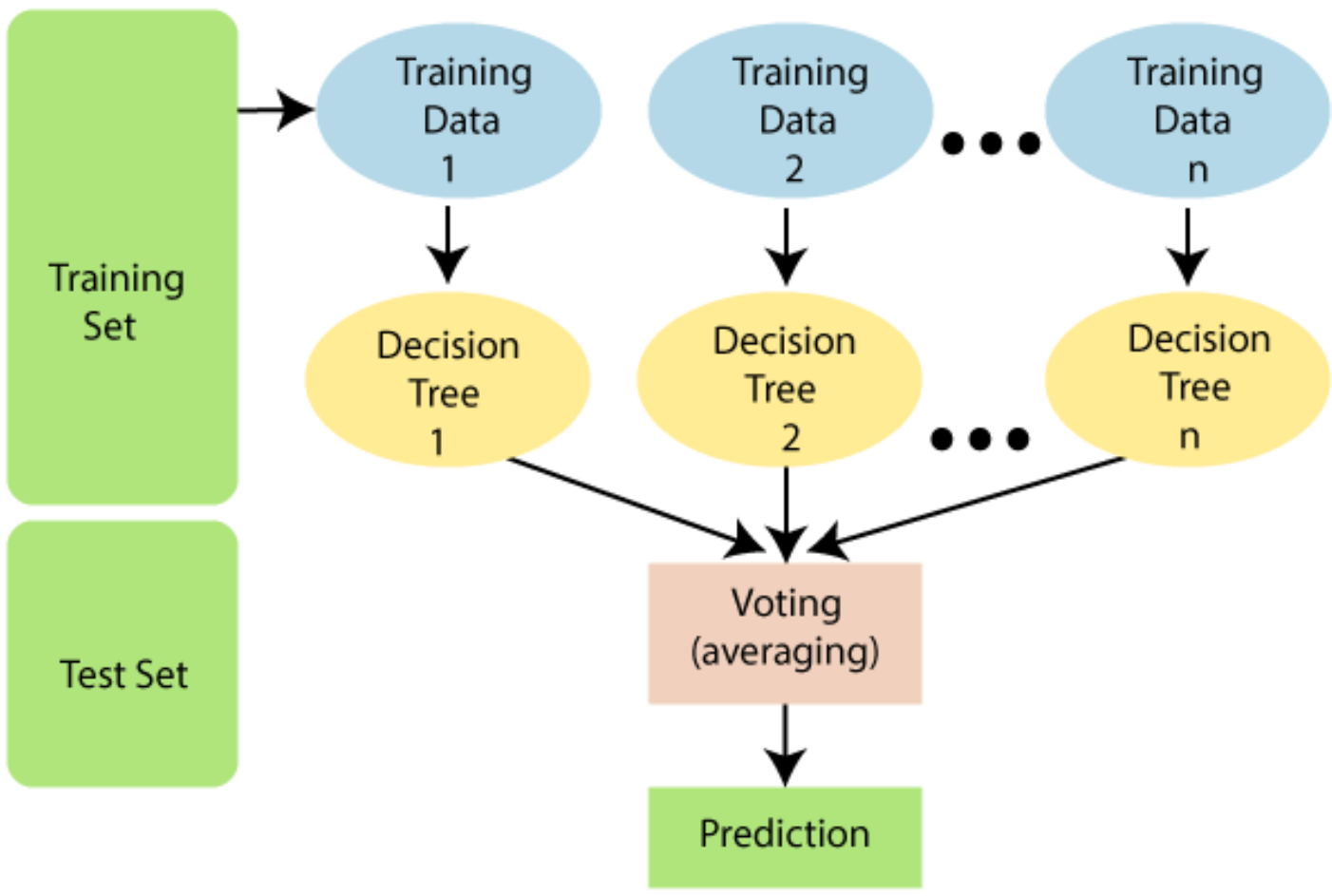




Model Training

RANDOM FOREST

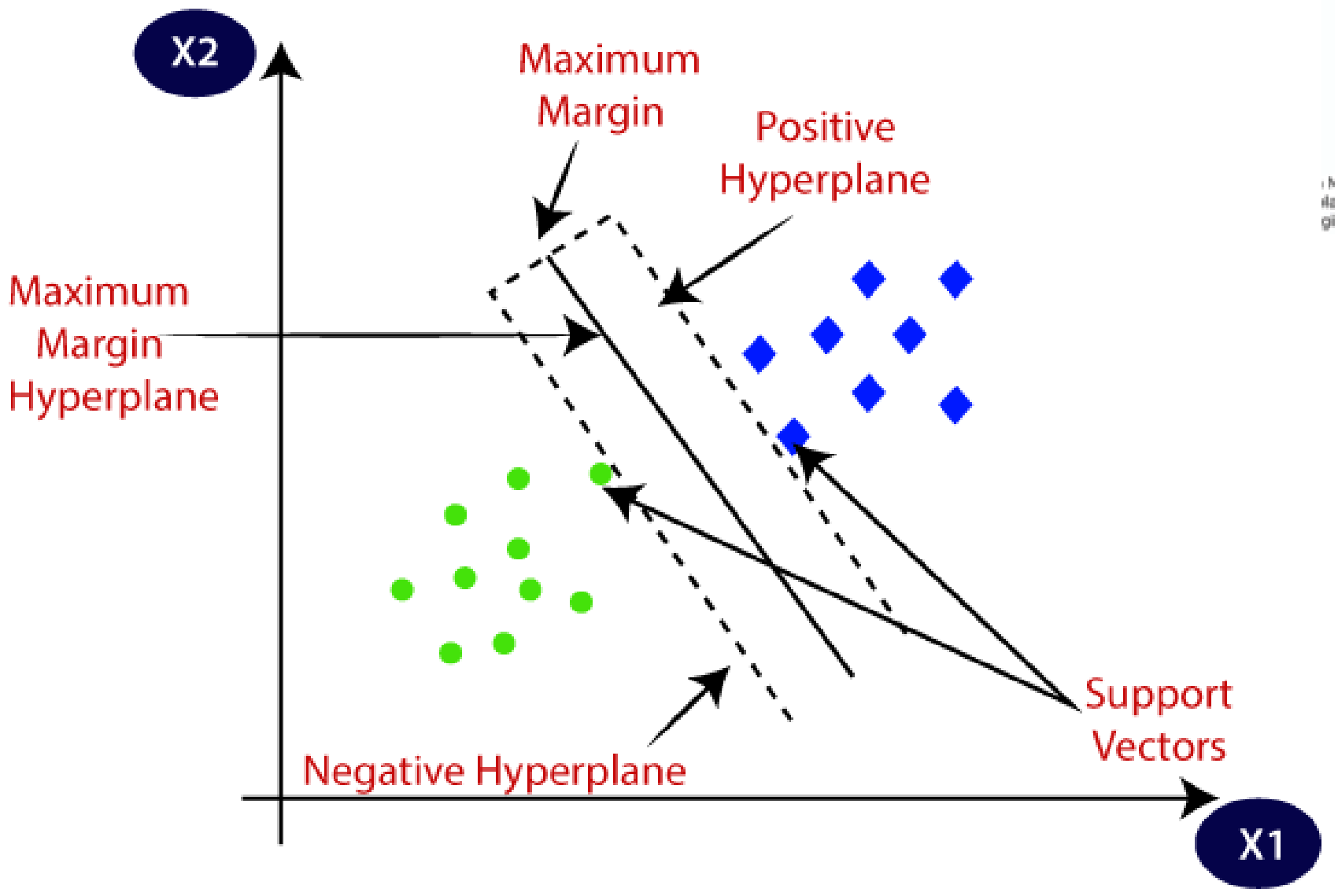
- used for both Classification and Regression problems in ML.
- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset



Model Training

SUPPORT VECTOR MACHINE (SVM)

- primarily, it is used for Classification problems in Machine Learning.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



Model Performance

The most common way to get a good model is to compare different models by calculating a performance metric on the **holdout data**.

PERFORMANCE METRICS FOR REGRESSION

Mean squared error (MSE)

- The mean model, which always predicts the average of the training data labels, generally would beused if there were no informative features

Median absolute error (MAE)

- If the data contains outliers, they can significantly affect the value of MSE. In such situations, it is better to apply MAE.

Almost correct predictions error rate (ACPER)

- It is the percentage of predictions that is within percentage of the true value.

Model Performance

A **confusion matrixis** a table that summarizes how successful the classification model is at predicting examples belonging to various classes. It shows how many True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) predictions.

PERFORMANCE METRICS FOR CLASSIFICATION

Precision

- is the ratio of TP predictions to the overall number of positive predictions

Recall

- is the ratio of TP predictions to the overall number of positive examples

Precision and recall are defined for binary classification, you can also use them toassess a multiclass classification model.
Some practitioners use a combination of precision and recall called **F-measure**, also known as **F-score**.

Accuracy

- is given by the number of correctly classified examples, divided by the total number of classified example.
- is a useful metric when errors in predicting all classes are judged to be equally important

Cohen’s kappa statistic

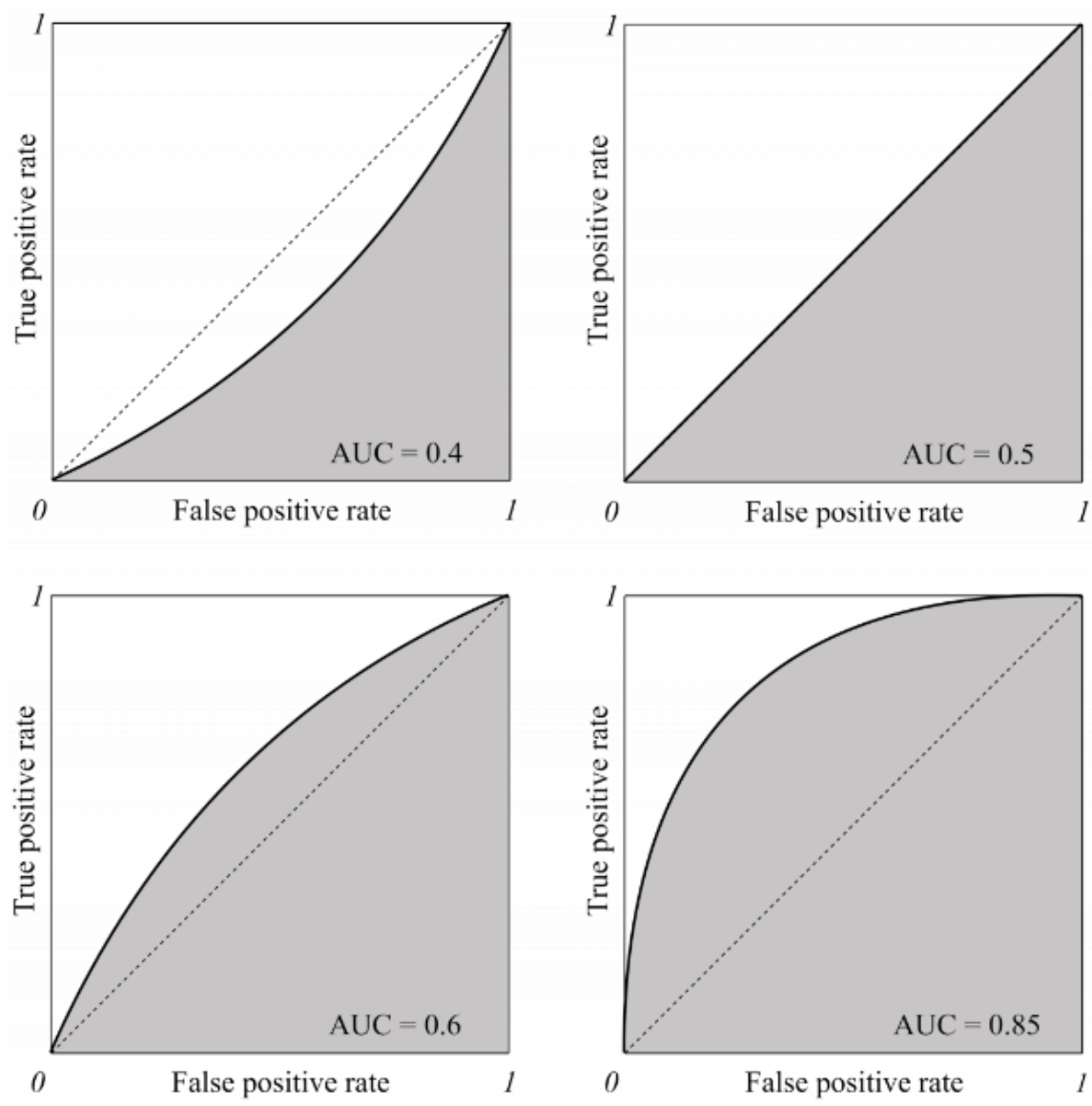
- is a performance metric that applies to both multiclass and imbalanced learning problems. The advantage of this metric over accuracy is that tells you how much better your classification model is performing, compared to a classifier that randomly guesses a class according to the frequency of each class.

Model Performance

PERFORMANCE METRICS FOR CLASSIFICATION

Area under the ROC curve (AUC)

- ROC curve is a commonly-used method of assessing classification models. ROC curves use a combination of the true positive rate and false positive rate, to build up a summary picture ofthe classification performance.
- ROC curves can only be used to assess classifiers that return a score (or a probability) ofprediction. For example, logistic regression, neural networks, and decision trees (and ensemblemodels based on decision trees) can be assessed using ROC curves.
- The greater thearea under the ROC curve(AUC), the better the classifier. A perfect classifier would have an AUC of1.



Model Performance

PERFORMANCE METRICS FOR RANKING

Precision and recall can be naturally applied to the ranking problem.

Discounted cumulative gain(DCG)

- is a popular measure of ranking quality in search engines. DCG measures the usefulness, or gain, of a document based on its position in the result list.

Cumulative gain(CG)

- is the sum of the graded relevance values of all results in a searchresult list

References

References

- http://www.mlebook.com/wiki/doku.php
- https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning
- https://www.javatpoint.com/machine-learning-random-forest-algorithm