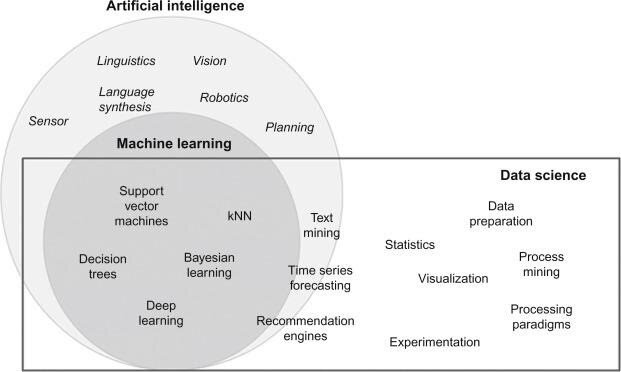
**Machine Learning with Python**

Machine Learning is the subfield of computer science that gives “computers the ability to learn without being explicitly programmed”.

AI is the greatest set where is inserted Machine Learning and which in turn contain the set of the deep learning algorithms.



The algorithms can be divided into supervised and unsupervised. Supervised algorithms are based on labeled data.

Examples of supervised algorithms are:

* Classification;
* Regression;

Unsupervised algorithms are based on non-labeled data.

Examples of unsupervised algorithms:

* Clustering;
* K-means;

Texto

Descrição gerada automaticamente com confiança média

**Regression:** The dependent variable (Y) must be continuous and can’t be a discrete value. However, the independent variable (X) can be either a categorical or a continuous value.

Observations about the methods of dividing data for training and test:

* Training and test the model on, the entire data set: May result in over-fit (the model is overly trained to the dataset, which may capture noise and produce a non-generalized model).
* Training and test the model on different portions of the data set: It’s just to partition the dataset in train and test samples that must be mutually exclusive. This will provide a more accurate evaluation on out-of-sample accuracy.

After analyzing the model accuracy be sure to train your model with the test data too. This is important for not to losing important data.

The method of split the data set on training and testing samples result in a more accurate out-of sample algorithm but it’s highly dependent on which dataset the data is trained and tested.

***Evaluation metrics in regression model:***

A model evaluation called K-fold cross validation resolves most of these issues performing multiple train/test splits where each split is different.

Mean absolute error: It’s just the average error.

Mean squared error: It’s the mean squared error and

Root mean squared error: It’s the square root of the mean squared error.

Relative absolute error: Also known as residual sum of square.

Relative squared error: It’s widely used because it is used to calculate the r square.

R-square: It’s not an error but represents how close the data values are to the fitted regression line.

The choice of metric depends on the type of model, the data type and the domain of knowledge.

Interface gráfica do usuário

Descrição gerada automaticamente com confiança baixa

**MULTIPLE LINEAR REGRESSION**

Is used When we want to measure the effect that the independent variables have on the dependent variable.

It’s also used when we want to predict the impact of changes.

There are some methods to find the parameters for the equation.

* Ordinary Least Square: Dataset < 10k rows.
* Gradient Descent: Dataset > 10k rows.
* Stochastic Gradient: Fastest than Gradient Descent
* Newton’s Method;

There are other methods too.

**\***Adding too many independent variables without any theoretical justification may result in an overfit model.

***How to prevent Overfitting?***

1. Training with more data:

Statisticians have conducted simulation studies\* which indicate you should have at least 10-15 observations for each term in a linear model.

For instance, if the regression model has two independent variables and their interaction term, you have three terms and need 30-45 observations. Although, if the model has multicollinearity or if the effect size is small, you might need more observations.

In short, the quality of the estimates deteriorates as you draw more conclusions from a sample.

1. Data augmentation;

An alternative to training with more data is data augmentation, which is less expensive compared to the former. If you are unable to continually collect more data, you can make the available data sets appear diverse. Data augmentation makes a sample data look slightly different every time it is processed by the model. The process makes each data set appear unique to the model and prevents the model from learning the characteristics of the data sets.

1. Cross-validation:

In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set.

1. Feature selection:

Some algorithms have built-in feature selection.

For those that don’t, you can manually improve their generalizability by removing irrelevant input features.

1. Regularization:

Regularization is a technique to discourage the complexity of the model. It does this by penalizing the loss function.

Regularization is three types

* L 1 or Lasso
* L 2 or Ridge
* L 3 or Elastic Net

If you’re really stuck in a situation where you ***have too many variables and too few observations, consider using principal component analysis*** to create a smaller set of indices you can model.

For linear regression, there is an excellent accelerated cross-validation method called ***predicted R-squared***. This method doesn’t require you to collect a separate sample or partition your data, and you can obtain the cross-validated results as you fit the model.

If there is a large discrepancy between the two values, your model doesn’t predict new observations as well as it fits the original dataset. The results are not generalizable, and there’s a good chance you’re overfitting the model.

**CLUSTERING**

Clustering means finding clusters ins a dataset, unsupervised.

A cluster is a group of objects that are similar to other objects in the cluster, and dissimilar to data points in other clusters.

***What is different between clustering and classification?***

Classification algorithms predict categorical classed labels while clustering is partitioning an unlabeled dataset into groups of similar objects.

Tabela

Descrição gerada automaticamente

***Uses of clustering***

* Exploratory data analysis
* Summary generation (reduce scale)
* Outlier detection
* Noise removal
* Finding duplicates
* Pre-processing step

***Examples of clustering***

Retail/marketing:

* Identifying buying patterns of customers.
* Recommending new books or movies to new customers.

Banking;

* Fraud detection in credit card use.
* Identifying clusters of customers.

Insurance:

* Fraud detection in claim analysis.

Publication:

* Auto-categorizing news based on their content.
* Recommending similar news articles

Medicine:

* Characterizing patient behavior

Biology:

* Clustering genetic markers to identify family ties.

***Clustering algorithms:***

* Partitioned-based Clustering:
  + Used for medium and large size database.
  + Relatively efficient
  + E.g: k-Means, k-Median, Fuzzy c-Means
* Hierarchical Clustering:
  + Used for small size dataset
  + Produces trees of clusters
  + E.g: Agglomerative, Divisive
* Density-based Clustering:
  + Especially good when dealing with spatial clusters or when there is noise at the dataset.
  + Produces arbitrary shaped clusters.
  + E.g: DBSCAN.

***k-Means:***

k-Means divides the data into ***non-overlapping*** subsets (clusters) without any internal division. Examples within a cluster are very similar and examples across different clusters are very different.

The algorithm tries to minimize the distance intra-cluster and maximize the distance inter-cluster.

How the distance is calculated depends on the data type and domain of knowledge to choose the meaningful distance measurement.

How this as it is an heuristic algorithm there is no guarantee that it will converge to the global optimum and the result may depend on the initial clusters. So, the result may be a local optimum, not necessarily the best possible outcome. To solve this problem***, it is common to run the whole process multiple times with different starting conditions***.

***k-Means accuracy***

One way to measure the accuracy of the model is to calculate the distance between data points within a cluster. So, we can change the numbers of initial number of centroids and plot a graph versus the mean distance of data point to cluster centroid.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

As increasing the k number will always decrease the mean distance so, we choose the right number of k when the gradient sharply change.

**Classification**

Classification is an unsupervised machine learning algorithm that classifies or categorizes some unknown items into a discrete set of classes.

The target attribute in classification is a categorical variable with discrete values.

Interface gráfica do usuário

Descrição gerada automaticamente

***Examples of classification***

* Loan default prediction.
* Spam filter.
* Image classification.
* Speech recognition.
* Handwriting recognition.
* Biometric identification.

***Examples of algorithms***

* Decision Trees (ID3, CD4.5, C5.0)
* Naïve Bayes
* Linear Discriminant Analysis
* K-Nearest Neighbor
* Logistic Regression
* Neural Networks
* Support Vector Machines (SVM)

***Pros and cons of some models:***

|  |  |  |
| --- | --- | --- |
| **Classification Model** | **Advantages** | **Disadvantages** |
| Logistic Regression | Probabilistic Approach, gives information about statistical significance of features. | The assumptions of logistic regression. |
| K – Nearest Neighbours | Simple to understand, fast and efficient. | Need to manually choose the number of neighbours ‘k’. |
| Support Vector Machine (SVM) | Performant, not biased by outliers, not sensitive to overfitting. | Not appropriate for non-linear problems, not the best choice for large number of features. |
| Kernel SVM | High performance on non – linear problems, not biased by outliers, not sensitive to overfitting. | Not the best choice for large number of features, more complex. |
| Naive Bayes | Efficient, not biased by outliers, works on non – linear problems, probabilistic approach. | Based in the assumption that the features have same statistical relevance. |
| Decision Tree Classification | Interpretability, no need for feature scaling, works on both linear / non – linear problems. | Poor results on very small datasets, overfitting can easily occur. |
| Random Forest Classification | Powerful and accurate, good performance on many problems, including non – linear. | No interpretability, overfitting can easily occur, need to choose the number of trees manually. |

***K-Nearest Neighbor***

Is an algorithm that takes a bunch of labeled points and use them to learn how label other points. Based on similar cases with same class labels are near each other. In this algorithm we must select the number of neighbors that will be used and the method to calculate the distance between the points.

One method that can be used is the Minkowski distance that is square root of the sum of the squared difference between each feature.

***How to choose the right number of ‘k’***

To choose the right number of ‘k’ we plot a graph of Accuracy vs k number.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

k-NN can also be used to predict continuous target. In this case the average or median target value of the nearest neighbors is used to predict.

The knn algorithm is a widely used algorithm. It is simple and easy to interpret. Since it does not make any assumption, it can be used to solve non-linear tasks as well.

On the negative side, we can say that the knn algorithm becomes very slow as the number of data points increases because the model needs to store all data points. Due to this reason, it is also not memory efficient.

Finally, it is sensitive to outliers because outliers also have a vote in decisions.

***Evaluation metrics in Classification***

***Jaccard index:*** It’s also known as Jaccard similarity coefficient.

Gráfico

Descrição gerada automaticamente

***F1-score:*** Can be shown as a confusion matrix which compare the model classification result against the true value. It considers the values of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), which in turn is used to calculate the precision and recall and then the F1-score. The average accuracy of the model is calculated as the average of the F1-score for all the labels.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

***Log loss:*** Measures the performance of a classifier where the predicted output is a probability value between 0 and 1. The classifier with lower log loss has better accuracy.

**Attention! \*\***Keep in mind that choose the best metric depends on several factors like if your data are highly imbalanced or not, if you care more about false positives or false negatives, if you want to rank models, etc. Look at the cheatsheet for choosing the right evaluation metric and the explanation about each one.

Diagrama

Descrição gerada automaticamente

**Decision Tree**

The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree.

To build a decision tree we use the training part in distinct nodes where one node contains all of or most of one category of the data.

Each internal node corresponds to a test.

Each branch corresponds to a result of the test.

Each leaf node assigns a classification.

***Decision Tree algorithm:***

1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go step 1.

This procedure is called recursive partition. What is important in making a decision tree, is to determine which attribute is the best or more predictive to split data based on the feature.

More predictiveness means less impurity on the leaves and lower entropy. Entropy, in this case, is a measure of randomness or uncertainty. The lower the entropy, the less uniform the distribution, the purer the node.

Entropy is calculated as follows: Uma imagem contendo Logotipo

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

**Logistic Regression**

Logistic regression is a statistical and machine learning technique for classifying records of a dataset based on the values of the input fields.

Logistic regression fits a special s-shaped curve by taking the linear regression function and transforming the numeric estimate into a probability with the following function, which is called the sigmoid function 𝜎.

Some characteristics of logistic regression are:

* Tries to predict a categorical or discrete target field instead of a numerical one.
* Independent variables should be continuous, if categorical they should be dummy or indicator coded.
* Can be used for binary or multi-class classification.

***Some applications:***

* Predicting the probability of a person having a hear attack.
* Predicting the mortality in injured patients.
* Predicting a customer’s propensity to purchase a product or halt a subscription.
* Predicting the probability of failure of a given process or product.
* Predicting the likelihood of a homeowner defaulting on a mortgage.

***When is logistic regression suitable?***

* If your data is binary: 0/1, yes/no, True/False.
* If you need probabilistic results.
* When you need a linear decision boundary.
* If you need to understand the impact of a feature.

***The training process***

Texto

Descrição gerada automaticamente

We need to change the value of theta in order to reduce the cost function. We can do this by using gradient descent function or other minimization functions.

We also need a parameter to stop the iterations. We can do this by using the accuracy of our model as the parameter.

To learn more about pros and cons of the solvers available for logistic regression in scikit-learn take a look at this: [https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-definitions/52388406#52388406](https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-definitions/52388406%2352388406)

**Support Vector Machines**

SVM is a supervised algorithm that classifies cases by finding a separator.

It’s a method used for classification.

***Advantages:***

* Accurate in high-dimensional spaces
* Memory efficient.

***Disadvantages:***

* Prone to over-fitting if the number of features is much greater than the number of samples.
* No probability estimates.
* Not computationally efficient. Not recommended if the data set contains more than 1000 rows.

***Some applications***

* Image recognition.
* Handwritten digits recognition.
* Text category assignment.
* Detecting spam.
* Sentiment analysis.
* Gene expression classification.
* Regression, outlier detection and clustering.

First it is mapping data to a high-dimensional feature space and then finding a separator for the data.

For this method to be used we must turn our linearly separable and this process is called kernelling and can be done by several functions. Unfortunately there is no easy way to know which one will perform better, so we usually choose different functions in turn and compare the results.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente