Introduction to Machine Learning

Eduardo Gomes (eduardo.gomes@m-iti.org)
GDG Workshop

m·iti
Madeira Interactive Technologies Institute

Contents

m·iti
Madeira Interactive Technologies Institute

What is Machine Learning?

Field of study that gives computers the ability to *learn* without being explicitly programmed.

Data

Information about the problem we are looking at

Algorithm

What we will train to give us our answers

Output

Answer of the algorithm

m·iti
Madeira Interactive Technologies Institute

What is Data?

Data

Information about the problem we are looking at

m·iti
Madeira Interactive Technologies Institute

What is Data?

- Information relevant to the problem at hand
- Requires a structure
  - E.g: Same scales used for data collection
- There are different types of data
  - NOIR

Data is crucial to a good ML system

NOIR - Nominal

Nominal data consists of labels but cannot be ordered nor can be calculated distances between

Example 1: *My favorite color is red*

Example 2: *My favorite color is blue*

We can calculate the **mode** of the data.

m·iti
Madeira Interactive Technologies Institute

NOIR - Ordinal

Ordinal data consists of labels that can be ordered there are no interpretable intervals between them.

Example:

- *Movie A got 2 stars*
- *Movie B got 3 stars*
- *Movie C got 4 stars*

It doesn't mean the Movie C is twice as better than Movie A!

We can calculate the **mode** and the **median** of the data.

m·iti
Madeira Interactive Technologies Institute

NOIR - Interval

Interval data is data measured using a scale, where every point is at the distance from one-another. However, there is no absolute zero.

Example:

- *Day 1 - 40 degrees*
- *Day 2 - 20 degrees*
- *Day 3 - 60 degrees*

The difference between Day 1 and Day 2 is the same as the difference between Day 1 and Day 3

We can calculate the **mode, median** and **mean** of the data.

m·iti
Madeira Interactive Technologies Institute

NOI**R** - Ratio

Ratio data is similar to Interval data, but with the existence of an absolute zero of the scale.

Example:

- *Person1: 20 years old*
- *Person 2: 40 years old*

Person 2 is twice as old as Person 1!

We can calculate what we could with Ordinal and do **multiplication** and **division** of variables

m·iti
Madeira Interactive Technologies Institute

NOIR - Summary

Offers:	Nominal	Ordinal	Interval	Ratio
The sequence of variables is established	–	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	–	Yes	Yes	Yes
Mean	–	–	Yes	Yes
Difference between variables can be evaluated	–	–	Yes	Yes
Addition and Subtraction of variables	–	–	Yes	Yes
Multiplication and Division of variables	–	–	–	Yes
Absolute zero	–	–	–	Yes

From: https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/

m·iti
Madeira Interactive Technologies Institute

What is Data - Structuring

	outlook	temp	humidity	windy	play
0	sunny	hot	high	False	no
1	sunny	hot	high	True	no
2	overcast	hot	high	False	yes
3	rainy	mild	high	False	yes
4	rainy	cool	normal	False	yes
5	rainy	cool	normal	True	no
6	overcast	cool	normal	True	yes
7	sunny	mild	high	False	no
8	sunny	cool	normal	False	yes
9	rainy	mild	normal	False	yes
10	sunny	mild	normal	True	yes
11	overcast	mild	high	True	yes
12	overcast	hot	normal	False	yes
13	rainy	mild	high	True	no

Attributes or Features

Samples or Instances

m·iti
Madeira Interactive Technologies Institute

Getting to know the Environment

1. Open Jupyter notebooks
2. Create a Python3 notebook
3. Import required packages:
    a. pandas
    b. numpy
    c. matplotlib.pyplot
4. Load the ‘tennis.csv’ file using pandas
    a. Tip: pandas.read_csv() can be useful
5. Look at the data
    a. Get some descriptive statistics
    b. Plot it!

m·iti
Madeira Interactive Technologies Institute

Getting to know the Environment - Tips

1. Create a pandas DataFrame by calling the DataFrame method
2. Can access DataFrame columns by either a dot or in the same way as a dictionary
    a. df[‘A’] or df.A are equivalent
3. Access row after column as in:
    a. df[‘A’][0] or df.A[0]
4. Filter DataFrame by values in a column:
    a. df.loc[df[‘A’] == 0]
5. Get descriptive statistics by using the describe() method on a DataFrame
6. There is a plot() method for DataFrames
    a. Uses matplotlib and can be tinkered with

m·iti
Madeira Interactive Technologies Institute

What are the Algorithms?

Algorithm

What we will train to give us our answers

m·iti
Madeira Interactive Technologies Institute

How to pick the algorithm to use?

First, define the problem:

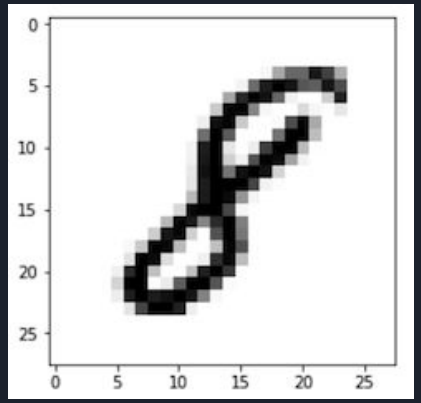
Is it a Classification or Regression problem?

m·iti
Madeira Interactive Technologies Institute

Classification

Where the output of the algorithm is a label

Outputs are discrete variables



Output:

‘eight’

m·iti
Madeira Interactive Technologies Institute

Regression

Where the output of the algorithm is a continuous value

Humidity previous hour: 71.2%

Humidity now: 67.2%

Rain Probability:

81.33%

m·iti
Madeira Interactive Technologies Institute

Types of Machine Learning

- Supervised Learning
  - Learns from examples (e.g. classifying an email as spam or not)
- Unsupervised Learning
  - Finding patterns in the data
- Reinforced Learning
  - Learn to perform actions (with rewards and penalties)

Supervised Learning

- Most explored type of ML
- Requires correctly labeled data
- Algorithms know their target

While training, an algorithm will always make use of the correctly labelled data (ground-truth) and try to obtain it using a set of inputs

Supervised Learning - Algorithms

Some algorithms that fall into this category:

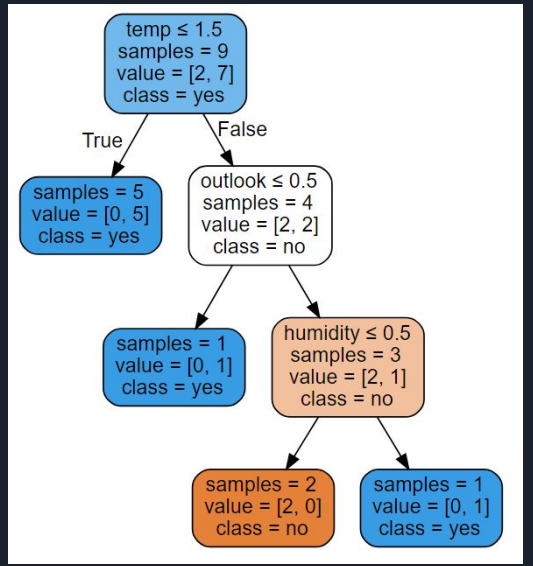
- Decision Trees
- Linear/Logistic Regression
- Support Vector Machines

In general:

- Algorithms in which you try to achieve a target value by a combination of any number of features as inputs

Supervised Learning - Decision Trees

- Builds a tree according to rules learned during training
- Usually used for classification
  - Can also be used for regression
- Simple and easy
  - Good for testing datasets
  - Can get surprisingly good results!



m·iti
Madeira Interactive Technologies Institute

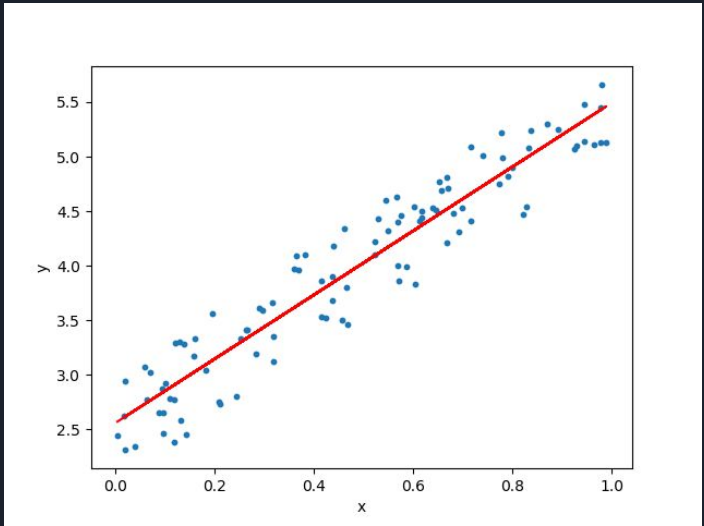
Supervised Learning - Decision Trees

1. Open a new Python3 notebook
2. Load the ‘tennis.csv’ file and turn it to categorical values
3. Instantiate a DecisionTreeClassifier
   a. Import it from sklearn.tree
4. Split the data between training and testing sets
   a. from sklearn.model_selection import train_test_split
   b. X_train, X_test, y_train, y_test = train_test_split(df[column], df[‘Target’])
5. Train! (hint: fit)
6. Woops!
7. Put the data in the right format (hint: LabelEncoder)
8. Train again!
9. Get the decision tree (tree.export_graphviz)

m·iti
Madeira Interactive Technologies Institute

Supervised Learning - Linear Regression

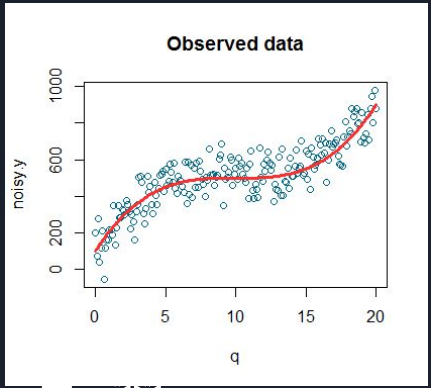
Fits a line through the data. Allows for the prediction of future inputs such as prices and other numerical values. Simplest case consist of a simple line equation.

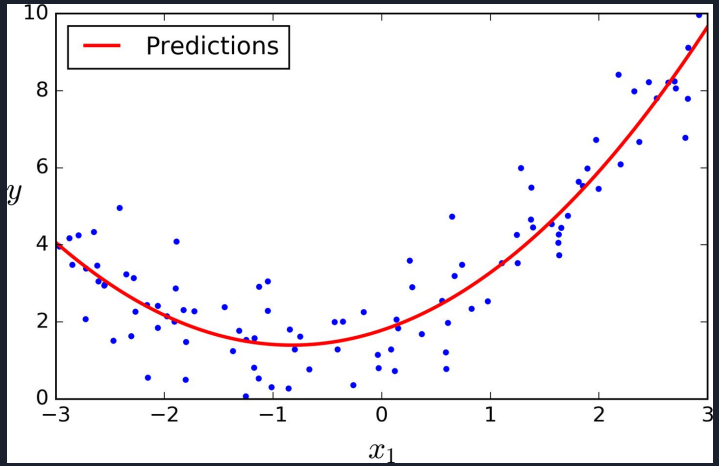


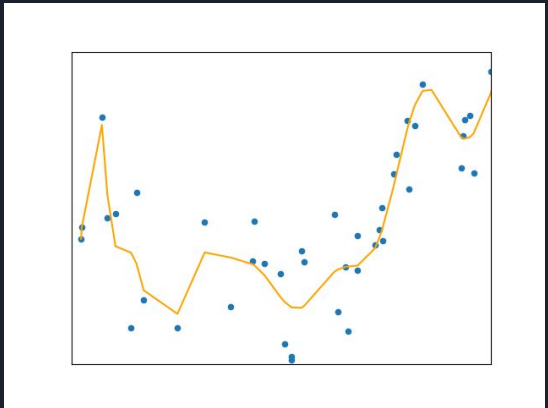
m·iti
Madeira Interactive Technologies Institute

Supervised Learning - Linear Regression

Often, one feature is not enough. Multiple Regressions and Polynomial Regressions allow the use of multiple features and can produce results like this:







Supervised Learning - Linear Regression

1. Open a new Python3 notebook
2. from sklearn.datasets import load_diabetes()
3. Use the imported function and save the dictionary
4. Use pandas DataFrame to turn the array of data into a DataFrame
5. Add the Target column
    a. hint: df['A'] = [.....]
6. Split data into training and testing
7. Instantiate a LinearRegression()
8. Fit
9. Score it! How did it perform?
    a. hint: model.score(X_train, y_train) or model.score(X_test, y_test)

m·iti
Madeira Interactive Technologies Institute

Supervised Learning - Linear Regression

1. Rerun the fit, but with all the variables in the X_train
2. Score it!
3. Plot it!

Supervised Learning - Logistic Regression

Exactly like Linear Regression but as a classifier. While using two classes, it translates into “squashing” the values of regression - typically between 0 and 1.

As with Linear Regression, it is possible to use multiple features to make a prediction.

Supervised Learning - Logistic Regression

1. Open a new Jupyter notebook
2. Import all the dependencies
3. from sklearn.datasets import load_digits
4. Try to plot first image
    a. Images in data are 8x8 but we have 64 value - rows….
5. Try to fit a LogisticRegression
6. How did it do?

m·iti
Madeira Interactive Technologies Institute

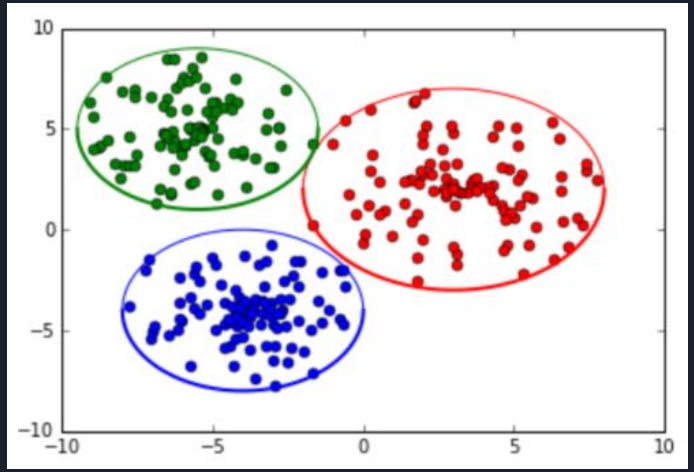
Unsupervised Learning

- Data is not labelled
- Mostly used to try to understand the underlying structure of the data
- Some more advanced ML systems use Unsupervised learning and then combine it with Supervised Learning

Unsupervised Learning - Algorithms

Most common form of unsupervised learning:

- Clustering
  - K-means

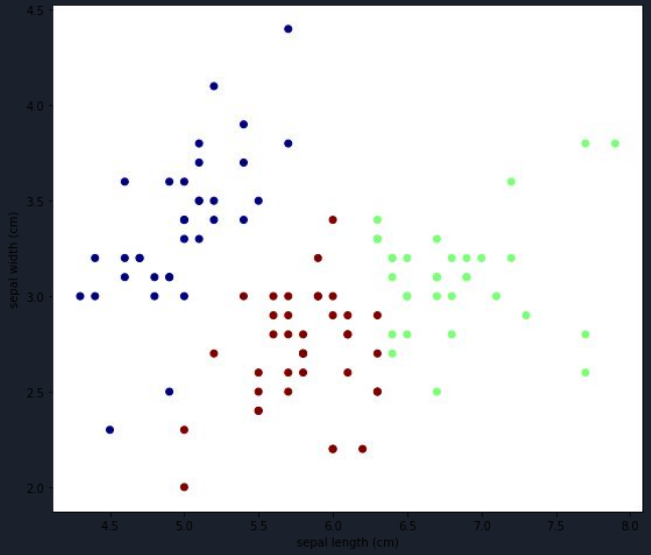


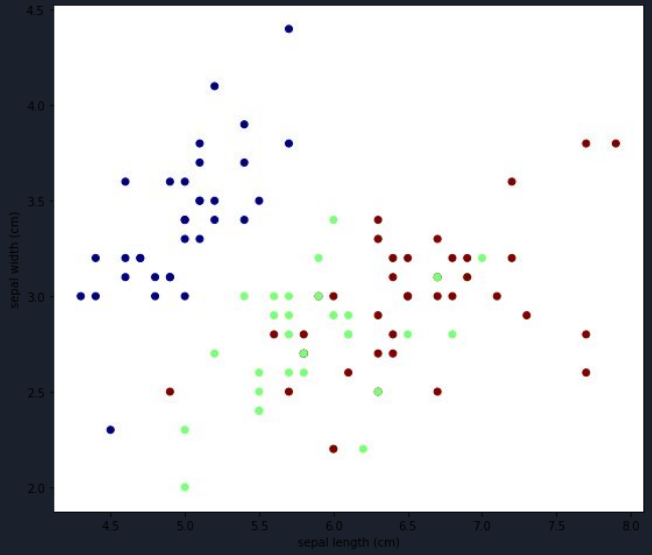
m·iti
Madeira Interactive Technologies Institute

Unsupervised Learning - K-means

The K-means algorithm iteratively re-arranges itself to find the optimal centroids of the cluster and defines new boundaries for instances.

Only the features are used during training. Labels were then compared after the training





m·iti
Madeira Interactive Technologies Institute

Unsupervised Learning - K-means

1. Open a new Jupyter notebook
2. Import all the required dependencies
3. Import the iris dataset
4. Split, instantiate, fit, etc…
5. Find the new labels on cluster.labels_
6. Compare them with ground-truth labels

m·iti
Madeira Interactive Technologies Institute

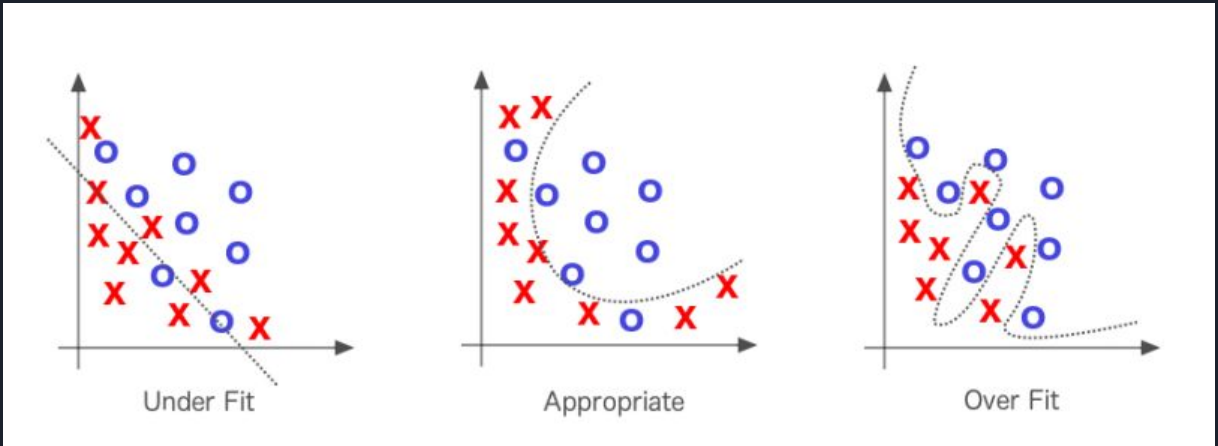
What is the Output?

Output

Answer of the algorithm

What is the Output?

Sometimes a model will not perform well. What happens then?

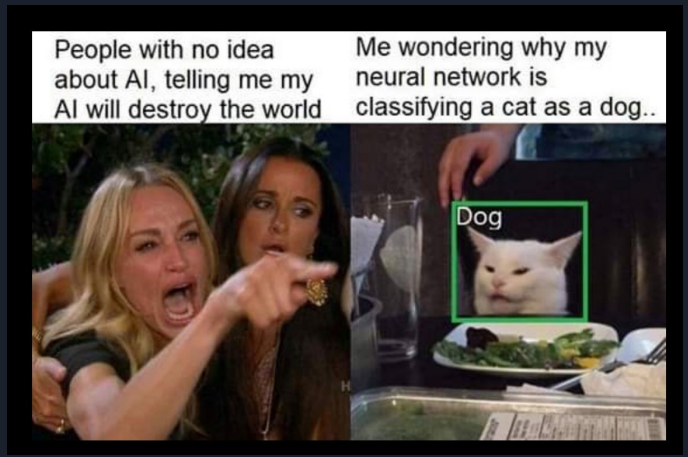


m·iti
Madeira Interactive Technologies Institute

What is the Output? - Underfitting

Underfitting:

- When the model did not learn the patterns in the data.
- Poor training performance - Poor testing performance
- Possible Reasons:
  - Not enough data
  - Not enough data quality
  - Not enough training
  - Bad algorithm choice



m·iti
Madeira Interactive Technologies Institute

What is the Output? - Underfitting

Underfitting:

- When the model did not learn the patterns in the data.
- Poor training performance - Poor testing performance
- Possible Reasons:
  - Not enough data - Get more data
  - Not enough data quality - Preprocessing may help
  - Not enough training - Train for more time
  - Bad algorithm choice - Try another algorithm

m·iti
Madeira Interactive Technologies Institute

What is the Output? - Overfitting

Overfitting:

- When the model learned “too well”.
- Great training performance - Poor testing performance
- Possible Reasons:
  - Trained for too long
  - Data might be too similar
  - Algorithm overfits easily

What is the Output? - Overfitting

Overfitting:

- When the model learned “too well”.
- Great training performance - Poor testing performance
- Possible Reasons:
  - Trained for too long - Train less
  - Data might be too similar - Cross-validation
  - Algorithm overfits easily - Regularizations

m·iti
Madeira Interactive Technologies Institute

What is the Output?

How can you evaluate and decide to trust (or not) an output?

Metrics!

Metrics

Metrics give us an idea of how well the model performs. There are different metrics for different tasks. Here are some examples:

- Classification:
  - Accuracy, Precision, Recall and F1-Score
- Regression:
  - MAE, MSE, RMSE and R-Squared

Metrics - Classification

Usually we use a confusion matrix to evaluate or results:

		Ground-Truth	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive (tp)	False Positive (fp)
	Negative (0)	False Negative (fn)	True Negative (tn)

m·iti
Madeira Interactive Technologies Institute

Metrics - Classification

Accuracy:

- How many predictions the model performed correctly in percentage.

Precision:

- The ability of the classifier not to label as positive a sample that is negative.

Recall:

- The ability of the classifier to find all the positive labels.

F1-Score:

- Weighted average of the precision and recall values.

Metrics - Classification

Precision:

- precision = tp / (tp + fp)
- How many positives we classified correctly

Recall:

- recall = tp / (tp + fn)
- Translates into how much we can trust a positive answer

F1-Score:

- f1 = (precision*recall) / (precision+recall)
- Harmonic mean of precision and recall

m·iti
Madeira Interactive Technologies Institute

Metrics - Regression

- Mean Absolute Error
  - Mean of the absolute differences of the predictions and ground-truth data
    - Value in the same units as the ground-truth
- Mean Squared Error
  - Mean of the squared differences of the predictions and ground-truth data
    - Very punishing for outliers in the data
- Root Mean Squared Error
  - Square root of the Mean Squared Error
    - Value in the same units as the ground-truth
- R-Squared
  - Goodness-of-fit indicating how well the model explained the data
    - Typically between 0 and 1 (can go negative depending on the platform where is is calculated)

Conclusions

1. Understand your data and formulate your problem
2. Classification or Regression?
3. Labels or no labels? Supervised vs Unsupervised
4. Model performance evaluation and Metrics
5. Iterate accordingly

Any questions/doubts?

m·iti
Madeira Interactive
Technologies Institute

Thank You!

m·iti
Madeira Interactive
Technologies Institute