Machine Learning Basics

As shown ML starts with existing data, so data is required to make any ML possible. This data is similar in the format later used as input when applying the model but is different in its concrete values. The model itself is trained once and then applied. Without any retraining, the model is not "learning" from any new data.

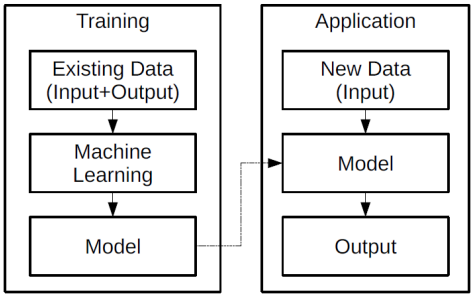

Figure 4: Basic ML overview (R. Findling)

Common Terms

There are many different terms introduced with machine learning here is an overview of some of the important terms:

- **Model:** relations in data that we model. In supervised learning: the regression/classification model we build = train from our data. Can be thought of as an "object"
- **Model type/class:** the class of the model, e.g. SVM, NN, KNN, etc.
- **Training:** process of modeling ("fitting") data using explicit model classes
- **Evaluation/Test:** evaluation of trained model object(s) on new data
- **Features or Predictors:** measurable property, e.g. timestamp, temperature, day-of-week, country code. In csv-files usually the columns (nr. of features = nr. of data dimensions = nr. of columns in .csv-files)
- **Sample:** one data instance, consisting of one concrete value for each property (in .csv-files the rows, which means all features of one row are one sample)
- **Label or target variable:** the property that should be predicted by a model.
- **Model parameters:** representation of a trained model (e.g. in- and export). This is what the model learns from data during training on its own
- **Hyperparameters:** parameters that influence how the model learns from data and the model's complexity. Speaking of parameter search/parameter tuning/model tuning usually means hyper-parameters. This is something we need/can specify for the model learning process

The Machine Learning Process

Those are the main steps in the machine learning process:

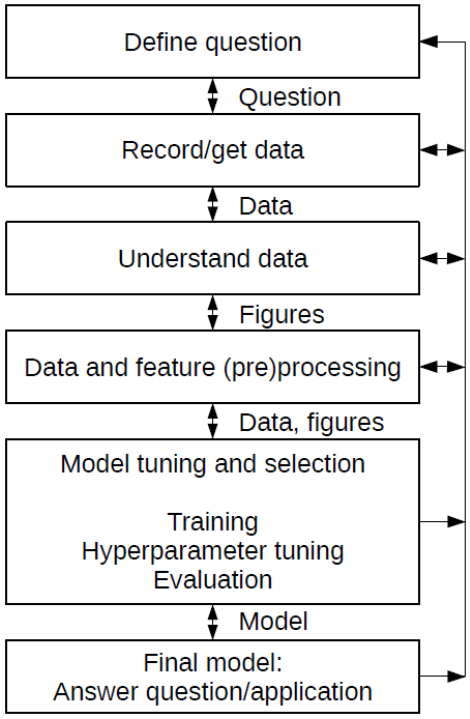

Figure 5: Machine Learning Process (R. Findling)

- **Question:** As in any other scientific process defining the question that should be answered is the first step. This is required to be sure that the right questions are answered by the created ML-models.
- **Acquire Data:** The data should be able to answer the defined question. This is often already recorded, but in some cases, it has to be recorded especially for the process.
- **Understand Data:** Understanding the data is key. The quality and quantity should be checked while exploring the data.
- **Preprocessing:** In many cases, the data has to be preprocessed. (e.g. interpolation, cleanup, filtering) In this step, additional high-level features can be introduced based on the existing features.
- **Model selection:** After the features were selected and preprocessed, (a) suitable model type(s) should be defined.
- **Model training:** During model training, the hyperparameters of the specific model type should be tuned to best fit the specified problem.
- **Evaluation:** To check the model's performance there is a wide variety of metrics available. Different metrics are used for different types of model/problem types.
- **Application:** The final model should be checked if it answers the question asked at the beginning. If it does it can be applied.

An example of how this process is applied: A visual introduction to machine learning (r2d3.us)

Ask the right questions!

When defining a question and using data, rethink if the results are plausible and realistic and answer your questions. ML algorithms try to find correlations but are not able to differentiate between correlation and causation. The human knowledge and intuition can't be replaced by an ML model.

An example question: **What are the side effects of drug X?**

- Asked for drugs for treating nausea
- Big database with drug usage and participant health records
- ... nausea often caused by drugs against leukemia, therefore
- patients who get leukemia drugs frequently get this drug too
- ... leads to an indirect correlation between the drug and leukemia
- An uninformed model might derive that drug causes leukemia > using ML outcomes without rethinking might be dangerous!

Those errors are called Type 3 errors: Providing the right answer to the wrong question.

Data Exploration

Data used for ML can have different sources and quality. Therefore the first step is to explore the data.

This is an example dataset which is represented in a tabular form as commonly provided by clients:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

Figure 6: Example Dataset (R. Findling)

This data is not yet clear enough for us to understand. It is not defined what each value represents - the values could be:

- Number of people dying from car accidents in Austria
- Average minutes of sun per day in Prague
- etc.

So to understand the data we need additional information on what each value represents and what unit those values have.

Graphs can also help to understand the data and the correlations. A simple line graph, as shown in Fig. 7, can show an overall trend in air passengers including some seasonal changes in the time series dataset. Those relations in the data are what ML uses to model our data. Understanding the data is crucial for using most ML tools and model types.

Values and Plots

To show the steps of data exploration we use Edgar Anderson's Iris Data (1935). The data is commonly used for ML examples and includes the measured length of different features of multiple iris species. It is a small and simple dataset including 5 features and 150 samples, a subset is shown in Fig. 8.

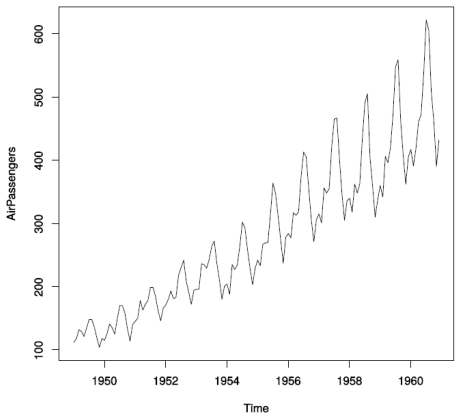

Figure 7: Example Dataset - Graph (R. Findling)

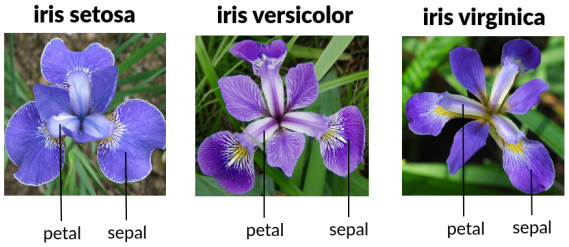

Figure 8: Iris Dataset (machinelearninghd.com)

The first steps for data exploration are:

- Is there any data missing? (NA etc.)
- Size of the Dataframe, data types, distribution of individual features
- Relations between the different features?

For the iris dataset, those first steps are shown in Fig. 9.

```
In [115]: iris.shape
Out[115]: (150, 5)

In [116]: iris.dtypes
Out[116]:
SepalLength    float64
SepalWidth     float64
PetalLength    float64
PetalWidth     float64
Name            object
dtype: object
```

```
In [117]: iris.describe()
Out[117]:
       SepalLength  SepalWidth  PetalLength  PetalWidth
count   150.000000  150.000000   150.000000  150.000000
mean      5.843333    3.054000     3.758667    1.198667
std       0.828066    0.433594     1.764420    0.763161
min       4.300000    2.000000     1.000000    0.100000
25%       5.100000    2.800000     1.600000    0.300000
50%       5.800000    3.000000     4.350000    1.300000
75%       6.400000    3.300000     5.100000    1.800000
max       7.900000    4.400000     6.900000    2.500000
```

Figure 9: Iris Dataset - first steps

Classification: This dataset includes 3 different classes: setosa, versicolor, viginica The samples are equally distributed between those classes with 50 samples each. An equal distribution makes problems easier. In many cases, the classes are inequally distributed (imbalanced).

Regression: A common regression problem could be to predict one dimension using all other dimensions/features of a sample. The distribution might be unequal or might not cover the complete target range. E.g. only bigger samples were recorded.

Statistical values

Mean, median, standard deviation (SD) and median absolute deviation (MAD) can help the scatter/dispersion of samples. Standard deviation is the mean difference of the values to the mean value. Median and MAD are considered more statistically robust. This means those values are more robust to outliers.

Scatterplot

A Scatterplot is a common way to show the relation between two different features by plotting all the samples. Commonly the single data points are color-coded to represent the different classes. As shown in Fig. 10 different features correlate differently and add more/less suitable for classification.

For each feature combination, we can create a unique scatterplot. This creates a scatterplot matrix as seen in Fig. 11. In this matrix a linear correlation between Petal.Width and Petal.Length can be seen. For Sepal.Length and Sepal.Width there is no such clear correlation. The different classes seem separable because there is not much overlap.

1D-Featureplot

A 1D featureplot only shows a single feature and its distribution including the different classes as seen in Fig. 12.

Density plot

The density plot shows the density of the distribution of a single variable. There is a kernel used to define how the density curve reacts to single points. The density differences can show the usefulness of a specific feature for classification. Overlapping densities show the similarity in the feature across classes which makes this feature less useful for classification.

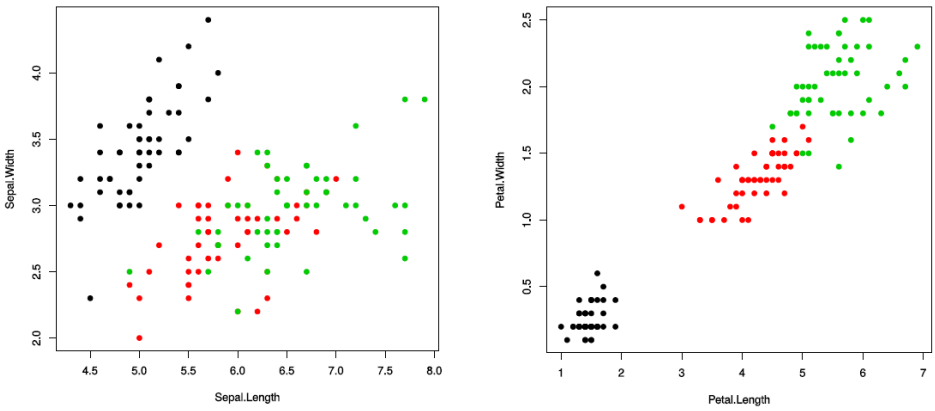

Figure 10: Iris Dataset - scatterplot

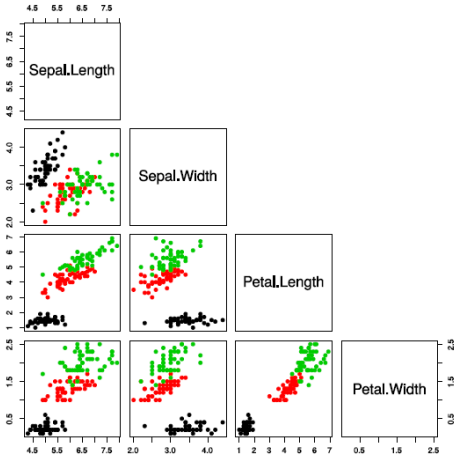

Figure 11: Iris Dataset - scatterplot matrix

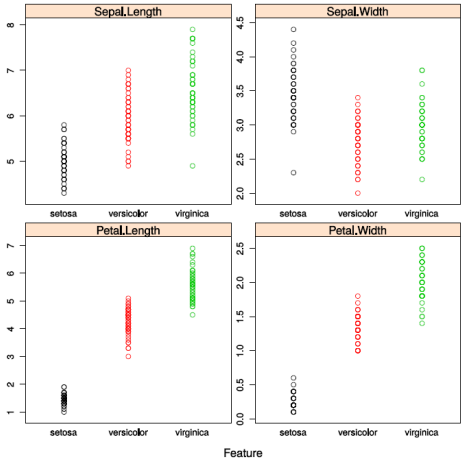

Figure 12: Iris Dataset - 1D featureplot

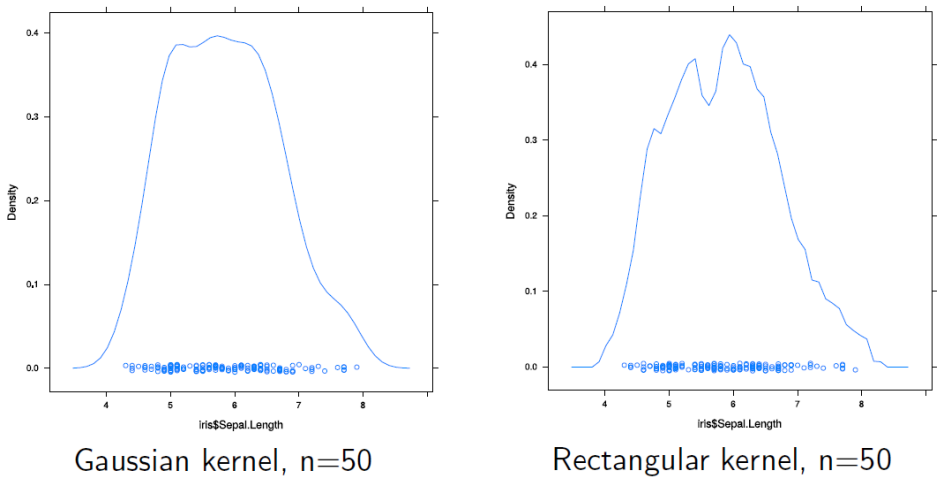

Figure 13: Iris Dataset - density plot

A special case for density plots is the histogram which uses ranges as "bins" to form single bars of frequency.

Boxplot

The boxplot (or box-and-whisker-plot) displays the distribution of a feature as quartiles. Each of the quartiles holds 25% of the data samples. The black line in Fig. 14 is the median with 25% of the values above and below in the box. Each whisker includes another 25%. The whiskers are usually max 1.5* the inner-quartile range of the closest quartile, so some values can be outside of those whiskers (also other definitions available). Those values are considered outliers.

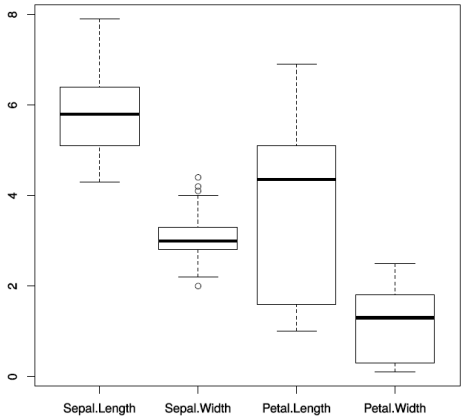

Figure 14: Iris Dataset - boxplot

Others

There are a lot of different plot types available, which are highly dependent on the data types and use cases. A level plot or contour plot can visualize 2D data on a coordinate system. Location-based data is best displayed on a map to show where the samples were recorded.