

Hands-on Machine Learning

Tutorial-2

- The next slide is a pointer to the different steps in the python notebook(*tutorial_2.1.ipynb*). This is just to give an overall picture of what we are doing in the notebook.
- Please make sure you are actually typing and going through the content to get a feel for it.

Contents

- **Get the data**
 - Create the workspace
 - Download the data
 - Take a quick look at the data structure
 - Create a test and training set
- **Discover and visualize the data to gain Insights**
 - Visualizing geographical data
 - Looking for correlations
 - Experimenting with attribute combinations
- **Prepare the data for machine learning algorithms**
 - Data cleaning
 - Handling text and categorical attributes
 - Feature scaling
- **Select and train a model**
 - Training and evaluating on the training set

Working with real data

- **Problem statement:** Build a model of housing price in California to **predict** the **median housing price** in any district
- **Data:** California census data

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0

median_income	median_house_value	ocean_proximity
8.3252	452600.0	NEAR BAY
8.3014	358500.0	NEAR BAY
7.2574	352100.0	NEAR BAY
5.6431	341300.0	NEAR BAY
3.8462	342200.0	NEAR BAY

Approach for solving the problem

- We are solving the problem as a supervised regression task, since the task is to predict a value rather than a class or category and the labels (the value to be predicted) are also available in the data

Reference

- Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems”
- <https://github.com/ageron/handson-ml>