



BERN WINTER SCHOOL – NATURAL LANGUAGE PROCESSING

Ahmad Alhineidi

Content

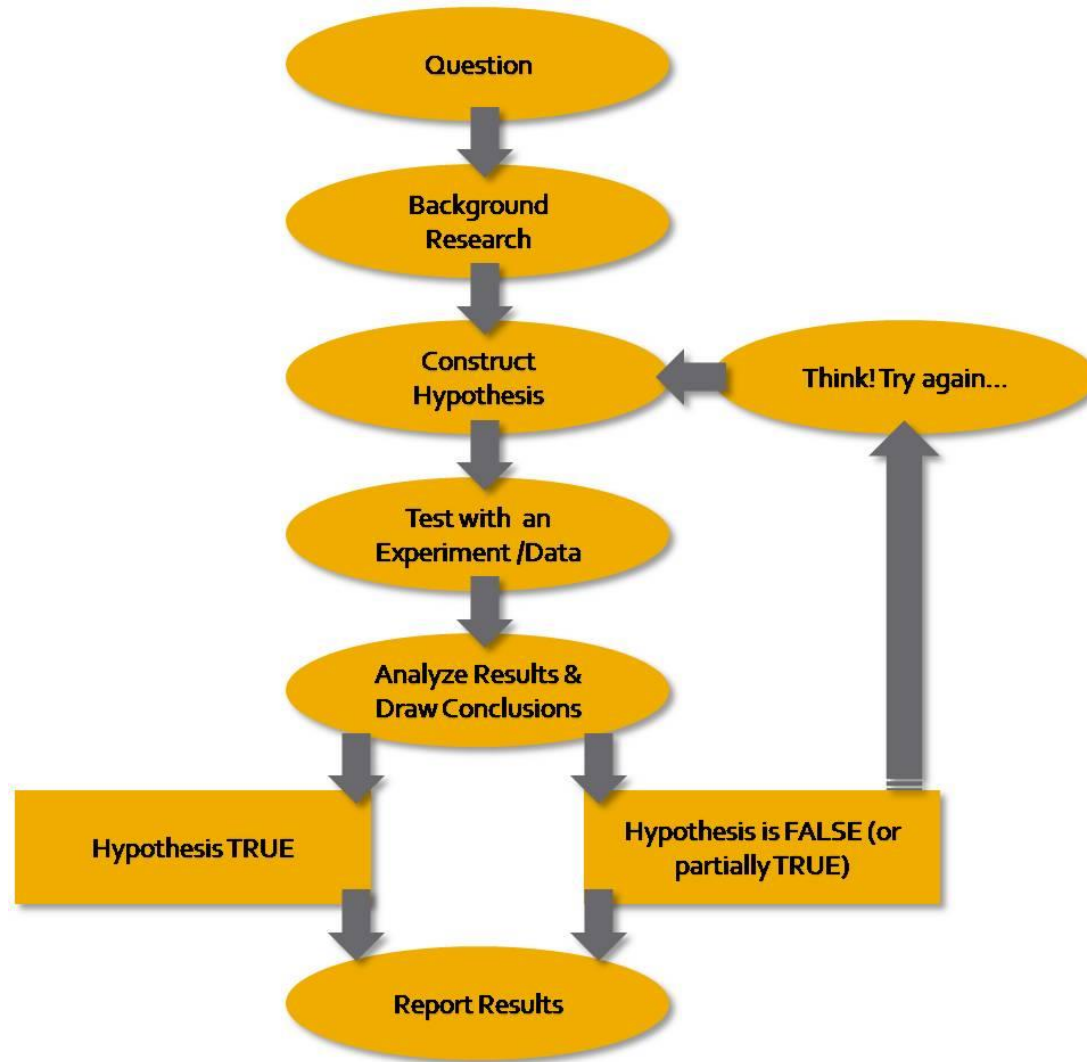
- Get to know each other
- Machine learning intro (Slides influenced or by PD Dr. Sigve Haug)
- Set up for the school (software installation, google colab, virtual machine, general info)

Materials on Github

- https://github.com/dsl-unibe-ch/Winter_School_NLP

Get to know each other

- Background (study - work - interest)
- Natural and unnatural languages you speak/know
- Interest in NLP
- Anything else?

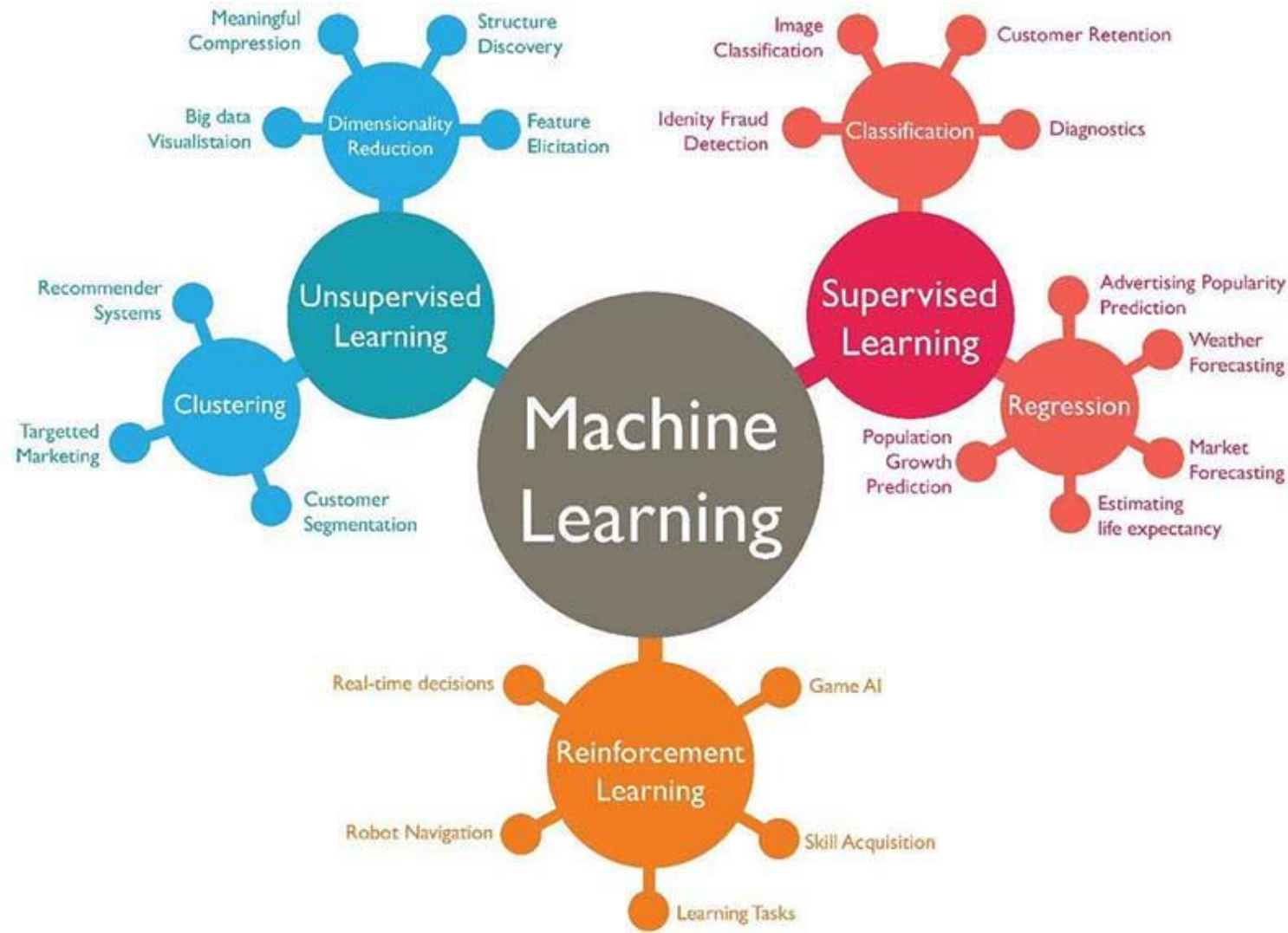


Machine learning Intro

- In science/data science the following loop represent the research process
- Machine learning has been popular method in the scientific process

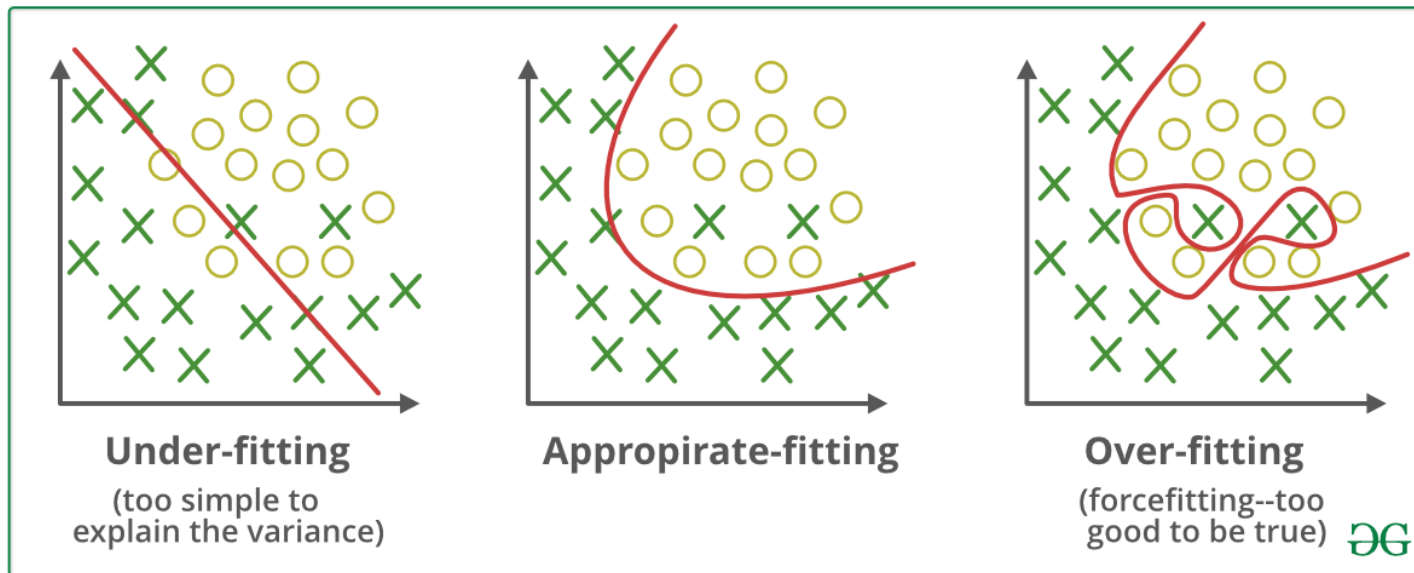
Machine learning Intro

- Typical machine learning task classification with some application examples



Machine learning intro

- Some (classical) machine learning algorithms:
- Supervised learning (Linear Regression, Logistic Regression, Decision trees, Random Forests, Support Vector Machine)
- Unsupervised learning (K-mean clustering, Principal Component Analysis, Apriori Algorithm:)
- **TASK:** Talk to your neighbor about 1 of the mentioned algorithms and fill out a slide [here \(click\)](#) with 2-3 sentences about what the algorithm does. If you don't know the algorithm, try to look it up using chatgpt or other language model. (15 – 20 min)



OVERFITTING, UNDERFITTING, APPROPRIATE FITTING

Source: <https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/>

Performance measure

- empirical measurements to evaluate the ML algorithm
- Some ML algorithms works better for certain tasks, while performs suboptimal in others
- ML algorithms performance is measured differently depending on its type, sometimes application
- Mean square error, mean absolute error, F1, accuracy, recall, precision, BLEU, others

Evaluation: Regression

We usually measure distance for evaluating regression

- Mean Square Error: $MSE = \frac{1}{n} \sum_i (y_i - \hat{y}(x_i^-))^2$
- Mean Absolute Error: $MAE = \frac{1}{n} \sum_i |y_i - \hat{y}(x_i^-)|$
- Median Absolute Deviation: $MAD = \text{median}(|y_i - \hat{y}(x_i^-)|)$
- Fraction of the explained variance: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}(x_i^-))^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y} = \frac{1}{n} \sum_i y_i$

		Prediction	
		Predicted 1	Predicted 0
Ground truth	Class 1	TP	FN
	Class 0	FP	TN

Evaluation: Classification

- We can use the confusion matrix to visualize the model prediction in a classification task

Evaluation: Classification

- We can use the F1 score as a good metric for evaluating a classification system.
- F1 takes into consideration both the precision and recall

- Accuracy = $\frac{TP+TN}{TP+FP+FN+TN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 = $2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP+FP+FN}$

Machine Learning example

- Google colab link in Ilias

Set up for school

- Install python

Why Python?

- Easy Syntax
- Free and Open Source
- Wide Support and Active Community
- String Handling
- Regular Expressions
- Natural Language Processing (NLP) Libraries
- Scraping Libraries
- Unicode Support
- Integration with Other Systems
- Machine Learning Libraries

```
from google.colab import files  
uploaded = files.upload()
```

```
from google.colab import drive  
drive.mount('/content/drive')
```

Use google colab?

- Don't need to install anything on your device
- Can load data from your device or connect colab to your google drive

Work on your machine?

- Use venv module to create virtual environment for the school
- `Python -m venv <env_name>`
- Activate the env: `source <env_name>/bin/activate`
- Run: `pip install -r requeriments.txt` #requirements.txt is in Ilias under “files, notebooks” folder
- Additionally run the following commands:

```
>>python -m textblob.download_corpora lite
```

```
>>python -m spacy download en_core_web_sm
```

```
>>python -m nltk.downloader stopwords punkt
```

- See you at the Aperero!