Introduction to Machine Learning

Mahammad Valiyev

15.01.2022

Contents and timeline

Recap

Part 1: Introduction to Machine Learning and use cases in O&G:

- ML vs traditional programming
- Key enablers of ML
- Major types of ML
- Intuition behind ML
- Power and limitations
- Use cases

Part 2: Overview of Machine Learning algorithms:

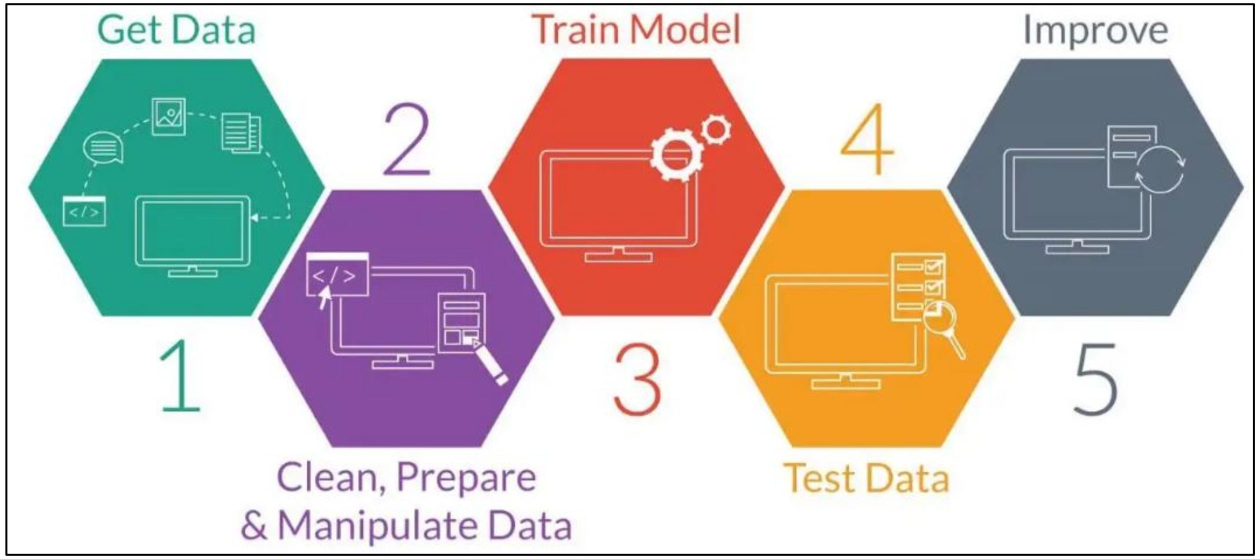
- Regression: Linear regression
- Classification: Logistic regression
- Clustering: K-means
- Deep Learning: Multilayer perceptron

Part 3:

Machine Learning Life Cycle

Machine Learning project workflow

1. Understanding the underlying the problem
2. Frame the problem into a Machine Learning problem
3. Get/collect data
4. Explore, visualize, prepare data
5. Select models (shortlist a few candidates), train and evaluate them:
6. Fine tune the best performing model
7. Present solution:



1. Understand the underlying problem

Key questions

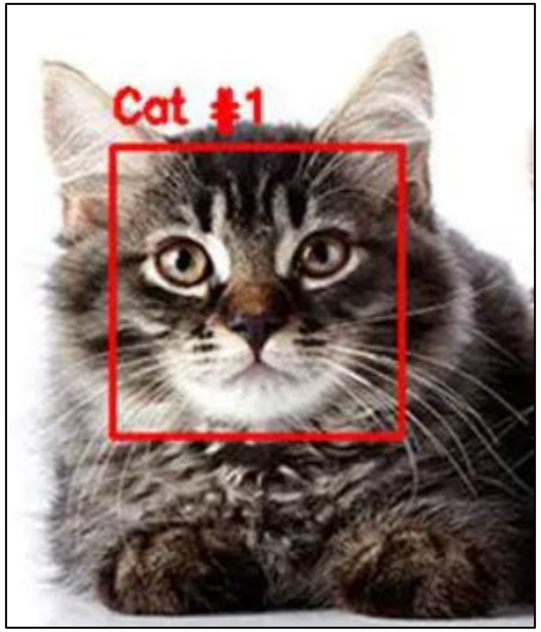
- What is the problem?
- What value will be derived from solution?
- How is the model going to be used? Who will use it?

Can the problem be solved with Machine Learning?

- Data of appropriate size and with useful features needed
- Input features need to have correlation with output

Example:

- Cat image detector



2. Frame the problem into Machine Learning problem

Key questions

- What are inputs and outputs?
- What type of Machine Learning task is it?
- What will be used as performance metric?

Performance metric

- Quantitative measure of degree of success
- Regression metrics:
  - MSE, MAE
- Classification metrics:
  - classification accuracy, precision/recall

Example:

- Cat image detector

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

3. Get/collect data

- Usually, one of the most time-consuming parts
- For supervised learning labeling is needed
- For most problems, data is much more important than algorithm!

Key features needed:
- Adequate size
- Representative of problem
- Informative features
- High-quality data (minimum errors, noise, outliers)

Source: freestampcatalogue.com

Example:
- Cat image detector

4. Explore, visualize, prepare data

Exploration

- Number and type of variables
- Range of values
- Missing data
- Outliers, anomalies

Visualization

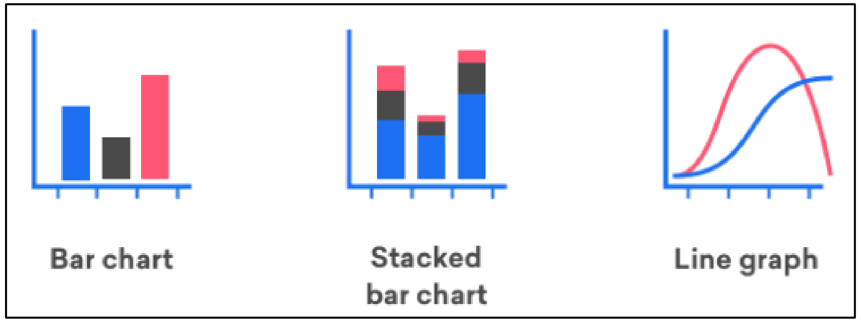
- Histograms
- Bar charts
- Specific plots depending on data

Prepare data

- Fix issues: e.g. missing data, outliers
- Normalize/standardize
- Feature engineering: select/create variables for modeling

Example:

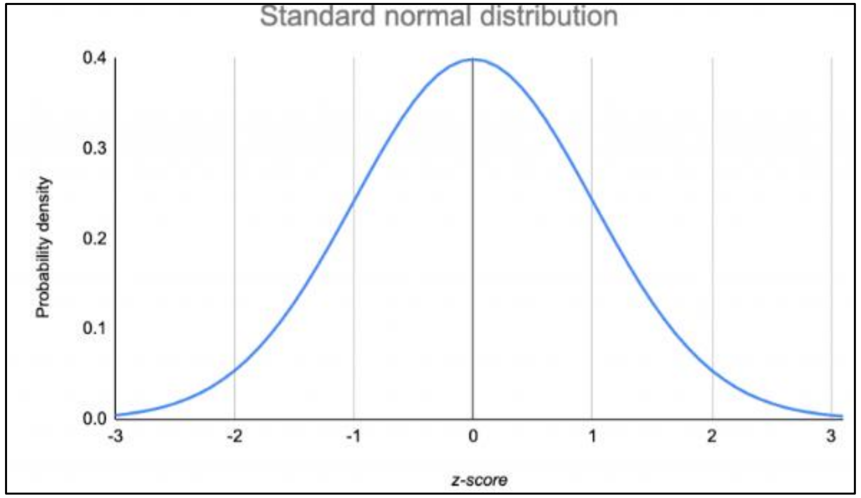
- Cat image detector



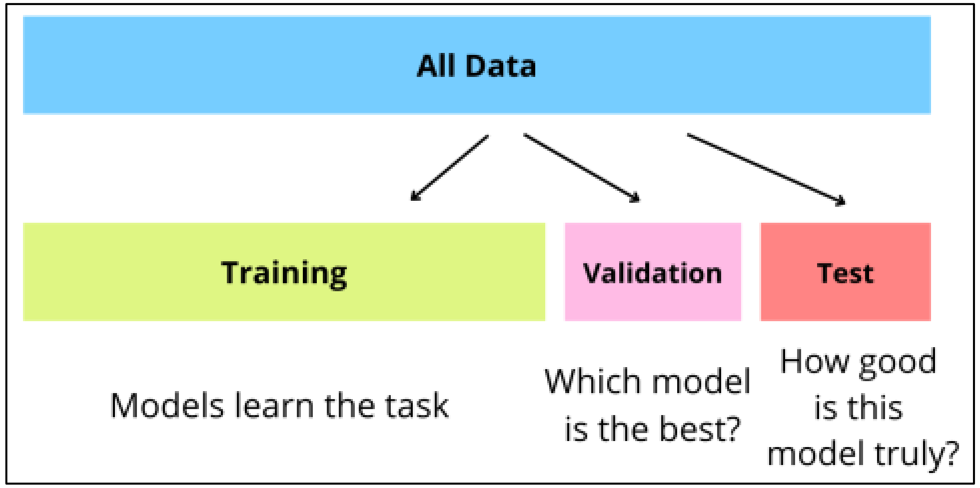
$$X_{normalized} = \frac{(X - X_{minimum})}{(X_{minimum} - X_{minimum})}$$

maximum

$$\text{Standardization: } z = \frac{x - \mu}{\sigma}$$



5. Select models, train, evaluate



Select a few possible models depending on:

- Machine learning problem type
- Complexity / Model interpretability
- Data availability

Train all shortlisted candidate models:

- Training refers to estimation of parameters of ML models

K Fold CV, K=5

Evaluate models:

- Using train/validation/test sets
- K-fold cross-validation

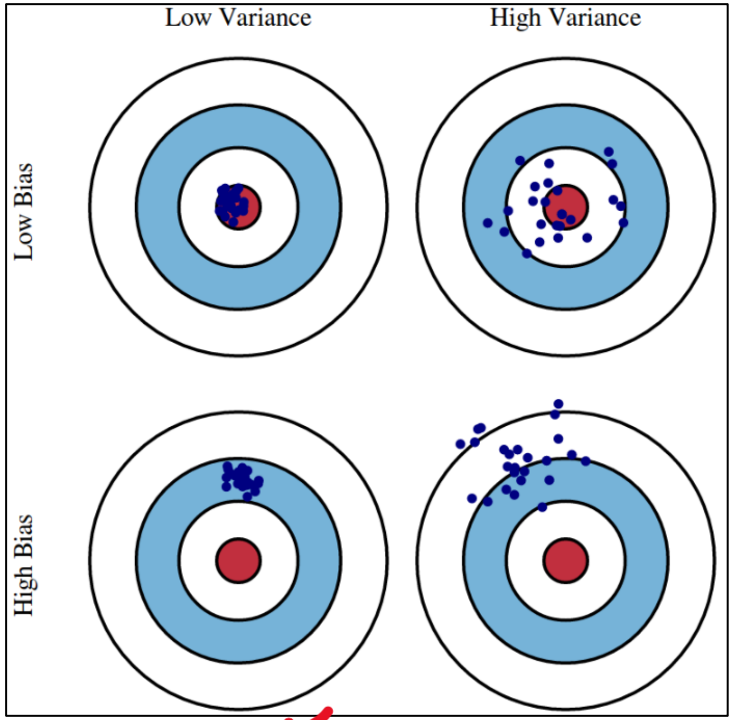
- Pick the best model

Example:

- Cat image detector

6. Fine tune the best performing model

- **Tune the best performing model** based on bias/variance trade-off
- **Tuning** means adjusting hyperparameters of a model
- **Hyperparameters** are not trainable parameters of a model
  - number of nodes and layers of a neural network model
  - number of clusters in k-means model

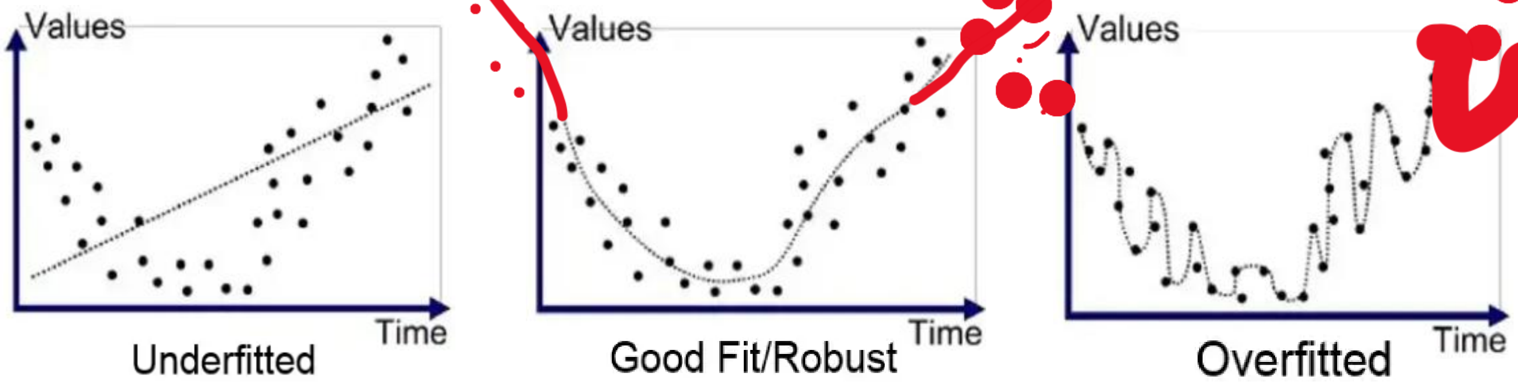


Underfitting: high bias, low variance
  - high training and test errors

Overfitting: low bias, high variance
  - Low training, high test error

Example

- Cat image detector



7. Present solution

Present key findings using:

- clear, easy to understand statements
- simple yet informative visualizations

Present/summarize for yourself:

- key lessons learned
- what worked, what did not
- what assumptions have been made
- scope for further improvement

Example

- Cat image detector

References and further resources

Books:

1. Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow, Aurélien Géron
- Chapter 1
- Chapter 2

2. Deep Learning with Python, François Chollet
- Chapter 6

3. The Hundred-page machine learning book, Andriy Burkov
- Chapter titled ‘basic practice’

Recap

Machine learning project workflow

1. **Understanding the underlying the problem**
2. Frame the problem into a ML problem
3. Get/collect data
4. Explore, visualize, prepare data
5. Select models, train and evaluate them: shortlist few candidates
6. Fine tune the best performing model
7. Present solution:

Thank you