Introduction to Machine Learning (in Education)

Dr. Jakub Kuzilek

Introduction

Today you will learn:

- How the machine learning can help educational institutions and what are its the most common applications.
- Definition of machine learning and its categorization.
- What is difference between supervised and unsupervised learning and how it is used.
- How to train the ML model and initial insights on its evaluation.

Introduction

How many university students fail to complete their degree?

Introduction

How many university students fail to complete their degree?

- It depends on the country, but in Germany approximately 25% of students do not complete their studies.

European Commission, Directorate-General for Education, Youth, Sport and Culture, Wollscheid, S., Stensaker, B., Jongbloed, B., et al., *Dropout and completion in higher education in Europe : main report*, Publications Office, 2015, **https://data.europa.eu/doi/10.2766/826962**

Why using Machine Learning in Higher Education?

- Make use of the data already collected at HEI
  - Student results, demographics, use of resources
  - Course portfolio, teacher data
  - …
- Building systems that get data to the right people

  → data informed decisions
- Supporting students during their educational path
- Personalize learning experience
- Improve drop-out rates
- Foster inclusivity
- Supporting teachers during their decision making
- Scientific knowledge about learning and learners

Learning Analytics

… the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs

- in short: research & practice area that uses computational analysis of data collected during the learning process to understand and improve learning
- the key fact is closing the gap between the data, analysis and intervention

Wolff; Hudson and Kortuem. The Role of Data Literacy within a MOOC Analysis. In: LAK'16: The 6th International Learning Analytics & Knowledge Conference, 25-29 Apr 2016, University of Edinburgh, Edinburgh

Educational Data Mining

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.

- the key is automated discovery with emphasis on reducing learning process to individual components
- focusing on automated adaption with no human in the loop

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Machine Learning

Machine Learning is:

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

~ Arthur Samuel, 1959

- Machine Learning is subfield of Computer Science
- Objective: *Generalize from experience*

- Machine Learning is learning model from set of observations

Terminology

- Sample
  - One realization of the observed phenomena (= one object).
  - Example: student
- Features
  - Set of possible measurements X of the observed phenomena
  - For 1 sample $x \in X$
  - Example: student results
- Labels
  - The set of possible categories or values Y assigned to the samples, which are of our interest
  - For 1 sample $y \in Y$
  - Example: indicator if student failed or passed the course
- Model
  - Decision function, which assigns each $x \in X$ decision $f(x) = \hat{y}$
  - $\hat{y}$ is the estimation of label delivered by model

Machine Learning

ML categories based on “feedback” available to learning system:

- Supervised learning
  - We know the right answers
  - Supervised learning algorithm is inferring decision function from labelled training data. The algorithm needs to generalize from training data to unseen data "reasonably".
- Unsupervised learning
  - We do not know right answers
  - Unsupervised learning algorithm is inferring function, which describes hidden structure of unlabelled data. We cannot estimate error of algorithm.
- Reinforcement Learning
  - Machine interacts with dynamic environment in which it needs to achieve certain goal without teacher telling it if it is close to the goal or not.

Unsupervised learning

- We do not know right answers.
- Unsupervised learning algorithm is inferring function, which describes hidden structure of **unlabelled** data.
- We cannot estimate error of algorithm

- Input:

$$T = \{x_1, x_2, x_3, \dots, x_N\}$$

- Output (depending on type):
  - Labels: $\{y_1, y_2, y_3, \dots, y_N\}$
  - Transformed data: $\{\hat{x}_1, \hat{x}_2, \hat{x}_3, \dots, \hat{x}_N\}$
  - Distribution: $f(x; \theta)$

Clustering Density estimation Dim. reduction

Unsupervised learning

- Applications:
  - News Sections
  - Computer vision
  - Medical imaging
  - Anomaly detection
  - Recommendation engines

  - Student profiles
  - Recommendation of study res.
  - Drop-out typology
  - …

- Challenges:
  - Computational complexity
  - Training time
  - Risk of inaccurate results
  - Needs human validation
  - Lack of transparency

Supervised learning

- We know the right answers
- Supervised learning algorithm is inferring decision function from **labelled** training data
- Input:

$$T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N)\}$$

- Output:

model

$$\hat{y}_i = f(x_i)$$

estimate

Classification

Regression

Supervised learning

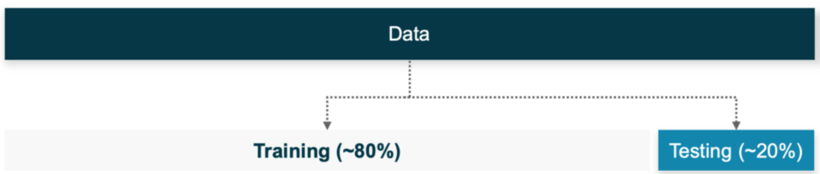
- Applications:
    - Image and object recognition
    - Predictive analytics
    - Customer sentiment analysis
    - Spam detection

    - Student pass/fail prediction
    - Assessment scores prediction
    - Estimation of drop-out rates
    - …
- Challenges:
    - Requires expertise
    - Time intensity
    - Human errors in labelling
    - Require the labels

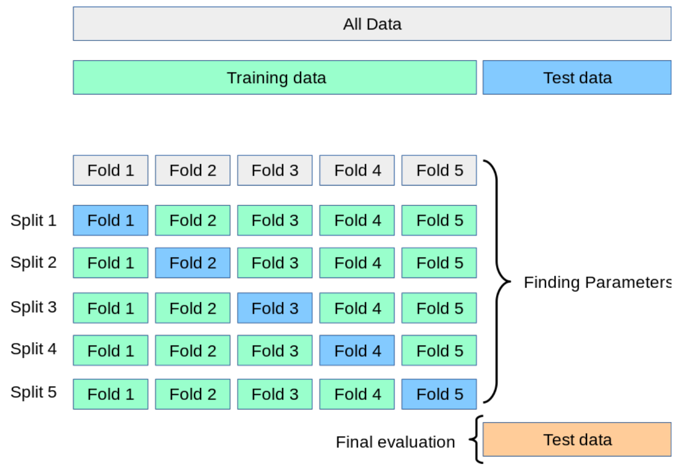
Model training

- During the model training **the sample of the population** is used
- It is critical to **estimate the model error** correctly

- To correctly train the model and estimate the error you need to **split data**:
  - Training data – used for estimation of model parameters
  - Validation data – used for evaluation (estimation of error) of model on “unseen” data

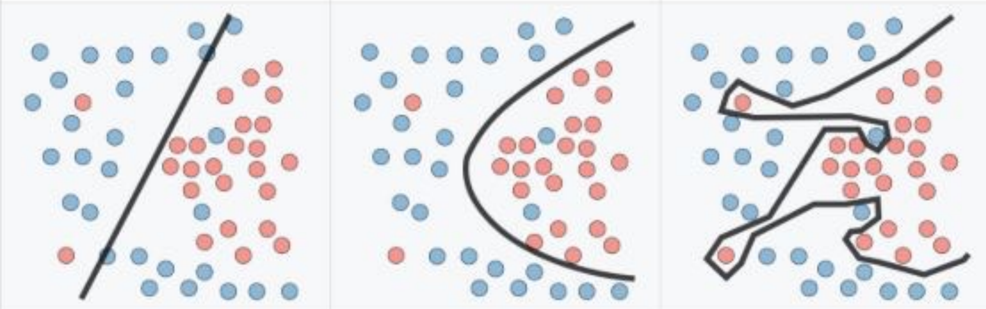
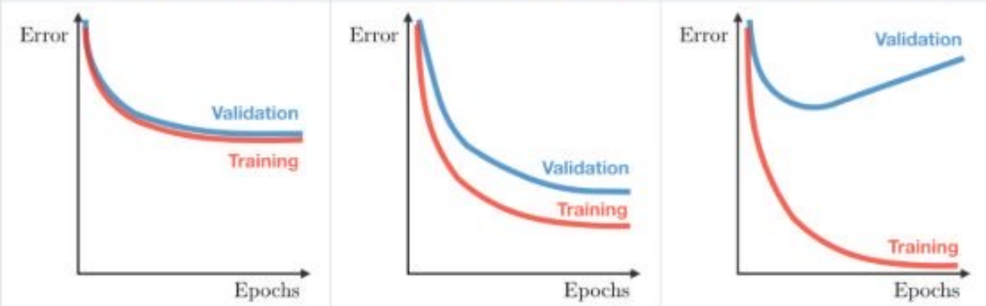


https://www.statworx.com/en/content-hub/blog/evaluating-model-performance-by-building-cross-validation-from-scratch/



https://scikit-learn.org/stable/modules/cross_validation.html

Model bias/variance

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

https://www.kaggle.com/getting-started/166897

Evaluation metrics

- For supervised learning

- Classification

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

- Regression
  - Mean Squared Error: $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
  - Root Mean Squared Error $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
  - Mean Absolute Error $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

Questions?