FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

Introduction to ML Modeling

LIAD MAGEN

Announcements

- Location Changes:
  - Saturday, 17.12.2022 – Online session
  - Friday, 13.02.2022 – Online session

- Reminder:
  - 1st Exercise Due on Monday

Last Week Recap

- What are the main challenges of NLP?
  - Variability, Ambiguity, Restrictivity
- What is the term for a model tendency towards an (often wrong) answer?
  - Model Bias
- What are the 5 levels of linguistic description?
  - Phonology, Morphology, Syntax, Semantics, Pragmatics

Today’s Agenda

> Canonical Learning Types

> Features Function // Feature Extraction

> Supervised Machine Learning Models:

> > Decision Tree

> > Random Forest

> Model Evaluation

> Statistical Models (Recap)

> > Maximal Likelihood Estimation (MLE)

Terms Today

- Feature Extraction
- Supervised Machine Learning
- Decision Tree
- Accuracy
- Precision
- Recall
- F1-Score
- Random Forest

Machine Learning

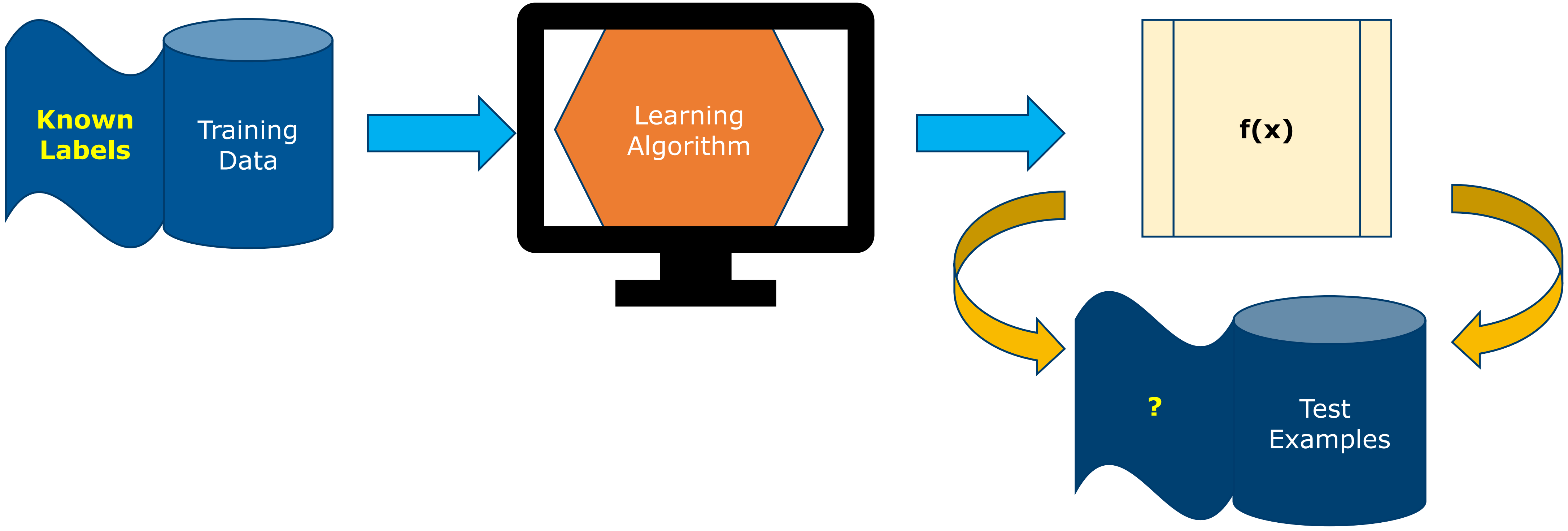
> What is "Learning"?
> How do we learn?
> How does one design a reasonable exam to evaluate learning?

Reduce **memorization** and encourage **generalization**

General Recipe for Modeling

- Definition of the problem
- Collecting (Historical) Data
- Analyzing the data (statistics)
- Model specification:
  - Feature Selection
  - Model Selection
  - Parameter Estimation
- Model application (inference // predictions)

General Recipe for Modeling with Machine Learning



Canonical Learning Types

Which ones are **continuous,** and which are **discrete**?

Regression

Classification

- Binary Classification
- Multiclass Classification
- Multilabel Classification

Ranking

Canonical Learning Types

Each has a different way of measuring the **error**

Regression
Continuous

Classification
Discrete

- Binary Classification
- Multiclass Classification
- Multilabel Classification

Ranking
Both/Depends

Canonical Learning Types

Each has a different way of measuring the **error**

Regression
Continuous

Classification
Discrete

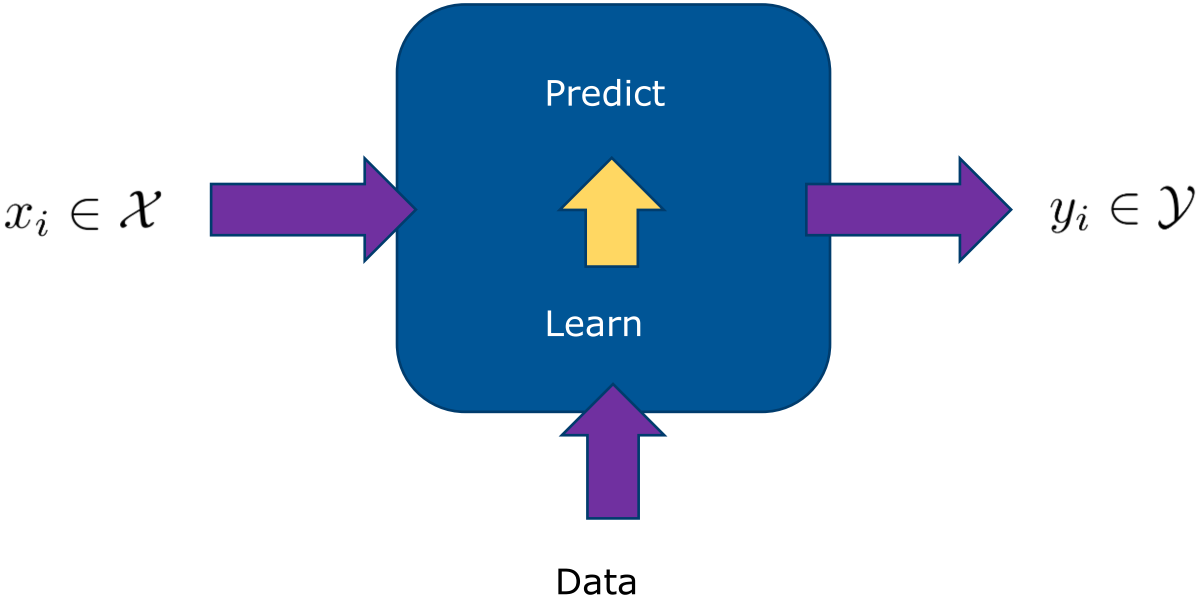
Ranking
Both/Depends

Binary Classification

Multiclass Classification

Multilabel Classification

Classification with ML



$$\{(x_i, y_i) | x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^{|D|}$$

Classification with ML

- We are given data samples: $x_1, x_2, \dots, x_n \quad x_i \in X$
- And their corresponding labels: $y_1, y_2, \dots, y_n \quad y_i \in Y$
- We train a function f: $f: x \in X \rightarrow y \in Y$
- The data-point x is represented by 'features': $f: \phi(\mathrm{x}) \in R^m \rightarrow y \in Y$

Feature Function

Feature Function

How do we represent an object?

f()

Feature Function

Perform **measurements** and obtain **features**

$\varphi($  $)$

= (1.3, 34, 8.2, #ff0000)

(Diameter, weight, softness, color)

Feature Function

Perform **measurements** and obtain **features**

> **Indicator** Features / **1-hot** vector / **binary** features

$\varphi($  $)$

= (0, 1, 0,   0, 1, 0,0,   0, 1, 0,   1, 0, 0, 0)   ← Buckets: (0-1, 1-2, 2-3…)

(Diameter,   weight, softness,   color)

Feature Function

For text?

What can we measure over **Text**?

Types of Classification Problems

- Binary: $y \in \{-1,\ 1\}$
- Multi-Class: $y \in \{1, 2, \ldots, k\}$
- Multi-Label: $y \in 2^{\{1,2,\ldots,k\}}$
- (Regression…?)

Types of classifiers

> Generative vs Discriminative
> Probabilistic vs Non-Probabilistic
> Linear vs non-Linear

$P(x, y)$

$P(y \mid x)$

$score(x, y)$

$f(x) = y$

FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

Types of classifiers

- > Generative vs Discriminative
- > Probabilistic vs Non-Probabilistic
- > Linear vs non-Linear

$P(x, y)$	Generative
$P(y \mid x)$	Discriminative
$score(x, y)$	Discriminative
$f(x) = y$	Discriminative

Types of classifiers

- Generative vs Discriminative
- Probabilistic vs Non-Probabilistic
- Linear vs non-Linear

prob	$P(x, y)$	Generative
prob	$P(y \mid x)$	Discriminative
Non-prob	$score(x, y)$	Discriminative
Non-prob	$f(x) = y$	Discriminative

Popular Classifiers

- kNN (k-Nearest Neighbors)
- Decision Trees
  - Decision Forests
  - Gradient-boosted Forests
- Logistic Regression
- SVM
- Neural Networks

Scikit-learn (sklearn):
a popular and good package for
those activities

Generic NLP Solution

> Find an annotated corpus
> Split it into train/dev & test parts
> Convert it to a vector representation
> Decide on the output type
> Decide on the features
> Convert each training example to a feature vector
> Train a machine learning model on the training set
> Apply your model on the test-set
> Measure the accuracy

Generic NLP Solution

> **Find an annotated corpus**

- Difficult to create your own corpus (expensive)
- *Decide* what are you classifying?
  What should the output classes be?
- *Consider*: is the problem even solvable?
  Can humans do that?
  At what level of accuracy can humans do it?

FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

Example #1

- Problem Definition:

  Given a person’s name, determine if it is a **Male** or a **Female**.

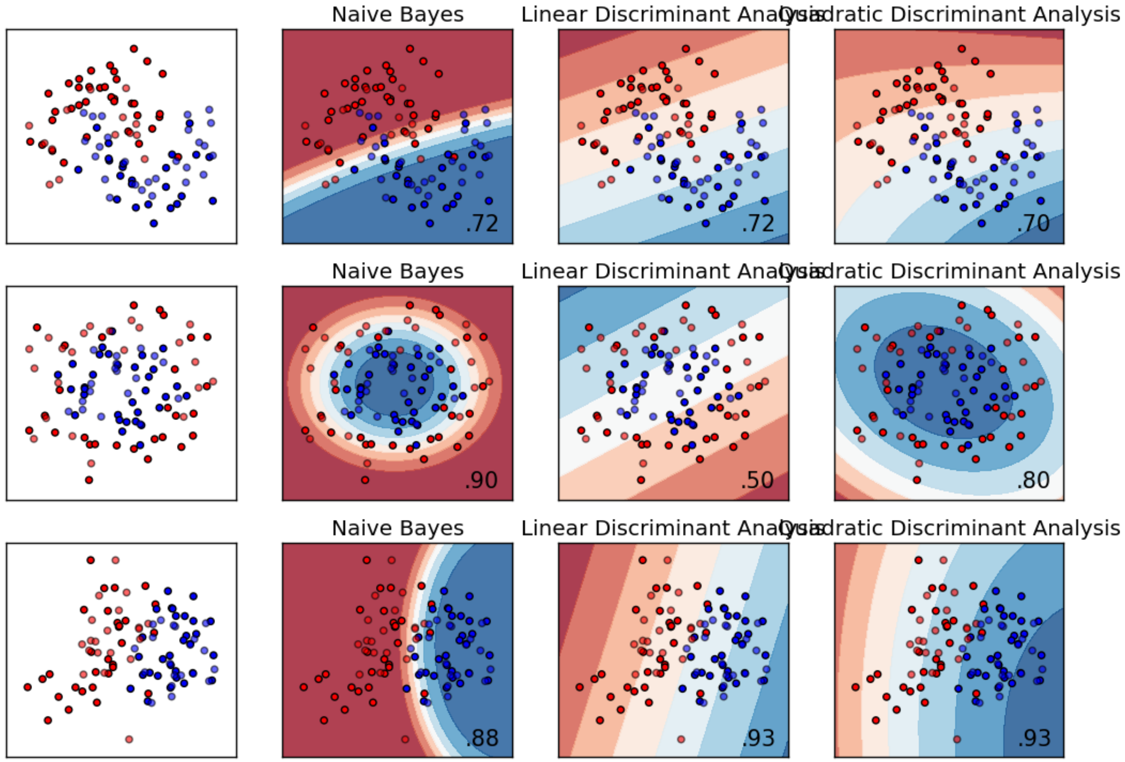
  Why?
  Possessive pronouns, Anaphora/Cataphora

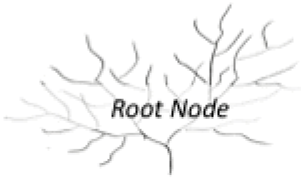
- Data:
  - A list of ~8000 names in *English* collected from a population administration data source
  - ~5k Female
  - ~3k Male

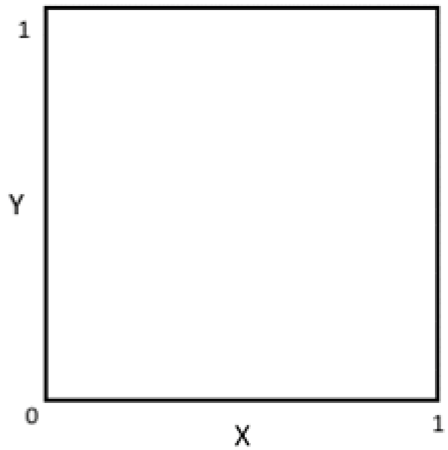
Decision Tree – Basic Idea

1. Begin the tree with a root node.
2. Identify a binary question for data splitting.
3. Split the data into two subsets based on the identified question
4. Repeat creating questions and splitting the remaining data, until you cannot further classify the nodes.
   Call the final node as a leaf node.

Decision Tree & Decision Boundary







For more tutorials: annalyzin.wordpress.com

Model Evaluation

> The performance of the learning algorithm should be measured on unseen **"test" data**.

> The data that our algorithm "sees" at **training** time and the one it "sees" at **test** time should be related:

- Drawn from the same distribution.
- (Hopefully represent the real-world data)

Model Evaluation

		Predicted Class	
Actual Class		Class =Yes	Class=No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

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

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

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

$$\text{F1 Score} = \frac{2*Recall*Precision}{Recall+Precision}$$

$$\text{MCC} = \frac{TN\times TP-FN\times FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Model Evaluation

- Performance measurement depends on the problem we are trying to solve:

  - *Classification*:
    - F-Score
    - Accuracy
    - Precision / Specificity
    - Recall / Sensitivity
    - AUC
    - …

  - *Regression*: Mean Squared Error (MSE)

Open Questions

> Does the model pick the best feature for splitting?
> Does the order of the features matter?
> Is there randomization involved?
> Can we do better than the Decision Tree?

Decision Tree Algorithm – Diving Deeper

- Step #2: Identify a binary question for data splitting.
- How?

- GINI-index or Entropy:
- Measuring the 'correctness' for every decision – of every feature.

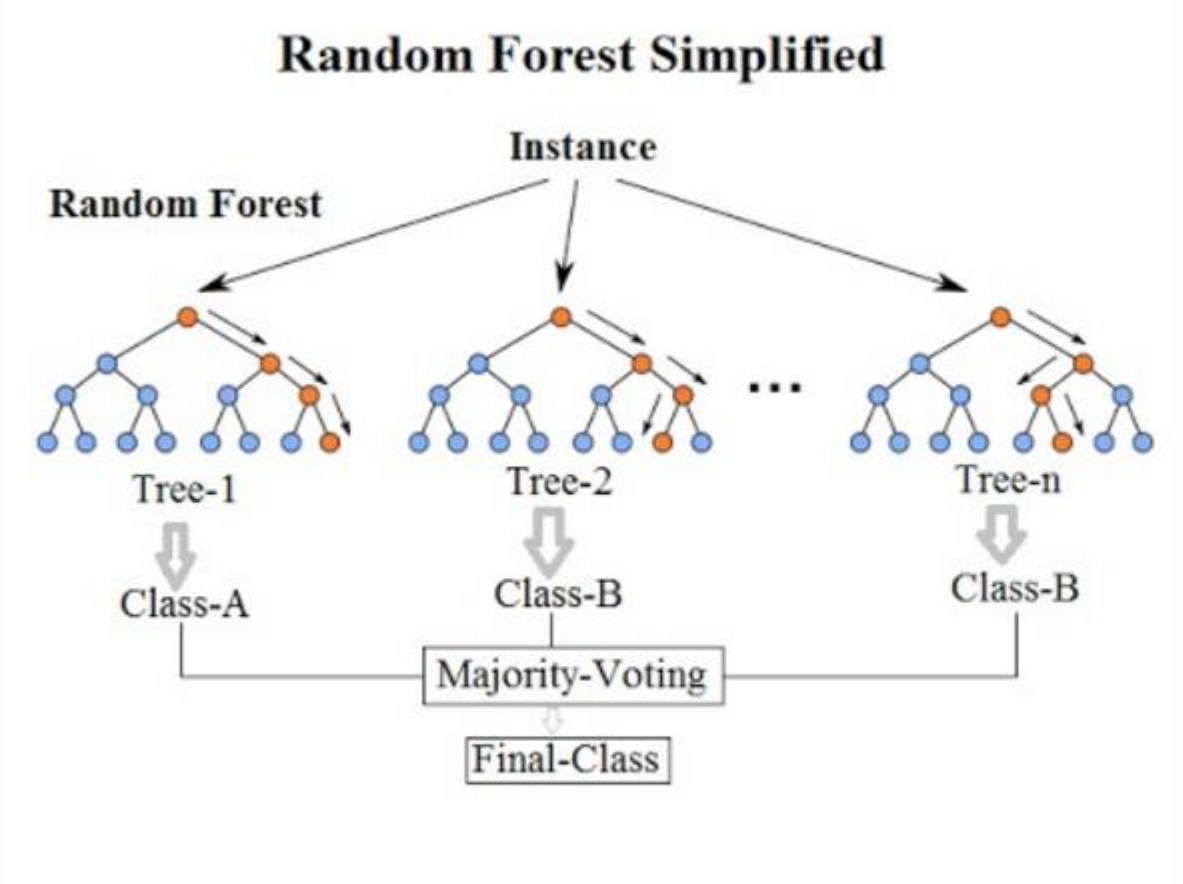
$$entropy = \sum_{i=1}^{c} -p_i * \log_2 p_i$$

- **Every** decision? **Every** feature?

FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

Random Forest

- Creating n Decision Trees
- Combining their outputs:
  - Voting
  - Weighted voting



Scavenger Hunt

- What is the meaning of:
  - Bootstrapping
  - Bagging
  - Boosting