Computational Communication Science 2
Week 8 - Lecture
»A Deep Dive into Supervised Machine Learning«

Marthe Möller
Anne Kroon

a.m.moller@uva.nl, @marthemoller
a.c.kroon@uva.nl, @annekroon

May, 2022

Digital Society Minor, University of Amsterdam

Today

Recap

SML in practice

Cross-validation

SML: Strenghts and Challenges

Recap

Recap

Last week, we talked about:

- Rule-based Text Classification
- Automated Text Classification: SML
- The principles behind SML
- The steps of SML
- Some commonly used ML models
- Validating models

At home, you:

- Worked with SML (homework assignment)

Recap

Today, we:

- Talk about the homework that you did
- Take a deep dive in SML
- Take a critical look at SML

SML in practice

Cross-validation

Let's review the homework assignment!

Cross-validation

Cross-validation

We talked about validating models.

Cross-validation

Overfitting: When a model fits *exactly* against the data.

When we calculate the metrics discussed above for multiple models on the same test dataset, we run the risk of overfitting on the test data.

Potential solution: Split the dataset into three smaller sets. A training dataset, a validation dataset and a test dataset.

However, this requires us to have a very large labeled dataset. In reality, this is not always the case!

Cross-validation

Cross-validation: A resampling procedure to evaluate ML models on a limited data sample.

k-fold cross-validation, where *k* refers to the number of groups or folds in which a sample is split.

k-fold cross-validation

k-fold cross-validation step by step:

1. Shuffle the data
2. Split the data into k folds (groups)
3. For each unique group
   - 3.1 Take the group as a test dataset
   - 3.2 Take the remaining groups as one training dataset
   - 3.3 Fit a model on the training set and evaluate it on the test set
   - 3.4 Retain the evaluation score and discard the model
4. Summarize the evaluation scores to assess the model

Cross-validation

Cross-validation is often used to compare many different model specifications, for example to find the best hyperparameters.

Hyperparameters: Parameters of the model that are not estimated from the data.

To do this, the Grid Search algorithm is often used.

More about hyperparameters in this week's tutorial!

Zooming out

We talked about:

- SML in practice
- Cross-validation

Next, we will talk about:

- Strengths and challenges associated to SML

SML: Strenghts and Challenges

Strengths and Challenges

Strengths:

- Easier to code large datasets
- Enhances replicable research
- Easier to study "natural" human behavior

Disadvantages:

- Resource constraints
- Ethical considerations
- Criticism required (see next slide)

Strengths and Challenges

Belastingdienst koos voor extra controle vooral burgers met lage inkomens

23 november 2021 00:47
Laatste update: 1 dag geleden

761 NUjij-reacties

Het systeem van de Belastingdienst koos ervoor om de kinderopvangtoeslag vooral bij mensen met een laag inkomen extra te controleren. Dat heeft de fiscus toegegeven in antwoord op vragen van *Trouw* en *RTL Nieuws*.

Trouw

Toeslagen

Belastingdienst ging vooral achter lage inkomens aan

Om toeslagen te controleren op fouten en fraude gebruikte de Belastingdienst een zelflerend algoritme. Dat selecteerde vooral lage inkomens voor controle.

Jan Kleinnijenhuis 22 november 2021

De Belastingdienst heeft jarenlang specifiek burgers met een laag inkomen geselecteerd voor extra controle op

Fraudejacht bij Toeslagen

Belastingdienst controleerde extra bij lage inkomens in jacht op fraude

22 november 2021 22:59

Zooming out

We talked about:

- SML in practice
- Cross-validation
- Strengths and challenges associated to SML

This week's tutorial:

- Hands-on approach to take a further look into the machine learning process
- Tutorial goal:
  - To provide a stepping stone so that you can (independently) advance your machine learning skills