Corpus Standards and Evaluation

Damir Cavar

January 2023

Indiana University (ISCA & NLP Lab)
2023 Datathon and ML Competition

Agenda

- JSON
- CoNLL

- Evaluation and data set segmentation
  - Partitioning
  - Cross-Validation

JSON

- See example

- Compared to XML
  - "self describing" (human readable)
  - hierarchical (values within values)
  - can be parsed and used by lots of programming languages
  - can be fetched with an XMLHttpRequest

JSON

- JSON is Unlike XML Because
  - JSON doesn't use end tag
  - JSON is shorter
  - JSON is quicker to read and write
  - JSON can use arrays

JSON vs. XML

- Parsing XML with an XML parser
- Parsing JSON using JavaScript, Python, Java, C++ etc. functions

- XML is more difficult to parse than JSON, JSON is parsed into a usable object

Examples for JSON

- JSON-NLP

- Validator:
  - https://jsonformatter.curiousconcept.com/

CoNLL-X

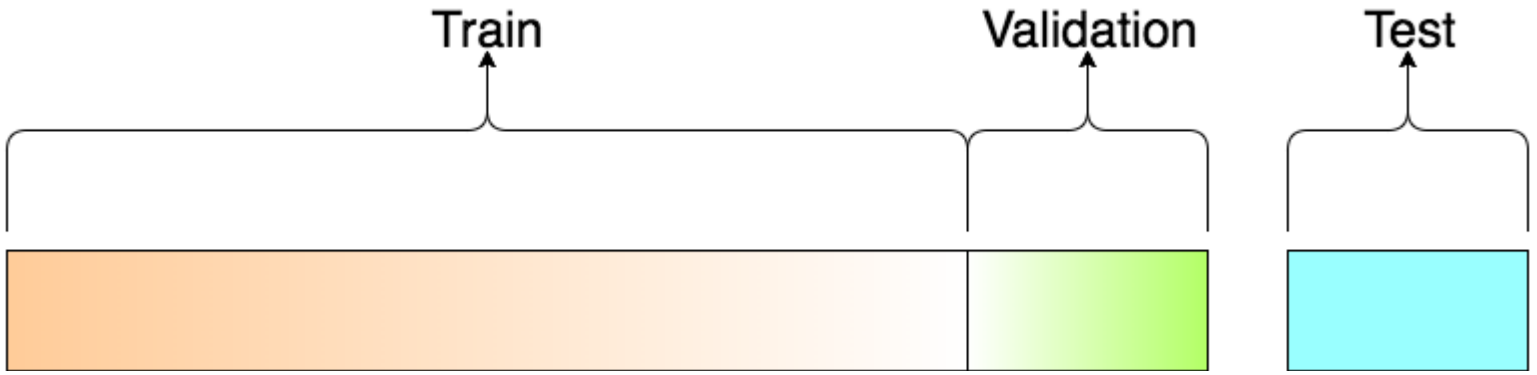
- See paper: Buchholz and Marsi
  - https://www.aclweb.org/anthology/W06-2920
- Plain text annotation
- Unicode encoding: UTF-8
- Line-break character: LF
- Word lines containing the annotation of a word/token in 10 fields separated by single tab characters
- Blank lines marking sentence boundaries
- Comment lines starting with hash (#)

CoNLL Example

- See example data sets:
  - CoNLL 2003 Shared Task
    - https://www.clips.uantwerpen.be/conll2003/ner/
  - CoNLL 2018 Shared Task
    - https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2899

Corpora Split

- Divisions
  - Training Set
  - Development Set
  - Test Set



- 5 and 10 parts for cross-validation

Data Sets

- Selection
  - Random
  - Development Set and Test Set reflect data one might expect in future unseen data sets
  - Size:
    - If data set small: dev and test set might be small and the variation might be high for different parts selected as dev set
    - If data set is large: dev set could be 1% or more
    - Select dev set to reflect accurately the performance of the model
  - Dev set used to measure overfitting

Overfitting

- Example:
  - Train a model on a data set with 100k documents and class labels
  - Result testing model on original data set: 99%
  - Run model on unseen data: drop to 50%
  - The model does not generalize to unseen data, it overfits
- Noise vs. Signal
  - Overfitting is learning noise
- Opposite effect:
  - Underfitting

Underfitting

- Too simple model
  - Too few features in the signal
- Tradeoff between
  - Bias
  - Variance

Detecting Overfitting

- Tests necessary to see how well a model performs on unseen data
- Create a small test set from data and do not use it in training
- Overfitting:
  - Model does better on training set than on test set

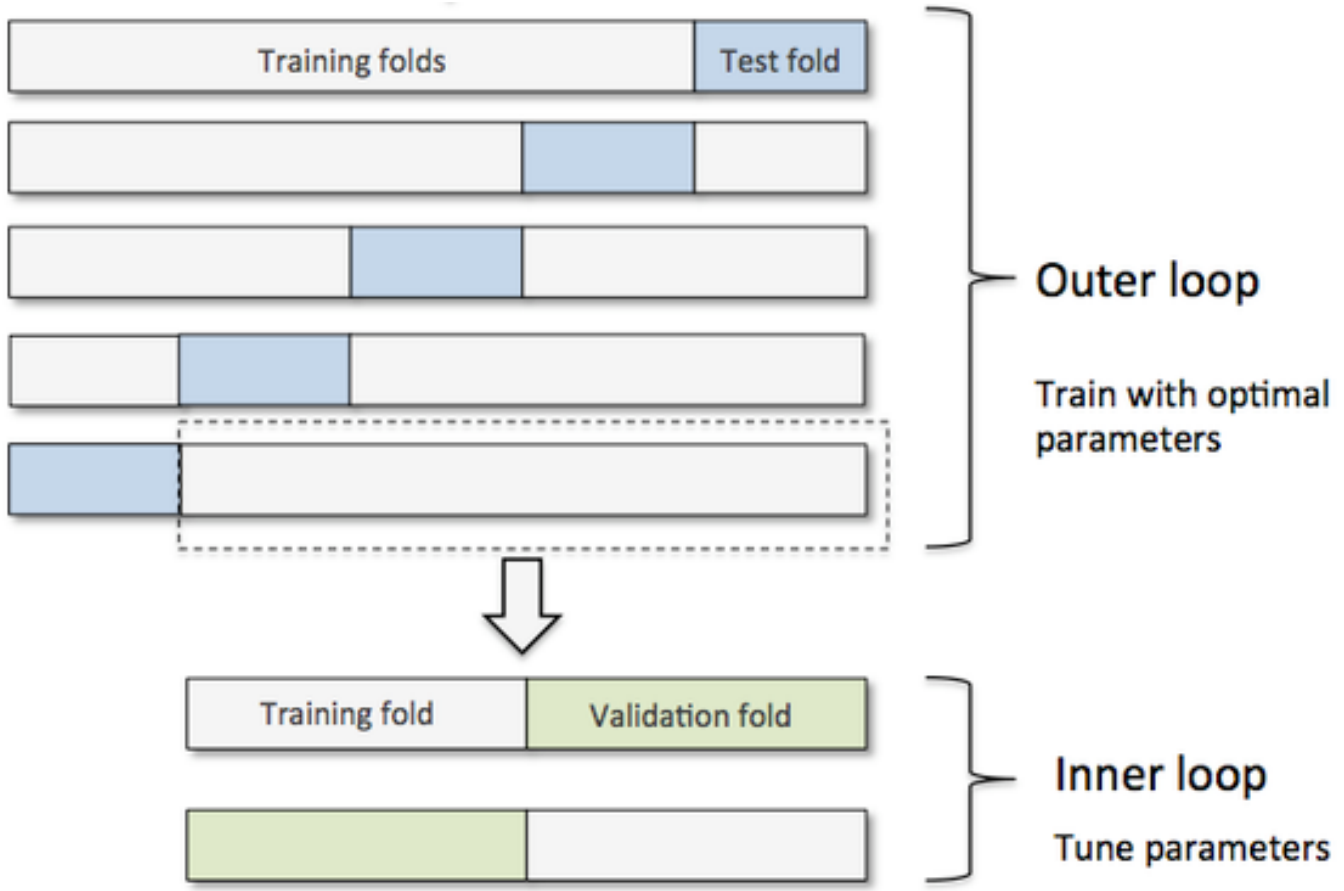
- Alternative approach:
  - Create a simple benchmark model and see whether some ML model performs better

Preventing Overfitting

- Cross-validation
  - K-fold cross-validation
- Approach:
  - Split data into 10 partitions of training, development (or validation), and test folds
  - Evaluate your approach by:
    - Training, optimizing, testing on the 10 different partitions
    - Sum up and average the test scores

Cross-Validation

- Outer and Inner Loop



Cross-Validation

- Average of test scores

Related Data Sources

- CLARIN Virtual Language Observatory
- GitHub repos on hate speech
- Hatespeechdata on GitHub
- Hatebase (inactive)
- Hate speech data on Kaggle
- English data and classification schema
- And many more…