What to do next

Learning Machine Learning

Nils Reiter

CENTER FOR REFLECTED TEXT ANALYTICS

September 26-27, 2018

Overview

Using Machine Learning at Home

Section 1

Using Machine Learning at Home

Using Machine Learning at Home

The Task

What kind of problem do you want to solve?

- Classification: Items to classes
- Sequence labeling: Sequential items to classes
    - By taking previous decisions into account
    - Used in many NLP tasks!
- Regression: Predict numeric values
- Clustering: Data exploration

Using Machine Learning at Home

The Classes

What are the classes?

- Can humans distinguish between them clearly?
- Are there more training instances than classes?
- How specific are the classes to one document/data set?
  - Can we learn something generic from them?
- How are they distributed in the data/in the world?

Using Machine Learning at Home

The Data

- How large is the data set?
- Is it representative of the real world?
- Is it representative for the application?

Using Machine Learning at Home

The Features

Which features to use?

- Features need to be
  - Relevant for the target category
    - Your own judgement
    - Data analysis on a data sample: Association
  - Applicable to large portions of the instances
  - Extractable from the instances
    - How much time do you have?
    - How much preprocessing can you afford?
    - How reliable is the preprocessing?
- Extracting features: Main task for you
  - You'll have to write code

Processing Text

- Languages are different
  - German vs. English vs. Chinese

Processing Text

- Languages are different
  - German vs. English vs. Chinese
- Text types are different
  - Newspaper vs. blog vs. scientific articles

Processing Text

- Languages are different
  - German vs. English vs. Chinese
- Text types are different
  - Newspaper vs. blog vs. scientific articles
- Domains are different
  - Business vs. sports

Processing Text

- Languages are different
  - German vs. English vs. Chinese
- Text types are different
  - Newspaper vs. blog vs. scientific articles
- Domains are different
  - Business vs. sports

Processing Text

Differences are different

- Domain: Vocabulary
- Text types: Vocabulary, syntax, perspective, …
- Language: Syntax, vocabulary, semantics, sign systems, …

Processing Text

Ambiguity

Time flies like an arrow

Processing Text

Ambiguity

Time flies like an arrow

- Texts/sentences/words can be ambiguous
- How many different meanings does the sentence have?

Processing Text

Ambiguity

Angela saw the man with the binocular

Processing Text

Ambiguity

Angela saw the man with the binocular

- Ambiguity reflected in different syntactic readings
- PP attachment ambiguity
  - ‘see with the binocular’
  - ‘man with the binocular’

Processing Text

Processing text is hard

- NLP tools (e.g., Stanford Core NLP)
  - almost always supervised
  - trained on newspaper/Wikipedia/social media
- This may be what you need, but there's no guarantee.

Processing Text

Processing text is hard

- NLP tools (e.g., Stanford Core NLP)
  - almost always supervised
  - trained on newspaper/Wikipedia/social media
- This may be what you need, but there's no guarantee.
- Tools focus on linguistic layers (e.g., parts of speech or coreference)
  - Dependencies between layers exist!
  - PoS tagging errors lead to subsequent errors
    - This gap can be large Reiter (2014)

Processing Text

Processing text is hard

- NLP tools (e.g., Stanford Core NLP)
  - almost always supervised
  - trained on newspaper/Wikipedia/social media
- This may be what you need, but there's no guarantee.
- Tools focus on linguistic layers (e.g., parts of speech or coreference)
  - Dependencies between layers exist!
  - PoS tagging errors lead to subsequent errors
    - This gap can be large Reiter (2014)
- Technical text quality matters
  - 'Garbage in, garbage out'
  - OCR is not perfect

Supervised vs. Unsupervised

Supervised vs. Unsupervised

Two strains of machine learning

Supervised Learning

- Goal: Replicate the gold standard
- Known classes
- Models trained on training data

$\rightarrow$ Classification

Supervised vs. Unsupervised

Two strains of machine learning

Supervised Learning

- Goal: Replicate the gold standard
- Known classes
- Models trained on training data

$\rightarrow$ Classification

Unsupervised Learning

- Goal: Identify groups of ‘similar’ items, similarity measured via features
  - Data exploration
- No gold standard, no training data

$\rightarrow$ Clustering

- Results not necessarily interpretable for humans!

Section 3

Data & Annotation

Data

- Supervised ML needs (training/testing) data
- For text: Annotations!

Data

- Supervised ML needs (training/testing) data
- For text: Annotations!
- Corpus annotation
  - Tradition/established in computational linguistics
  - Explicitly marked linguistic categories
    - e.g., parts of speech (verb/noun/adjective/…)

Getting Annotated Corpora

- LDC: Linguistic Data Consortium
  - https://www.ldc.upenn.edu
  - Intransparent business model ...
- ELDA: European Language Resources Association
  - http://www.elra.info
- Open Access
  - Oxford Text Archive: http://ota.ox.ac.uk
  - Deutsches Textarchiv: http://www.deutschestextarchiv.de
  - TextGrid Repository: https://textgridrep.org

Creating Annotated Corpora

- Non-trivial
  - Difficult decisions
  - Large list of special cases, exceptions
- Expensive
  - Multiple annotators
  - Supervision
- Time-consuming
  - Concentration fades quickly

Creating Annotated Corpora

- Non-trivial
  - Difficult decisions
  - Large list of special cases, exceptions
- Expensive
  - Multiple annotators
  - Supervision
- Time-consuming
  - Concentration fades quickly

$\Rightarrow$ Annotated data is valuable

Creating Annotated Corpora

Best Practice

- Annotation guidelines mediate between theory and annotators
  - Not every annotator needs to be an export on syntactic theory
- Parallel annotation: Multiple annotators annotate the same text
  - Allows estimation of annotation quality
  - Regularly measure inter-annotator agreement
- Iteratively improve the annotation guidelines
  - This might invalidate previous annotations!

Annotation Guidelines

- Mediator between theory and annotations
- Applicability is important
  - Self-contained
  - Clarity
  - Work of reference

Part-of-Speech Tagging Guidelines for the Penn Treebank Project

Beatrice Santorini

March 15, 1991

2 List of parts of speech with corresponding tag

Adjective—JJ

Hyphenated compounds that are used as modifiers are tagged as adjectives (JJ).

EXAMPLES: happy-go-lucky/JJ
one-of-a-kind/JJ
run-of-the-mill/JJ

Figure: Part of Speech Guidelines used in the Penn Treebank

Inter-Annotator Agreement

Motivation

- IAA expresses agreement between annotators/raters quantitatively
- Often used as an upper bound in NLP:
  Computers can't be expected to perform better than human agreement
- Annotations with high IAA are considered more reliable
- Sometimes used to steer guideline/resource development
  - '90% solution': Remove word senses for which annotators achieve less than 90% Hovy et al. (2006)
- Corpus releases should be accompanied by IAA values, to allow estimation of annotation quality

Inter-Annotator Agreement

Different Metrics

- Not all annotation tasks are the same
  - PoS tagging: Assign each word to a category
    - Only categorizing
  - Sentence splitting: Mark sentence boundaries
    - Only unitizing
  - Named entities: Select a span *and* assign it to a category
    - Unitizing, categorizing
- Different metrics for different tasks!

Cohen 1960; Fleiss 1971; Fournier and Inkpen 2012; Mathet et al. 2015

Inter-Annotator Agreement

Different Metrics: Common Properties

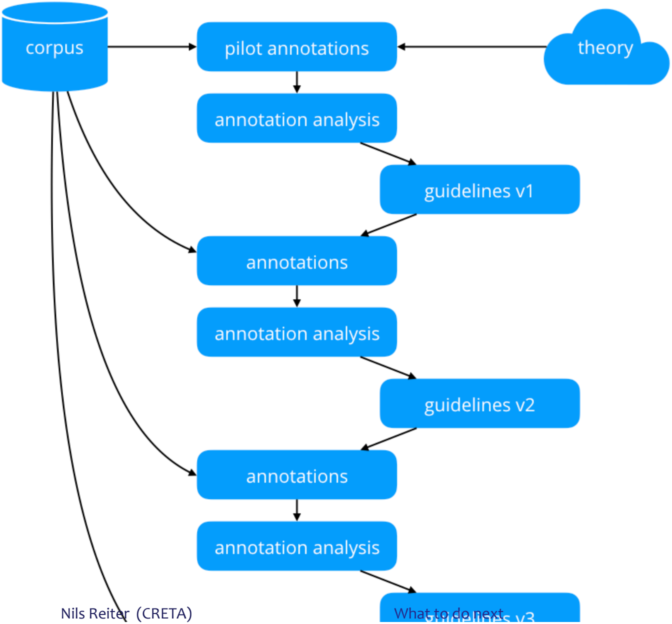
- All metrics incorporate *observed* and *expected* agreement
- Observed agreement: Extracted from the annotations
- Expected agreement: Agreement to be expected by chance annotations
  - Indicates difficulty of the annotation task
  - Allows comparing agreement values with different numbers of categories!

Inter-Annotator Agreement

Expected Agreement

If two annotators assign word classes (noun, verb, adjective, other) by throwing a 4-sided die, they achieve a certain level of agreement (this is a categorization task).

Annotation Workflow



Section 4

Resources

Continue Learning

- Coursera online course
  - Andrew Ng, Stanford University
  - https://www.coursera.org/learn/machine-learning
  - Lecture and exercises, generic (not only text/language)
- Books
  - Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press, 1999
  - I. H. Witten and Eibe Frank. *Data Mining*. 2nd ed. Practical Machine Learning Tools and Techniques. Elsevier, Sept. 2005
  - Dan Jurafsky and James H. Martin. *Speech and Language Processing*. 2nd. Prentice Hall, 2008

Start Coding

- You do not have to implement everything by yourself
  - Frameworks and APIs are faster, more tested, better documented
- Python
  - Natural Language Toolkit (NLTK): https://www.nltk.org
  - scikit-learn http://scikit-learn.org/
  - Industrial-Strength NLP https://spacy.io
- Java
  - Weka https://www.cs.waikato.ac.nz/ml/weka/
  - Mallet http://mallet.cs.umass.edu
  - Apache UIMA http://uima.apache.org
  - ClearTk http://cleartk.github.io/cleartk/
- R
  - caret https://topepo.github.io/caret/

References I

- Cohen, Jacob. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.
- Fleiss, Joseph L. “Measuring nominal scale agreement among many raters”. In: *Psychological Bulletin* 76.5 (1971), pp. 420–428.
- Fournier, Chris and Diana Inkpen. “Segmentation Similarity and Agreement”. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montrèal, Canada: Association for Computational Linguistics, 2012, pp. 152–161.
- Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. “OntoNotes: The 90% Solution”. In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Ed. by Robert C. Moore, Jeff Bilmes, Jennifer Chu-Carroll, and Mark Sanderson. New York City, USA: Association for Computational Linguistics, June 2006, pp. 57–60.

References II

- Jurafsky, Dan and James H. Martin. *Speech and Language Processing*. 2nd. Prentice Hall, 2008.
- Manning, Christopher D. and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press, 1999.
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3 (2015), pp. 437–479.
- Reiter, Nils. "Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms". Submitted on Aug 13, defended on Nov 27. PhD thesis. Heidelberg University, June 2014.
- Witten, I. H. and Eibe Frank. *Data Mining*. 2nd ed. Practical Machine Learning Tools and Techniques. Elsevier, Sept. 2005.