
DATA
LAB

Brief Tutorial on Statistical Methods in Natural Language Understanding

Example of Text Classification

Charlotte Gils

Natural language processing (NLP)

- NLP: Process and analyze large amounts of natural language data
- Common problems in NLP:
  - Natural language generation
  - Speech recognition
  - Natural language understanding
- Methods:
  - Rule-based
  - Statistical: Automatically learn rules and patterns by processing large corpora
    - Fred Jelinek "Anytime a linguist leaves the group the recognition rate goes up"

Natural Language Understanding

- Some common tasks:
  - Machine translation
  - Automatic summarization
  - Part-of-speech tagging
  - Named-entity recognition
  - Named-entity disambiguation
  - Syntactic annotation (parsing)
  - Text classification

Text Classification using Supervised Machine Learning

- Obtain **training data**
  - Documents $d_1, d_2, \ldots, d_N$
  - Corresponding category(ies) for each document $c_1, c_2, \ldots, c_N$
- From the training data, **learn** a function f that maps a document d to category(ies) c
- **Infer** category(ies) of new documents using *f(d)*

Text Classification Basics I: Bag-of-Words

- Create a set of variables from document d (**feature engineering**)
- Simplest approach: **Bag-of-words** (BOW)
  - Suppose there are M words $w_1, w_2, \dots, w_M$ in all the documents
  - Word j appears k times in in d_i: variable $x_{ij} = k$
- Example:
  - d_1 = «The dog is in its house.»
  - d_2 = «That house is old.»
  - d_3 = « The dog is old. That old.».
  - $X = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$

Text Classification Basics II: Algorithm

- Suppose two categories c = 0 and c=1
- A simple algorithm is **logistic regression**
- Divide the variable data space by hyperplanes
  - $\mathrm{f(x)} = \frac{1}{1+\exp(x^T w + w_0)}$ probability that x in category 1
  - $x = (x_1, x_2, \ldots, x_M)$ features of new document
  - $w = (w_0, w_1, .., w_M)$ model coefficients that were learned
- Maximize log-likelihood to learn w from observed data X, c
  - $L = Log \prod_{i=1}^{N} f(x_i)^{c_i} (1 - f(x_i))^{1-c_i}$
  - Use optimization methods

Text Classification Basics III: Evaluation

- Precision(c) = # samples correctly predicted to be in class c / # samples predicted to be in class c
- Recall(c) = # samples correctly predicted to be in class c / # samples actually in class c
- $\mathrm{F1(c)} = 2 * \mathrm{precision(c)} * \mathrm{recall(c)} / (\mathrm{precision(c)} + \mathrm{recall(c)})$
- See whiteboard example
- Business requirements typically dictate the appropriate evaluation metrics (and how high it must be)
- A model may be tuned to optimize the chosen metric (typically at the expense of other metrics; e.g., recall vs precision, class c_1 vs class c_2).

What Determines How Well a Model Works

1. Data
   - Are the inputs and the labels related in the first place ?
   - Amount of data
   - Quality of data: Incorrect and missing inputs, incorrect labels, imbalanced classes, biased training data, etc.
   - Are the distributions of training data and data that the model is applied to the same (model drift/covariate shift)?
2. Feature engineering
3. Machine Learning Algorithm

Aspects of Text Classification Project

- Obtain labeled data
- Data preprocessing and management
- Model training: See jupyter notebook
- Productionize model
- Monitor model outputs
- Improve, re-train , update model