Text Classification

(SNLP tutorial)

Vilém Zouhar

June 4, 2021

Overview

- Task, approaches
- Features
- - Document Frequency
- - Information Gain
- - Pointwise Mutual Information
- - Pointwise Mutual Information
- - χ^2
- - Term Strength
- Classification Algorithms
- - Decision Trees
- - Naïve Bayes
- - kNN
- - Perceptron
- - SVM
- Homework

TODO multicolumn

Text Classification

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow \{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow \{\text{SPAM}, \text{BENIGN}, \text{MARKETING}\}$
- Author identification/profiling: Document(s) $\rightarrow \{\text{F. Bacon}, \text{W. Shakespeare}, \ldots\}$
- Native language identification: Document $\rightarrow \{\text{German}, \text{Polish}, \ldots\}$
- POS Tagging: Sentence $\rightarrow \{NN, VERB, PART., \ldots\}^{|S|}$
- Sense Disambiguation: Word+sentence $\rightarrow$ Senses of Word

Classification vs. Clustering

	Classification	Clustering
Method	Supervised	Unsupervised
Classes	Given	Unknown
# Classes	Given	(Mostly) unknown

Binary vs. Multi-Class Classification

Multi-Class

- $f : D \to \{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$

Binary

- $f_1 : D \to \{\text{politics}, \text{not politics}\}$
- $f_2 : D \to \{\text{NLP}, \text{not NLP}\}$
- $f_3 : D \to \{\text{healthcare}, \text{not healthcare}\}$
- $\ldots$

Flat vs. Hiearchical

TODO

Single-Category vs Multi-Category

TODO

Feature Extraction

- Move from text to more processable domain

Binary/indicator features

$$f_b(doc) = \begin{cases} 1 & \text{Contains string } \texttt{"Super free \$\$\$ discount"} \\ 0 & \text{Otherwise} \end{cases}$$

Integer features

$f_i(doc) =$ Number of occurences of `"buy"`

Real-valued features

$$f_r(doc) = \frac{\text{Number of occurences of } \texttt{"buy"}}{|doc|}$$

Feature Selection

TODO

Document Frequency

DF

$$df(term) = \frac{|\{doc|term \in doc, doc \in D\}|}{|D|}$$

- Remove rare items ($df \leq \frac{2}{|D|}$)
  Won't occur in new documents anyway
- Remove frequent items ($df = 1$)
  Usually stop words
  No information

- Sometimes not a good idea (interaction with other terms, etc.)
- Stopword distribution gives information in author identification

Information Gain

- Information gained (reduction in entropy) by knowing term present or not

$$
\begin{aligned}
G(C,t) =& H(C) - H(C|t) \\
=& -\sum_i p(c_i) \log p(c_i) \\
&+ \; p(t) \sum_i p(c_i, t) \log p(c_i, t) \\
&+ \; p(\overline{t}) \sum_i p(c_i, \overline{t}) \log p(c_i, \overline{t})
\end{aligned}
$$

Pointwise Mutual Information

- Difference between observed distribution and independent

$$\text{pmi}(c_i, t) = \log \frac{p(c_i, t)}{p(c_i) \cdot p(t)}$$

- TODO (expansion using Bayes)
- TODO (average, max)
- TODO (relation to mutual information)

χ^2

$\chi^2(c_1, c_2) = \sum_{tt,tf,ft,ff}(O - E)^2$

- TODO example
- TODO table
- χ^2 avg vs. χ^2 max (multiple categories)

Term Strength

- Two documents: d_1, d_2
- Term t
- $p(t \in d_2 | t \in d1)$
- *What is the probability that the term t will be in d_2 given that it is in d_1?*
- If two documents related $\rightarrow$ high probability
- If two documents not related $\rightarrow$ low probability
- "Constant" with stop words

Classification

TODO

Decision Trees

TODO

Naïve Bayes

TODO

kNN

TODO

SVM

- Find a boundary that maximizes the distance to closest vectors
- If not possible, find one that minimizes the error
- Add the kernel trick

Perceptron

- Binary classification
- Linear boundary in feature space
- $\hat{y} = \text{sign}(wx + b)$

Algorithm:

- $w_0 = \overrightarrow{0}$
- For every data point x_i
- ▸ $\hat{y}_i = \text{sign}(w_k x_i + b)$
- ▸ if $\hat{y}_i \neq y_i$:
- ▸ ⋆ $w_{k+1} = w_k - \hat{y}_i \cdot x$
- ▸ else:
- ▸ ⋆ $w_{k+1} = w_k$
- TODO: illustration
- TODO: advantages/disadvantages

Resources

1. UdS SNLP Class, WSD: https://teaching.lsv.uni-saarland.de/snlp/
2. Classical Statistical WSD: https://www.aclweb.org/anthology/P91-1034.pdf