Assignment 7, 8 + Text Classification Basics

(SNLP Tutorial 8)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

15th, 17th June

Assignment 7

- Exercise 1: Count Tree
- Exercise 2: Kneser-Ney Smoothing
- Bonus: Smoothing Techniques

Text Classification

Fill in the classes:

- 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

Issues with this?

Classification vs. Clustering

	Classification	Clustering
Method	???	???
Classes	???	???
# Classes	???	???

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

Binary vs. Multi-Class Classification

Multi-Class

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

How to turn this into a binary classification?

Binary

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

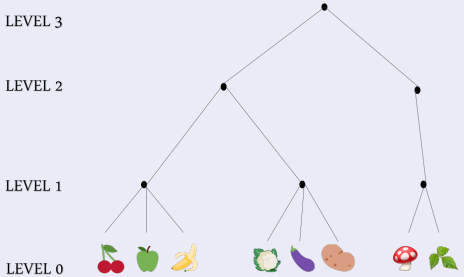
How to turn multiple $\{\text{multi-class}|\text{binary}\}$ into a single multi-class?

Flat vs. Hiearchical

Flat Classification

$f_1 : D \rightarrow$ 🍌 🍏 🍆 🍒 🍇 🥔 🥦 🌿

Hierarchical Classification



What is this structure similar to in classification?

Feature Extraction

- Move from text to more processable domain
- How? (at least three "approaches")

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

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

Name a scenario where you can use each of these...

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
- Why not always a good idea?
- How is this feature used in retrieval?

Term Frequency - Inverse Document Frequency

TF-IDF

$$tf(term, doc) = \frac{count_{doc}(term)}{|doc|}$$

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

$$idf'(term) = \frac{|D|}{df(term)}, idf(term) = \log_2\left(\frac{|D|}{df(term)}\right)$$

$$tf - idf(term, doc) = tf(term, doc) \times idf(term)$$

- How can we use tf-idf for document similarity?

Information Gain

- Information gained (reduction in entropy) by knowing whether a term is present

Information Gain

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

Questions

- When is Information Gain 0? When is it positive? Can it be negative?
- Term t occurs in all classes equally. Is it a good feature?

Pointwise Mutual Information

- Difference between observed distribution and independent

PMI

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

Another formulation

$$PMI(t, c) = \log \frac{A \cdot D}{(A + C) \cdot (A + B)}$$

where

A = co-occurrence of c and t

B = #times t occurs without c

C = #times c occurs without t

D = #documents in c

Pointwise Mutual Information

Relation to Mutual Information

MI = weighted pmi = expectation of pmi over all events

Questions

- When is PMI 0?
When is it positive?
Can it be negative?

Chi Square χ^2

$$\chi^2_n(c_1, c_2) = \sum_j \frac{(O_j - E_j)^2}{E_j} \tag{1}$$

- n : Degrees of freedom
- O_j: Observed absolute frequency of the feature j
- E_j: Estimated absolute frequency of the feature j
- $E_j = p_j \cdot N$
- N: Number of observations in one class

- Null Hypothesis: ?
- The two events are independent.

χ^2 Example

$$\chi^2(c_1, c_2) = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{2}$$

Imagine a language with the following syllable structure: CV, $C \in \{p, k\}$, $V \in \{a, u\}$:

C/V	k	p	
a	75	33	108
u	31	61	92
	106	94	200

- What is expected number of occurences of `ka`?
- $p_a = \frac{108}{200} = 0.54$, $N_k = 75 + 31 = 106$
- $E_{ka} = p_a \cdot N_k = 0.54 \cdot 106 = 57.24$
- $\frac{(O_{ka} - E_{ka})^2}{E_{ka}} = \frac{(75 - 57.24)^2}{57.24} \approx 5.51$

χ^2 Example, continued

$$\chi^2(c_1, c_2) = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{3}$$

- And so forth for all other cells

C/V	k	p
a	5.51	*pa*
u	*ku*	*pu*

χ^2 Example, continued

- And so forth for all other cells:

C/V	k	p
a	5.51	6.21
u	6.47	7.29

- $\chi^2 = 5.51 + 6.21 + 6.47 + 7.29 = 25.48$
- Degrees of freedom: $df = (\#_{rows} - 1) \cdot (\#_{cols} - 1) = (2-1) \cdot (2-1) = 1$
- Choose significance level α
- What are common significance levels?
- Look up χ^2-value in a χ^2-table
- Reject H_0 if $\chi^2 > \chi^2_{(\alpha,df)}$

χ^2 Table Lookup

- Calculated $\chi^2 = 25.48$
- $df = 1$
- $\alpha = 0.05 \rightarrow \chi^2 = 3.84$

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.790
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582	41.610	43.820
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997	43.072	45.315

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

Questions

Can we use term strength for

- Stopword removal
- Document Clustering

Resources

1. UdS SNLP Class: https://teaching.lsv.uni-saarland.de/snlp/
2. Information Gain in decision trees: https://en.wikipedia.org/wiki/Information_gain_in_decision_trees#Example
3. PMI in classification: https://towardsdatascience.com/multinomial-na%C3%AFve-bayes-classifier-using-pointwise-mutual-information-9ade011fcbd0
4. χ^2 table: https://www.medcalc.org/manual/chi-square-table.php
5. χ^2 example: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
6. Term Strength: http://mlwiki.org/index.php/Term_Strength
7. Comparison of Feature Selection Techniques: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9956&rep=rep1&type=pdf