Statistical Natural Language Processing (SS-2018)
Submission Deadline: 08.06.2018, 23:59

June 1, 2018

Feature Selection (6 points)

1) (1.5 points) Miscellaneous Problems

- Part a (0.75 points):
  (i) Consider a document containing 100 words wherein the word 'Eutopia' appears 3 times. What is the term frequency for the term 'Eutopia'? → $3/100$
  (ii) Now, assume we have 100 million documents and the word 'Eutopia' appears in ten thousand of these. What is the inverse document frequency (idf) of this term? $\log_{10}$ (100 Mio / 10.000)
  Note: Take log to the base 10. Click on this link to learn about tf-idf
  (iii) Finally, calculate the tf-idf weight from the values obtained from Part a and Part b.
  → $3/100 \cdot \log$ (100 Mio / 10.000)
- Part b (0.25 points):
  Choose only one best option and explain your choice.
  When training a language model, if we use an overly narrow corpus, the probabilities
  a. Doesn't reflect the task
  b. Reflect all possible wordings
  c. Reflect intuition
  d. Dont generalize
  ⇒ the text is ~~too~~ specific and does not cover many characteristics of the language
- Part c (0.5 points)
  You are an English Literature teacher and you ask your class to write a play in the style of Shakespeare. You want to score their plays using a trigram language model you computed from a corpus of all Shakespeare plays but you find that the data is too sparse and most of your students sentences receive a score of zero.
  How would you use a back-off model to alleviate this problem? Your short answer should be between 50-100 words. → Many points trigram less probabilit less points bigram less probabilit | a certain point unigram used by students spe[illegible] | no point | unigram does not appear in Shakespeare

2) (2 points) Imagine we have a predefined set of class labels 'Coffee' and 'Tea'. The following table containing the entry of the counts is given for this purpose:

	Σ
500	500
500	500
Σ 1000	

	black	beans	leaves	rest
Class = Coffee	500	1000	100	400
Class = ∼Coffee	500	50	1200	9450
Class = Tea	750	110	1300	400
Class = ∼Tea	1000	1500	200	7350

Σ

$\frac{1000}{13.200} \cdot \frac{2000}{1310}$ [illegible]

- (1 point) What are good features for predicting class 'Coffee'? Explain your findings.
- (1 point) What are good features for predicting class 'Tea'? Explain your findings.
  Hint: Do this task by checking χ^2 value for all possible features and perform the feature selection.
  Ex: χ^2(black, Coffee). ↳ nehmen wo χ^2 hoch ist

3) (2.5 points) You learnt about the feature selection for unsupervised learning (Slide 23 onwards, Chapter 6) in the lecture. Observe the two plots below and comment about the following.

- (1 point) In Plot 1, are features x and y redundant/useful/irrelevant for defining the clusters? Explain your choice.

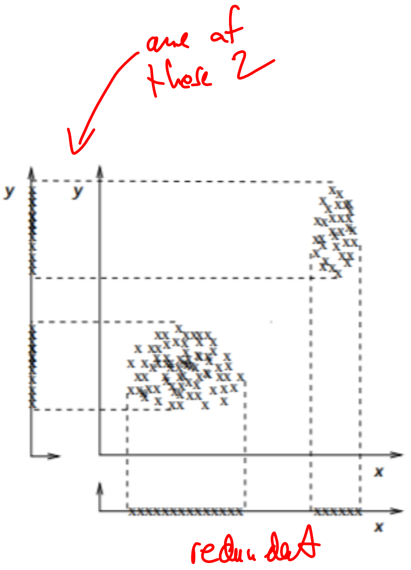


Figure 1: Plot 1

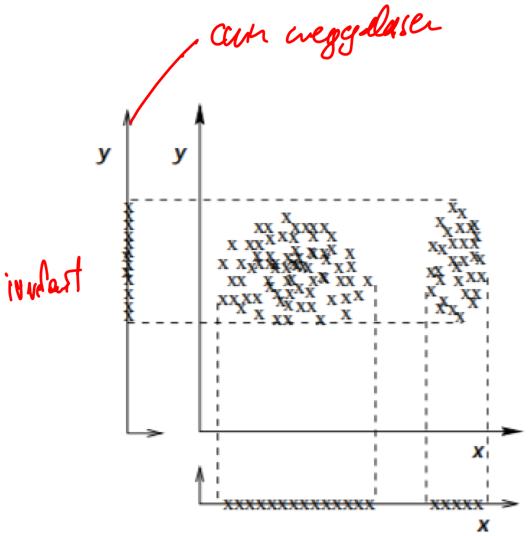
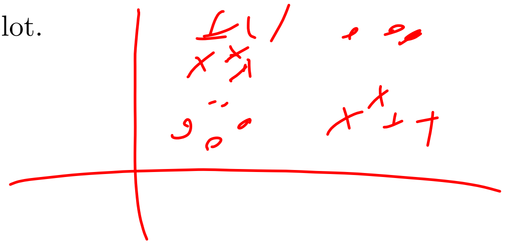


Figure 2: Plot 2

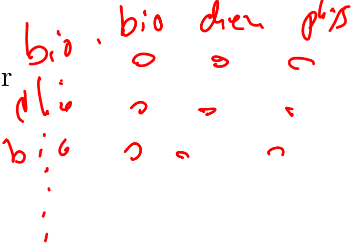
- (1 point) In Plot 2, are features x and y redundant/useful/irrelevant for defining the clusters? Explain your choice.
- (0.5 point) Draw an example plot which portrays feature x and y as the remaining choice from 'redundant/useful/irrelevant' for defining the clusters which was not chosen as an answer in above two parts for plot 1 and plot 2. Explain your plot.



Mutual Information (4 points)

4) (4 points)

- Use the documents provided in "Materials/train" to construct the vocabulary. You need this vocabulary for the next exercise as well. Remember to do the text preprocessing:
  - stopword removal with the stopwords.txt given in the Materials Folder
  - lowercasing (Can use NLTK 3.3: Reference Link)
  - lemmatization + stemming Reference Link (Can use NLTK 3.3: Reference Link)
  - tokenization (Can use NLTK 3.3: Reference Link)



- (1 point) Find the mutual information between each term and each class (topic). Compute *pmi*(t) in the case we want each term to discriminate well for a single category. see slide 50
- (1 point) Use the *pmi*(t) s to do the feature selection such that it results in 10 features and report them. How much has your problem's dimension decreased?
- (1 point) Do the feature selection this time by MI [1] and select the 10 terms with greatest MI. How do these features differ from the previous part of the question? Report these features and their differences with previous part.
- (1 point) Use the features obtained from each case separately to classify each test file by Naiive Bayes Classifier.
  - Compute the likelihoods for each word (after feature selection) in each class (topic)
  - Assume uniform prior probability for classes
  - Classify by posterior probability

[1] https://en.wikipedia.org/wiki/Mutual_information in which each term is a random variable

Submission Instructions

- You must form groups of 2 to 3 people
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

  Exercise_07_MatriculationNumber1_MatriculationNumber2.zip
- Provide in the archive:
  i. your code, accompanied with sufficient comments
  ii. a PDF report with answers, solutions, plots and brief instructions on executing your code
  iii. a README file with the group member names, matriculation numbers and emails
  iv. Data necessary to reproduce your results

- The subject of your submission mail **must** contain the string [SNLP] (including the braces) and explicitly denoting that it is an exercise submission, e.g:

  [SNLP] Exercise# Submission MatriculationNumber1 MatriculationNumber2
- Depending on your tutorial group, please send your assignment to the **corresponding tutor**:
  - Mo. 16-18: Lukas Lange *s9lslang@stud.uni-saarland.de*
  - Th. 14-16: Harshita Jhavar *snlp18.thursday@gmail.com*
  - Fr. 10-12: Marimuthu Kalimuthu *mkalimuthu@lsv.uni-saarland.de*

- If two teams submit same solutions, both teams will be given 0 points and no presentation chance will be given to the two team members. If you do this again, you will be disqualified from the exam.